Available online at www.sciencedirect.com
SCIENCE DIRECT®
Journal of Molecular Graphics and Modelling 22 (2004) 487–497
Journal of Molecular Graphics and Modelling
www.elsevier.com/locate/JMGM

Modelling in vitro hepatotoxicity using molecular interaction fields and SIMCA

Robert D. Clark [a,*], Philippa R.N. Wolohan [a], Edward E. Hodgkin [a], James H. Kelly [b], Norman L. Sussman [b]

[a] *Tripos Inc., 1699 South Hanley Road, St. Louis, MO 63144, USA*
[b] *Amphioxus Cell Technologies Inc., 11222 Richmond Avenue, Houston, TX 77082, USA*

Accepted 4 March 2004

Available online 27 April 2004

Abstract

There is currently a great deal of interest in creating computational tools for predicting the pharmacological properties of drug development candidates, ranging from physicochemical properties such as pK_a and solubility to more complex biological properties such as oral bioavailability and toxicity. The limiting factor in many cases is a shortage of good data from which to construct training sets. In other cases, large amounts of data are available, but they use surrogate end-points or are comprised of compounds very different from those usually encountered in drug discovery and development. In such cases large training sets and global models are not necessarily better than local models based on smaller data sets. Such considerations make it as important to examine the available data carefully so as to avoid over-interpretation of the models obtained as it is to minimise errors in prediction per se. The kinds of complications likely to be encountered for in vitro hepatotoxicity modelling are discussed in general terms and illustrated in particular by SIMCA analysis of data obtained from assays of cultured hepatocytes for a large, structurally diverse data set and a smaller, much more focussed one.

Keywords: In silico toxicology; In vitro hepatotoxicity; SIMCA; CoMFA; Molecular field analysis; HepG2/C3A

1. Introduction

There is currently a great deal of academic and industrial interest in anticipating problems that might cause a drug candidate to drop out late in the development cycle, particularly after expensive clinical trials have begun. Hence early identification of drug metabolism, pharmacokinetic, and toxicological (DM-PK/Tox) liabilities has become an area of particular interest, and has been the focus of many publications and symposia in recent years. Approaches to the problem have included large-scale implementation of high-throughput solubility assays as well as in vitro model systems such as Caco-2 permeability and microsomal oxidation screens. Medium-throughput assays such as cytotoxic effects on hepatic cells in culture are also being used. Unfortunately, all such methods can only be applied *after* compounds have actually been synthesised, and the advent of combinatorial synthesis has made it possible to generate a very large number of compounds with quite poor properties rather quickly. Even assuming the sheer number of compounds produced was not a problem in itself, then the assays involved consume an amount of material that is often prohibitive, especially in the context of combinatorial chemistry.

Such considerations have combined to make development of predictive models based on molecular structure a high priority for computational chemists. One major use anticipated for such models is to avoid ever synthesising compounds that are likely to eventually fail, thereby obviating the need to drop them from development later. The direct savings expected in terms of wasted synthesis and testing would be substantial, but reducing diversion of resources from better leads and better candidates is probably more important in the long run. Anything that will serve to prevent crowding out of lead chemistries that may be less potent in vitro but that have good DM-PK/Tox profiles is likely to be of significant value.

Reliable in silico systems for identifying all potential problems are still a long way off [1,2], but several programs have been developed to address individual aspects thereof. These are usually grouped under the rubric of ADME/Tox (absorption, distribution, metabolism, excretion and toxicol-

* Corresponding author. Tel.: +1-314-951-3365; fax: +1-314-647-9241.
E-mail address: bclark@tripos.com (R.D. Clark).

1093-3263/$ – see front matter © 2004 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmgm.2004.03.009

ogy; "pharmacokinetics" is generally reserved for more integrated properties such as bioavailability and clearance half life). Examples of more or less successful efforts published to date include many solubility models as well as models for predicting intestinal absorption [3,4], blood–brain barrier permeability [5,6], and oxidative metabolism by cytochrome P450s [7–10].

A reductionist approach may be effective in those cases where some single factor is responsible for bad pharmacokinetics, even when the reliability of each constituent model is modest. In fact, one of the best-performing methods for eliminating compounds likely to suffer from poor oral bioavailability is Lipinski's "Rule of 5" [11], which is based on a set of four simple property filters. Such approaches, however, fail to address the greater challenge: to identify the all too common hard cases where no *single* property is disqualifying but the aggregate is, and where structurally idiosyncratic effects such as metabolism and active transport determine whether a compound is a good candidate for further development or a bad one.

Molecular interaction fields are excellent candidates as structural descriptors for such an holistic approach, in that they have the potential to account for all possible interactions between a ligand and its environment in an extremely general way. Their use for effectively capturing interactions with enzymes and receptors underlies much of the 3D QSAR field [12–14], and they have been used to good effect for modelling binding to some specific cytochrome P450s [8]. The QSAR and DM-PK/Tox contexts differ in at least one respect, however: in the former case, some common substructure or shared pharmacophoric pattern exists with respect to which each molecule can be aligned [15], whereas no analogous *extrinsic* alignment rule can exist for the latter application.

The VolSurf program [4][1] was created, in part, to address this complication. It generates alignment-independent descriptors by extracting characteristic values from molecular field contours calculated using the GRID force field [16][2] across a range of probe types. Cruciani et al. have demonstrated that these descriptors (together with molecular volume, surface area, volume-to-surface ration, and globularity) are quite effective for predicting protein binding to serum protein, skin and Caco-2 cell layer permeation [4]; passive blood–brain barrier permeability [6]; and, in conjunction with a protein homology model, likely sites of oxidation by CYP2C9 [10]. These field autocorrelation descriptors effectively capture the *overall* nature of a structure and so are very effective for modelling the individual physical properties involved in applications early in the drug development cycle. They discard the spatial correlations *between* fields, however, which necessarily limits their usefulness for modelling complex anisotropic interactions such as toxicity that are important later in development.

We have taken a third approach. CONCORD[3] is used to generate 3D structures in characteristic conformations, just as is done for VolSurf. Rather than aligning structures to each other based on common substructure, an internal coordinate system is defined for each individual molecule based upon its geometric centre, principal axes, and the dipole moments along those axes. Molecular fields (either standard van der Waals (steric) and coulombic (electrostatic) CoMFA fields or Gaussian CoMSIA [17] fields) are then generated with that orientation as a frame of reference. Such idiotropic field orientation for comparative molecular field analysis (IFO-CoMFA) was originally described purely in terms of steric alignment [18], but has recently been extended to encompass orientation based on electrostatic fields as well [19]. The net result of the procedure is generation of a characteristic pose—a sort of molecular mug shot of a molecule—useful for identifying similar molecular structures. In this regard, it has a lot in common with topomer technology utilised in ChemSpace™ [20–22].

The data set can then be subjected to soft independent modelling of class analogy (SIMCA [23]), which uses principal components analysis to identify patterns of field intensity shared by compounds within a property class. SIMCA is a classification technique that minimises assumptions about the linearity of relationships between descriptors within and between classes. This is appropriate for DM-PK/Tox applications, where the data are generally categorical in nature. More importantly, decision points tend to be categorical: exactly how insoluble or toxic or orally bioavailable a candidate is not critical. That a compound is soluble or non-toxic or orally bioavailable enough is what matters.

We have successfully applied SIMCA with IFO-CoMFA (IFO-SIMCA) to literature data sets for human intestinal absorption, blood–brain barrier permeability, and bioavailability, with considerable success in terms of both model fitting and predictivity [19]. Here we examine the usefulness of the method for obtaining global toxicity models based on in vitro hepatotoxicity data collected for a large set of structurally diverse pharmacologically active agents specifically for that purpose.

We also pursue an alternative approach: collecting data for a particular class of pharmacological target, then constructing a model for predicting the properties of other examples from that class. This is attractive for the kinds of low to medium throughput considered here. In this situation, the cells set an overall metabolic context for the particular chemistry involved, and the model discriminates between toxicity classes based only on idiosyncratic nuances of metabolism and toxicology that differentiate one particular compound from another.

[1] VolSurf is distributed by Molecular Discovery Ltd., Oxford, UK, and by Tripos Inc., St. Louis, MO.

[2] The GRID program was developed by Peter Goodford and is distributed by Molecular Discovery Ltd., Oxford, UK.

[3] CONCORD was developed by R.S. Pearlman, A. Rusinko, J.M. Skell, R. Balducci at the University of Texas, Austin, TX and is available exclusively from Tripos Inc., St. Louis, MO.

The results are of interest for their own sake, in that hepatotoxicity is a common toxicological problem in drug development, but also because the cell line we are using is derived from human tissue. In addition, the analyses serve to highlight the point that interpretation of the responses observed is not always as straightforward as one might wish.

2. Materials and methods

2.1. Data sets and structure generation

The Sigma-RBI Library of Pharmacologically Active Compounds (LOPAC) [24] was purchased from Sigma–Aldrich. The 640 compounds in the LOPAC fall into eight broad pharmacological classes: adenosines and purinergics; adrenergics and histaminergics; cholinergics and other ion channel modulators; dopaminergics; glutaminergics; signal transduction agents and opioids; enzyme inhibitors and GABAergics; and serotonergics. Twenty-one compounds were unresolved racemic mixtures that were treated as enantiomeric pairs, and assay data were missing for seven compounds. This left us with 654 molecules in the data set. The 26 NSAIDs tested were purchased from a variety of commercial sources. Chemical structures and chirality information were taken from the Sigma catalogue and the Merck Index [25] and converted into 3D structures on a Silicon Graphics Inc. workstation using the SYBYL sketcher[4] and CONCORD 4.0.[5]

Each structure was relaxed using the Tripos molecular mechanics force field. Atomic partial charges were calculated using the GAST_HUCK option in SYBYL, which invokes an extension of the method described by Gasteiger and Marsili [26] for estimating the distribution of charge over sigma bond networks.

2.2. Cell line and assays

The Amphioxus Technologies ACTIVTOX® human C3A hepatocyte cell line (ATTC #CRL-10741) used is a subclone of the HepG2 cell line. These cells retain most functions of adult hepatocytes, including the capacity for glucogenesis, albumin production and drug metabolism, as well as the capacity for cytochrome P450 induction [27]. This adult phenotype makes it a particularly attractive model for ADME screening. Four different assays were run, each in triplicate, with data collected from tests at 10 and 100 μM concentrations for each compound in LOPAC.

A homogenous colorimetric assay was used to estimate cell numbers for determination of *cell proliferation*. MTS [3-(4,5-dimethylthiazol-2-yl)-5-(3-carboxymethoxyphenyl)-2-(4-sulfophenyl)-2H-tetrazolium] is reduced to soluble formazan in the presence of the electron coupling reagent phenazine ethosulfate as a result of dehydrogenase activity in healthy, metabolically active cells. The assay is performed over a 12 h incubation period. The conversion of MTS to formazan results in an absorption of light at 490 nm that is proportional to the number of *healthy* cells.

Membrane integrity was determined by measuring the amount of lactate dehydrogenase (LDH) activity accessible from the medium. This is a homogenous fluorometric assay based on the reduction of exogenously supplied resazurine to the fluorescent resorufin by cytosolic lactate dehydrogenase. In healthy cells the plasma membrane is impermeable to LDH, but damage leads to leakage. Hence the amount of fluorescence measured is proportional to the number of *unhealthy* cells.

Intracellular ATP levels are also a useful measure of cell vitality. These were determined using a homogenous bioluminescent assay entailing addition of luciferase and D-luciferin to cell lysates. The intensity of the luminescence obtained is proportional to the amount of ATP present and, therefore, to the number of *healthy* cells.

Finally, a fluorometric assay was used to measure two key *caspase* levels—caspases 3 and 7—as indicators of the number of cells going into apoptosis. The combined levels of the two enzymes were measured in permeabilised cells by following the release of fluorescent rhodamine from a tetrapeptidyl conjugate; the Apo-ONE™ assay from Promega Corporation was used for this purpose.[6] The caspase 3/7 levels found are proportional to the number of *unhealthy* cells present.

Table 1 shows the mean, standard deviation and range of values found for each assay across all compounds in LOPAC, along with values for untreated controls and the average standard error of determination for each assay. The majority of compounds did not exhibit significant toxicity in individual assays, so the population means did not differ significantly from the control values.

Raw data for all experiments reported here are available in electronic form from the authors.

2.3. IFO-CoMFA and SIMCA analyses

Steric idiotropic field orientation was carried out using the Selector™ module in SYBYL, whereas electrostatic IFO was done using a modified version of SYBYL in which each atom in the connection table fed into the ORIENT BEST_VIEW command was weighted by the absolute value of its partial atomic charge (the atoms are unweighted for steric IFO [18,19]). Fields were then generated using the de-

[4] SYBYL® is distributed by Tripos Inc., 1699 South Hanley Road, St. Louis, MO 63144, USA. The version used for the work described here was the 6.9 release. http://www.tripos.com.

[5] CONCORD® was developed by R.S. Pearlman, A. Rusinko, J.M. Skell and R. Balducci at the University of Texas, Austin, TX and is available exclusively from Tripos Inc., 1699 South Hanley Road, St. Louis, MO 63144, USA. http://www.tripos.com.

[6] Promega Corp., 2800 Woods Hollow Road, Madison, WI 53711, USA. http://www.promega.com.

Table 1
Distribution of assay values across the LOPAC data set ($N = 654$)

Assay[a]	Proliferation (absorbance)		LDH (fluorescence)		Caspases (fluorescence)		ATP (luminescence)	
	10 μM	100 μM	10 μM	100 μM	10 μM	100 μM	10 μM	100 μM
Mean	0.68	0.43	0.75	1.41	911	3232	13349	14655
S.D.	0.09	0.12	0.13	0.47	212	121	1330	1972
Minimum	0.09	0.06	0.57	0.86	554	331	615	135
Maximum	0.97	0.88	2.85	3.04	6739	1.1×10^5	17595	26336
Control ± S.E.[b]	0.67 ± 0.03	0.43 ± 0.01	0.77 ± 0.02	1.25 ± 0.05	850 ± 70	1300 ± 96[c]	13000 ± 954	15000 ± 1365

[a] Fluorescence and luminescence in arbitrary units.
[b] Root mean square error of the mean calculated across all triplicate determinations, unless otherwise noted.
[c] RMSE of the mean calculated for triplicate determinations below 2000.

fault settings from the SYBYL QSAR module—i.e., a 2 Å rectilinear lattice extending 5 Å beyond any atoms in any molecule, and an sp^3 carbon atom bearing a unit positive charge as probe for calculating steric and electrostatic interaction potentials at each lattice point. Electrostatics were suppressed at lattice points falling within the steric envelope of each molecule, and both steric and electrostatic fields were included in the final model. CoMSIA fields were examined but including them failed to improve any of the models described here appreciably (data not shown).

The various responses indicative of cytotoxicity for the LOPAC data set were clustered manually for each assay based on the variances found for each of the two test compound concentrations. Observations falling within 1 standard deviation (S.D.) of the mean response at both concentrations formed one cluster,[7] with observations falling progressively farther from the mean binned at intervals of 1 S.D. Compounds that shared bins for responses at both concentrations were then grouped together. The initial clustering obtained was then modified as appropriate after visual inspection—e.g., by merging proximal pairs of clusters when each alone was too sparsely populated to be useful.

Classification analyses based on these clusters were run using SIMCA as implemented in SYBYL 6.9, with the quality of each model determined using the QSAR ANALYSIS PREDICT command. The statistical quality of each model is reported in terms of the number of correct predictions within each class and in terms of overall accuracy. The latter statistic can be misleading, however, when the distribution of observations among classes is skewed, as it is here and as is generally the case. If 90% of the compounds in a data set are non-toxic, for example, a model that simply classes everything as non-toxic will not be very useful but will nonetheless have an overall accuracy of 90%. Therefore the average accuracy *across* classes—i.e., the average of the accuracies found within each class—is also reported. In the non-discriminating case cited above, this statistic will be the average of 100% for the non-toxics and 0% for the toxics, yielding an average across classes of 50%.

Each set of responses was clustered separately. The number of clusters identified differed between assays, and the cluster IDs assigned are at best partially ordered. Hence the compounds making up cluster 2 in the ATP level results are not, in general, the same as those making up cluster 2 in the proliferation assay.

3. Results

3.1. LOPAC results for ATP levels

Of the four cytotoxicity indicators examined, the cultured hepatocyte ATP content proved least informative and, perhaps not surprisingly, yielded the weakest IFO-SIMCA model. The ATP levels found fell into two clear classes, mostly based on results obtained at 100 μM (Fig. 1A). Only tyrphostin A9 showed a clearly deleterious effect at 10 μM, whereas 47 depressed ATP levels significantly when tested at the higher concentration alone. Accuracy of the model obtained with respect to the toxic compounds was reasonably high (32 out of 47 were correctly classified, i.e., 68%). The accuracy of classification is illustrated graphically in Fig. 1B, where the symbol for points is based on the toxicity class to which the corresponding compound was assigned by the IFO-SIMCA model.

The high positives rate (12%) evident in Fig. 1B probably reflects, at least in part, the rather conservative toxicity criterion used here (see Footnote 7). The fact that the toxic tyrphostin A9 was categorised correctly by the model, even though the magnitude of its effect was not included, argues against this as the sole explanation.

3.2. LOPAC results for caspase induction

A high value in the caspase assay indicates that a relatively large number of unhealthy cells are present. Hence any compound that was cytotoxic at both tested concentrations would appear in the upper right-hand corner of the plots shown in Fig. 2A. None do, but two (sepiapterin and W-7) do appear at the lower right (cluster 8). These are anoma-

[7] Note that the bins are based on S.D. found across all treated cells; this is necessarily larger than the assay standard error (S.E.) due to the skewness of the populations, which inflates the S.D. disproportionately. Examination of the actual distributions of responses obtained shows that taking one S.D. from the mean is not unduly conservative given the high cost of toxicological false negatives.

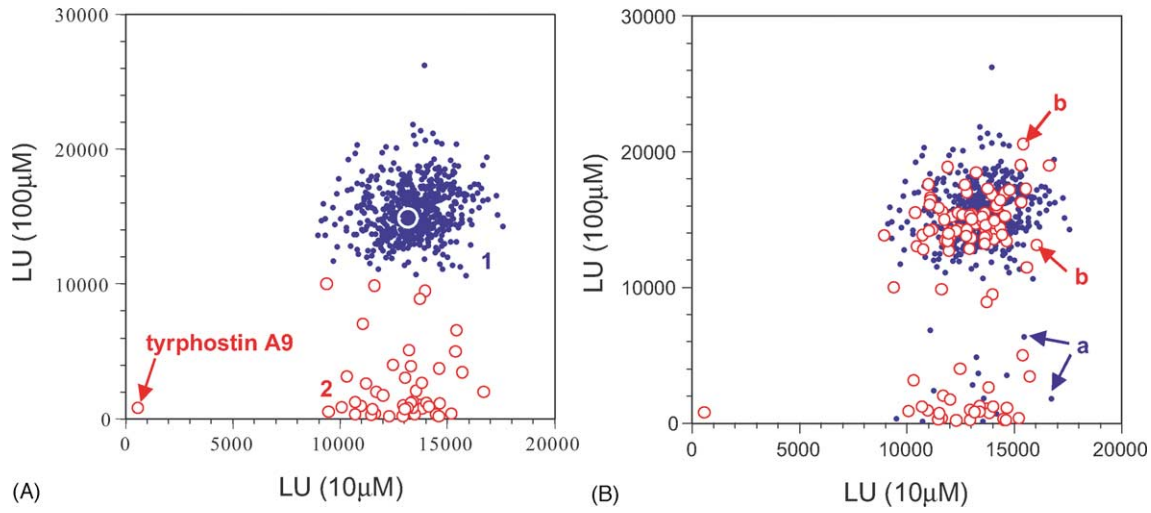

Fig. 1. ATP level in HepG2/C3A cells after treatment with compounds from the LOPAC data set at a concentration of 100 μM vs. the level found after treatment at 10 μM. LU: luminescence units. Compounds were classified as non-toxic (blue) or as toxic (red). (A) Original division into two toxicity classes based on observed ATP levels. The white circle in cluster 1 indicates the ATP levels found in untreated control cells. (B) Derived classifications from the SIMCA model generated using steric and electrostatic fields based on steric IFO alignments. Examples of false negative and false positive assignments are indicated by 'a' and 'b', respectively.

lous, in that they appear to be toxic at 10 μM but not at 100 μM.

The average classification accuracy across all eight toxicity clusters was 83% (Table 2) versus the 12.5% correct expected for assignment at random. Moreover, most misclassified compounds were assigned by the model to a toxicity class qualitatively similar to that into which is was originally placed based on the assay data itself (Fig. 2B versus Fig. 2A). If the compounds in clusters 1–4 are all considered non-toxic and those in clusters 5–8 are taken as being toxic, the respective accuracies are 99 and 86%. IFO-SIMCA failed to produce good models based on these broader categorisations, however, probably because the structural variation within each class was too great.

3.3. LOPAC results for LDH release

The distribution of values obtained for the LDH assay is shown in Fig. 3, with the toxicity classifications used to build the model indicated in Fig. 3A and the assignments generated from the IFO-SIMCA model indicated in Fig. 3B.

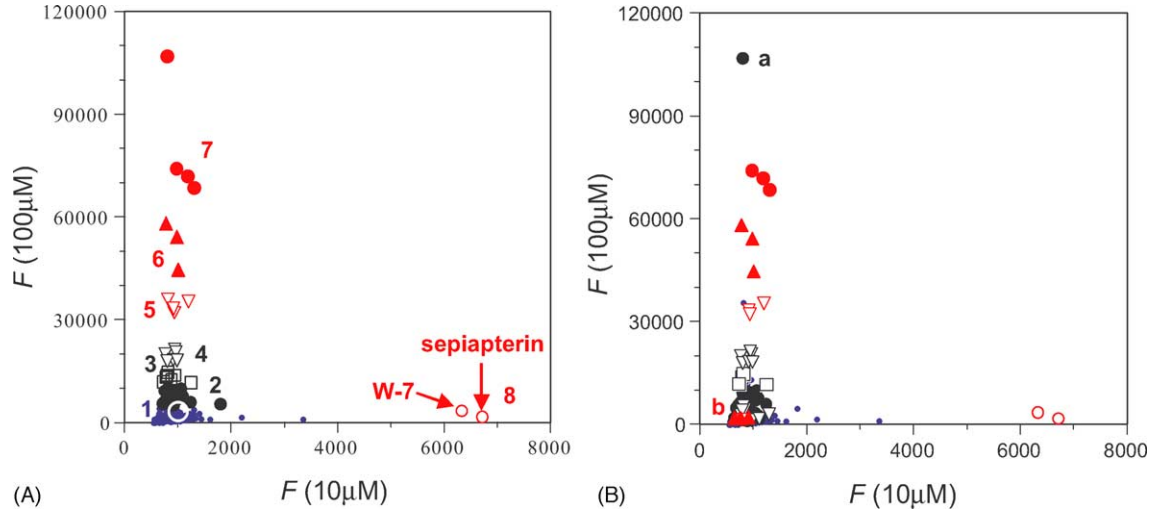

Fig. 2. Dependence of the combined caspase 3 and 7 levels after treatment with compounds from the LOPAC data set at 100 μM on levels found after treatment at 10 μM. F: fluorescence intensity (arbitrary units). Blue points represent non-toxic compounds (class 1), black symbols indicate mildly toxic compounds (classes 2–4), and red symbols indicate toxic compounds (classes 5–8). (A) Original division into eight toxicity classes based on observed caspase levels. The white circle indicates the levels found in untreated control cells. (B) Toxicity class assignments in the SIMCA model generated using steric and electrostatic fields based on steric IFO alignments. Examples of false negative and false positive assignments are indicated by 'a' and 'b', respectively.

Table 2
Caspase model results for the LOPAC data set

Toxicity class	n[a]	Number correct	Percentage correct
1	586	514	88
2	40	29	73
3	8	4	50
4	6	6	100
5	5	4	80
6	3	3	100
7	4	3	75
8	2	2	100
Total	654	565	86 (83%)[b]

[a] Number of compounds in each class.
[b] Average percentage correct across classes.

In this assay, low values are indicative of healthy, intact cells, whereas high values indicate that cells have lysed or have become leaky. The point at the upper right-hand corner of the plot corresponds to tyrphostin A9, in this case the only LOPAC compound exhibiting high toxicity at both concentrations tested.

Detailed statistics of fit for these data are given in Table 3. By this in vitro hepatotoxicity criterion, classification of toxics (classes 3–5) exhibited an accuracy of 70%, whereas non-toxics (classes 1 and 2) were correctly categorised 85% of the time. These class-wise accuracies for LDH are considerably lower than are the corresponding values for classification according to caspase induction (Table 2). The classes themselves are quite a bit larger than for caspase induction, however, which sharply reduces the tendency to over-fit the data. Then, too, caspase induction measures an inherently more specific mode of cytotoxicity.

Table 3
Classification accuracy for the LDH assays on the LOPAC data set using steric IFO for alignment and both steric and electrostatic fields as descriptors

Toxicity class	n[a]	Number correct	Percentage correct
1	467	317	68
2	89	58	65
3	35	25	71
4	55	35	64
5	8	6	75
Total	654	441	69 (69%)[b]

[a] Number of compounds in each class.
[b] Average percentage correct across classes.

3.4. LOPAC results for cell proliferation

The most complex results were obtained from measurements of cell proliferation (Fig. 4), but these were also the most informative. Some compounds (clusters 2 and 3) actually enhanced proliferation at 100 μM, albeit to a modest degree. Others were inhibitory at both test concentrations (clusters 7 and 8), and others only showed clear signs of growth inhibition at one concentration (clusters 4 and 6). Compounds that inhibited proliferation significantly at 10 μM but not at 100 μM constituted class 5. Compounds in this toxicity class were all correctly assigned by the IFO-SIMCA models. This class accounted for a large number of the false positives generated by the model (indicated by the many red triangles scattered amongst the blue points in Fig. 4B), which suggests that the cytotoxicity seen for these compounds at 10 μM may represent an experimental artifact of some kind.

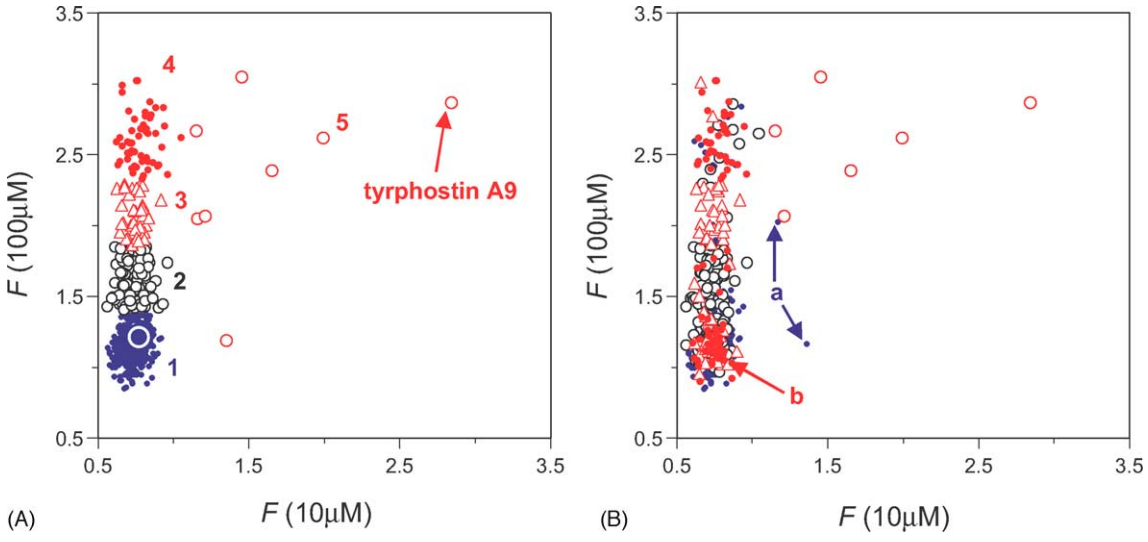

Fig. 3. Distribution of LDH assay values for the LOPAC data set. Resorufin fluorescence F after treatment at 100 μM is shown as a function of fluorescence levels found after treatment at 10 μM. Blue points represent non-toxic compounds (class 1), black circles indicate mildly toxic compounds (class 2), and red symbols indicate toxic compounds (classes 3–5). (A) Original division into six toxicity classes based on observed release of LDH. The white circle indicates the extent of LDH release found for untreated control cells. (B) Derived classifications in the SIMCA model generated using steric and electrostatic fields based on steric IFO alignments. Examples of false negative and false positive assignments are indicated by 'a' and 'b', respectively.

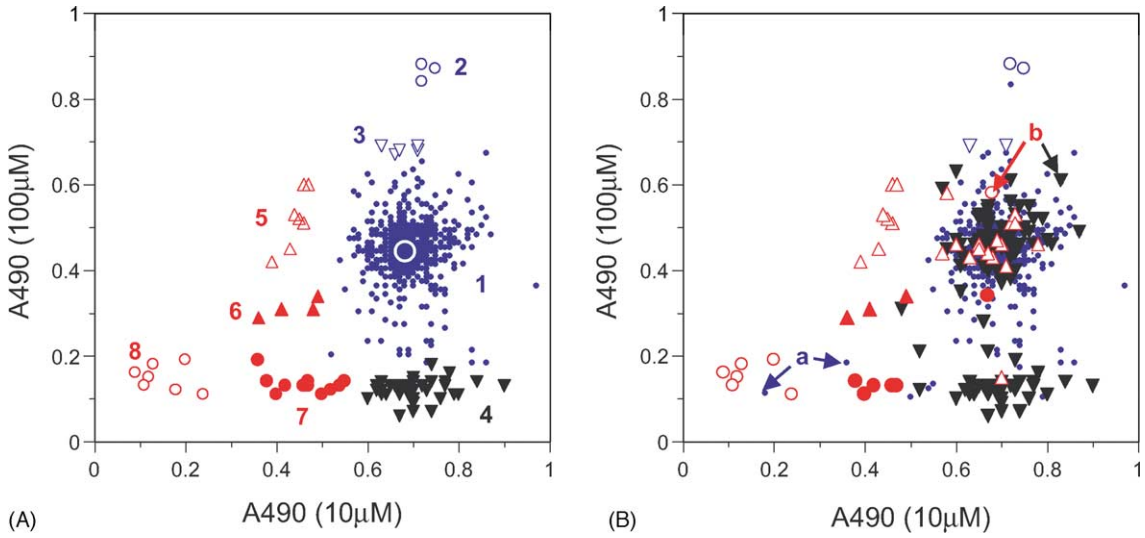

Fig. 4. Dependence of cell proliferation for the LOPAC data set after treatment at 100 μM as a function of that found after treatment at 10 μM. Blue symbols represent non-toxic compounds (classes 1–3), black symbols indicate mildly toxic compounds (class 4) and red symbols indicate toxic compounds (classes 5–8). (A) Original division into eight toxicity classes based on effects on proliferation. The white circle indicates data from untreated control cells. (B) Derived classifications from the SIMCA model generated using steric and electrostatic fields based on steric IFO alignments. Examples of false negative and false positive assignments are indicated by 'a' and 'b', respectively.

Table 4 presents quantitative performance statistics for the IFO-SIMCA model shown graphically in Fig. 4 for both steric and electrostatic orientation. This assay was the only one in which electrostatic IFO outperformed steric alignment in terms of overall accuracy. The biggest gain in accuracy was for toxicity class 7 (91% versus 45%; Table 4). Accuracy for classes 2 and 3 improved slightly, whereas toxicity classes 5 and 8 were somewhat less accurately assigned.

The right-hand columns in Table 4 show in detail how compounds incorrectly classified by the models were distributed across the other toxicity classes. For both models, non-toxic compounds from cluster 1 that were incorrectly assigned to class 4 accounted for most of the false positives seen. In the case of steric alignment, most of the balance came from toxicity class 5 (mildly toxic). For electrostatic alignment, on the other hand, the balance involved mistaken assignment to class 7 (toxics). The increase in sensitivity for class 7 compounds, then, comes at the cost of a greater than 20-fold degradation in selectivity.

Despite the large size of the full data set, most of the individual clusters of cytotoxic compounds were rather sparsely populated, making any sweeping cross-validation impractical. Instead, 10% of the compounds were withdrawn from each of the three largest classes (1, 4 and 7) to serve as a test set. The reduced steric IFO-SIMCA model generated from the remaining 589 compounds was similar in quality to the model for the complete data set (Table 4), with a classification rate of 82% overall and 78% when averaged across classes. Predictive performance on the 65 compounds in the test set was only slightly worse, at 79% correct overall. Most tellingly, however, the value obtained by averaging across classes was only 61%. The discrepancy reflects the high false negative rate—60%—for the five toxic compounds in the test set, which serves to underscore the importance of con-

Table 4
Classification distributions for SIMCA models of cell proliferation based on steric (S) and electrostatic (E) IFO alignment

Toxicity class	*n*	Number correct		Percentage correct		Number erroneously assigned to class															
						1		2		3		4		5		6		7		8	
		S	E	S	E	S	E	S	E	S	E	S	E	S	E	S	E	S	E	S	E
1	577	482	480	84	83	–	–	0	0	0	6	78	64	15	2	0	0	1	23	1	2
2	3	2	3	67	100	1	0	–	–	0	0	0	0	0	0	0	0	0	0	0	0
3	5	2	4	40	80	3	0	0	0	–	–	0	0	0	0	0	0	0	0	0	0
4	40	36	36	90	90	3	0	0	0	0	0	–	–	1	0	0	0	0	0	0	0
5	7	7	6	100	86	0	1	0	0	0	0	0	0	–	–	0	0	0	0	0	0
6	4	3	4	75	100	1	0	0	0	0	0	1	0	0	0	–	–	0	0	0	0
7	11	5	10	45	91	5	0	0	0	0	0	1	0	0	0	0	0	–	–	0	0
8	7	6	5	86	71	1	1	0	0	0	0	0	0	0	0	0	0	0	1	–	–
Total	654	543	548	83 (73)[a]	84 (88)[a]																

[a] Average percentage correct across classes.

sidering selectivity across classes in data sets with skewed distributions.

3.5. Consolidated results

As may be clear from the above discussions for each individual set of assay results, no consensus was readily apparent as to which compounds are hepatotoxic. Tyrphostin A9, for example, exhibited toxicity in the ATP and LDH assays but had no significant effect on proliferation. There is, in fact, very little direct correlation among the assay results (data not shown). Given that the other criteria are compromised to a greater or lesser extent when cells fail to proliferate, we tried normalising ATP, caspase, and LDH levels to the extent of proliferation. Doing so did not increase concordances between the different measures.

Nor was simple conversion to ranks with summation across assays or across some subset of assays effective. Some variation of this approach would be desirable, since no single assay produced clusters big enough to support full cross-validation studies. As an alternative approach, we applied a variation of a hierarchical ranking technique originally developed for non-parametric analysis of titration data from herbicide screening [28,29]. Each set of assay results was converted to a decile scale and sorted using a particular assay as primary toxicity criterion. Ties with respect to the primary assay criterion were then broken by reference to a second criterion. Ties with respect to both the primary and secondary criteria were resolved where possible by reference to a third criterion, and so forth. Only results from tests carried out at the higher concentration were used, and the assay priorities were set primarily on the basis of resolution: proliferation, LDH, ATP and caspase level, in order of descending priority. Clustering and subsequent SIMCA analysis gave the results shown in Table 5.

These are not very promising statistics, and cross-validation experiments suggest that they may be overly optimistic. When a diverse representative test set of 100 compounds was chosen using optimisable *k*-dissimilarity (OptiSim [30,31]) selection and set aside, the reduced model based on the remaining 554 compounds performed slightly better than the full model did, with an overall accuracy of 57%. Accuracy for categorisation of compounds in the test set was only about 20%, however, whether calculated overall or across classes. This is the same as the 20% accuracy expected by chance.

Table 5
Global hepatotoxicity model based on hierarchical ranking

Ranks	Class	*n*	Number correct	Percentage correct	Number correct ± one class
515–654	1	140	97	69	101 (72%)
412–514	2	111	41	37	90 (81%)
237–390	3	167	86	51	108 (65%)
148–234	4	93	51	55	68 (73%)
1–142	5	143	65	45	86 (60%)
Total		654	340	52 (52%)[a]	453 (69%)

The order of priorities used was proliferation > LDH > ATP > caspase level, with a rank of 1 being the most cytotoxic. Steric IFO fields were used as descriptors.

[a] Average percentage correct across classes.

In this case, the response classes are explicitly ordered, so one can ask how many are almost correct—i.e., how many are assigned by the model to a category to one side or the other of the one where it was originally found [32]. When that is done, one obtains the results shown in the last column of Table 5. This approach needs to be taken with great caution for skewed data, but here classes 1, 3 and 5 are the largest toxicity classes, which reduces the risk of distortion considerably.

The top-ranked tested compound by this procedure is thioridazine, which is a known hepatotoxin. Terfenidine ranks fourth on the list and was withdrawn from the market in 1998, though not because of hepatotoxicity. Interestingly, propranolol ranks second in terms of the in vitro hepatotoxicity indicators considered here, but is predicted to not be hepatotoxic by both the steric and electrostatic IFO-SIMCA models obtained from that data. Propranolol is not, in fact, known to exhibit hepatotoxicity in humans.

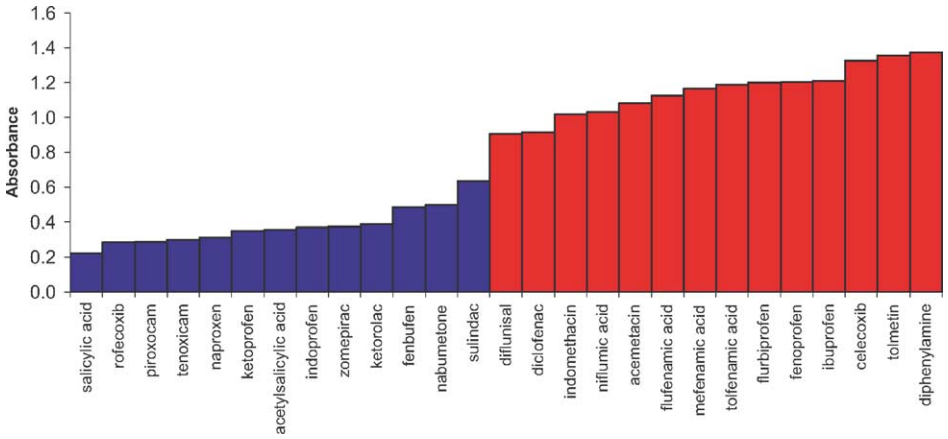

Fig. 5. Distribution of LDH assay results for the NSAID data set ($N = 27$). Blue indicates compounds classified as non-toxic for purposes of further analysis. Red indicates compounds assigned to the toxic class.

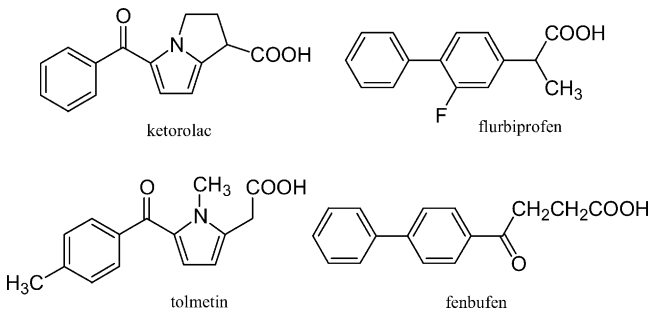

Fig. 6. Structures of misclassified NSAIDs and some structural analogues included in the NSAID data set. Ketorolac is toxic in the LDH assay but is categorised as toxic by the IFO-SIMCA model. Flurbiprofen is classed as toxic by assay but is categorised as non-toxic by the model. Tolmetin is toxic and fenbufen is non-toxic; both are correctly placed by the model.

3.6. NSAID results for LDH release

LDH assays were carried out for 27 NSAIDs applied to HepG2/C3A cells, generating the activity profile shown in Fig. 5. Based on this profile, 13 compounds were classified as non-toxic and 14 were classified as toxic. Applying IFO-SIMCA to these data produced a model that correctly assigned 11 of 13 compounds classified as non-toxic (92%) and 13 of 14 toxic compounds (93%). The single false negative was flurbiprofen and the single false positive was ketorolac (Fig. 5). Of these two, ketorolac is a borderline case in terms of its effect on the hepatocytes. It is no coincidence that the closest analogue to ketorolac is the toxic tolmetin, and that the closest analogue to flurbiprofen is the (borderline) non-toxic fenbufen (Fig. 6).

Even though the NSAID data set is small, the performance of the full model was good enough to justify attempting cross-validation. Results of carrying out such an experiment for a test set of six compounds are shown in Table 6. Not surprisingly, the accuracy of the reduced model is less than that of the full model (71 versus 92%), but the accuracy in categorising the test is quite good—five of six (83%) correctly classified. It is interesting that the one failure in the test set is aspirin, which often comes up as an outlier in DM-PK/Tox analyses.

Were this a full-fledged development project, we would probably feel justified in using IFO-SIMCA to evaluate new compounds proposed for synthesis in this area, or to suggest which other compounds already in hand should be pushed forward for further efficacy and toxicological testing. That would presumably start an iterative process, in which some of those new compounds would themselves be screened for in vitro hepatotoxicity, thereby allowing the model to be refined further.

4. Discussion

CoMFA is best known for its use in QSAR, particularly in connection with linear regressions and PLS. There, as in IFO-SIMCA, the key value of using molecular interaction fields for similarity analysis is their combination of generality and specificity. The generality comes from the fact that interactions between molecules, whether ligand and protein or ligand and solvent, are mediated by molecular fields, whereas the specificity comes from characteristic anisotropies in the pattern of electron distributions in space.

The IFO-CoMFA fields of two molecules can only be similar if they have homologous 3D structures *and* similar field profiles about *each* principal axis. Using a rule-based 3D builder (CONCORD) to generate 3D structures is critical to ensuring that similarity of structure translates reliably into similarity of conformation. These considerations make the combination of IFO-CoMFA with SIMCA a promising extension of earlier similarity-based ADME modelling based on derived 2D descriptors [33,34] into the more generalised world of 3D molecular interaction fields.

Whatever the descriptor, no model can be better than the data upon which it is based. Published in silico DM-PK/Tox models have been hobbled by a lack of reliable published data for areas of chemistry relevant to drug discovery [35,36]. Most publicly available solubility data, for example, are for thermodynamic rather than kinetic solubilities and are dominated by unlikely drug candidates such as long-chain alcohols and waxes. Conversely, data sets used to train oral bioavailability models are usually (though not always [36]) composed entirely of commercial

Table 6
Cross-validation of the steric IFO-SIMCA model for LDH release after treatment with NSAIDs at 100 μM

Training set	Reduced model statistics		Test compound	Observed class	Predicted class
Class	*n*	Number correct (percentage)			
Non-toxic	10	6 (60%)	Acetylsalicylic acid	Non-toxic	Toxic
Toxic	11	9 (82%)	Piroxocam	Non-toxic	Non-toxic
			Fenbufen	Non-toxic	Non-toxic
			Indomethacin	Toxic	Toxic
			Celecobix	Non-toxic	Non-toxic
Total	21	15 (71%)	Tolfenamic acid	Toxic	Toxic

Six compounds were held back, and the remaining 18 compounds were used to construct a reduced model, yielding the classification statistics indicated on the left side of the table. That model was then used to make the test set predictions shown on the right.

drugs. The negative examples in such data sets are inherently anomalous—they all managed to became commercial drugs *despite* their poor bioavailability. It seems unlikely that such compounds are really representative of the wide range of structures that fall by the wayside during drug development.

We have addressed the problem of data quality by generating our own, internally consistent data rather than relying on literature compilations. Global models were constructed from data obtained for a diverse collection of biologically active compounds that includes but is not limited to drugs (in fact, the estimated log *P* and molecular weight profiles for the LOPAC data set are similar to those for marketed drugs (data not shown), though some of the nucleotides, nucleosides, and other compounds contained therein would violate the donor and acceptor limits imposed by Lipinski's Rule of 5 [11]). The generalised models obtained exhibited good sensitivity and specificity, but more data is needed if they are to be made adequately robust and predictive. The local model obtained for a single pharmacological class, in contrast, proved itself both accurate and predictive, at least within the class in question—i.e., NSAIDs.

Taken together, our results demonstrate that combining careful biochemistry with IFO-CoMFA makes it possible to generate predictive models for quite complex toxicological endpoints for the kinds of compounds typically encountered in drug discovery and development.

Acknowledgements

Martin Bohl (Tripos GmbH) has provided extensive moral and technical support for our DM-PK/Tox work over the last several years, as have David Patterson and Trevor Heritage (Tripos Inc.).

References

[1] H. van de Waterbeemd, E. Gifford, ADMET in silico modeling: towards prediction paradise? Nat. Rev. Drug Discov. 2 (2003) 192–204.
[2] T.R. Stouch, J.R. Kenyon, S.R. Johnson, X.-Q. Chen, A. Doweyko, Y. Li, In silico ADME/Tox: why models fail, J. Comput.-Aided Mol. Des. 17 (2003) 83–92.
[3] D.E. Clark, Rapid calculation of polar molecular surface area and its application to the prediction of transport phenomena. 1. Prediction of intestinal absorption, J. Pharm. Sci. 88 (1999) 807–814.
[4] G. Cruciani, P. Crivori, R.-A. Carrupt, B. Testa, Molecular fields in quantitative structure-permeation relationships: the VolSurf approach, J. Mol. Struct. (Theochem.) 503 (2000) 17–30.
[5] D.E. Clark, Rapid calculation of polar molecular surface area and its application to the prediction of transport phenomena. 2. Prediction of blood–brain barrier penetration, J. Pharm. Sci. 88 (1999) 815–821.
[6] P. Crivori, G. Cruciani, P.-A. Carrupt, B. Testa, Predicting blood–brain barrier permeation from three-dimensional molecular structure, J. Med. Chem. 43 (2000) 2204–2216.
[7] M.J. de Groot, M.J. Ackland, V.A. Horne, A.A. Alex, B.C. Jones, A novel approach to predicting P450 mediated drug metabolism. CYP2D6 catalyzed *N*-dealkylation reactions and qualitative metabolite predictions using a combined protein and pharmacophore model for CYP2D6, J. Med. Chem. 42 (1999) 4062–4070.
[8] S. Rao, R. Aoyama, M. Schrag, W.F. Trager, A. Rettie, J.P. Jones, A refined 3-dimensional QSAR of cytochrome P450 2C9: computational predictions of drug interactions, J. Med. Chem. 43 (2000) 2789–2796.
[9] R. Snyder, R. Sanger, J. Wang, S. Eakins, Three-dimensional quantitative structure activity relationship for CYP2D6 substrates, Quant. Struct.-Act. Relat. 21 (2002) 357–368.
[10] I. Zamora, I. Afzelius, G. Cruciani, Predicting drug metabolism: a site of metabolism prediction tool applied to the cytochrome P450, J. Med. Chem 46 (2003) 2313–2324.
[11] C.A. Lipinski, F. Lombardo, B.W. Dominy, P.J. Feeney, Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings, Adv. Drug Del. Rev. 23 (1997) 3–25.
[12] R.D. Cramer III, S.A. DePriest, D.E. Patterson, P. Hecht, The developing practice of comparative molecular field analysis, in: H. Kubinyi (Ed.), 3D QSAR in Drug Design, vol. 1, ESCOM, Leiden, 1993, pp. 443–485.
[13] Y.C. Martin, 3D QSAR: current state, scope, and limitations, in: H. Kubinyi, G. Folkers, Y.C. Martin (Eds.), 3D QSAR in Drug Design, vol. 3, Kluwer Academic Publishers/ESCOM, Dordrecht, 1998, pp. 3–23.
[14] C.J. Blankley, Recent developments in 3D-QSAR, in: H. van de Waterbeemd (Ed.), Structure–Property Correlations in Drug Research, Academic Press, Austin, 1996, pp. 111–177.
[15] R.D. Clark, J.M. Leonard, A. Strizhev, Pharmacophore models and comparative molecular field analysis (CoMFA), in: O. Güner (Ed.), Pharmacophore Perception, Development, and Use in Drug Design, International University Line, La Jolla, 2000, pp. 151–169.
[16] P.J. Goodford, A computational procedure for determining energetically favorable binding sites on biologically important molecules, J. Med. Chem. 28 (1985) 849–857.
[17] G. Klebe, U. Abraham, T. Mietzner, Molecular similarity indices in a comparative analysis (CoMSIA) of drug molecules to correlate and predict their biological activity, J. Med. Chem. 37 (1994) 4130–4146.
[18] R.D. Clark, A.M. Ferguson, R.D. Cramer, Bioisosterism and molecular diversity, in: H. Kubinyi, G. Folkers, Y.C. Martin (Eds.), 3D QSAR in Drug Design, vol. 2, Kluwer Academic Publishers/ESCOM, Dordrecht, 1998, pp. 211–224.
[19] P.R.N. Wolohan, R.D. Clark, Predicting drug pharmacokinetic properties using molecular interaction fields and SIMCA, J. Comput.-Aided Mol. Des. 17 (2003) 65–76.
[20] R.D. Cramer, D.E. Patterson, R.D. Clark, F. Soltanshahi, M.S. Lawless, Virtual compound libraries: a new approach to decision making in molecular discovery research, J. Chem. Inf. Comput. Sci. 38 (1998) 1010–1023.
[21] R.D. Cramer, M.A. Poss, M.A. Hermsmeir, T.J. Caulfield, M.C. Kowala, M.T. Valentine, Prospective identification of biologically active structures by topomer shape similarity searching, J. Med. Chem. 42 (2000) 3919–3933.
[22] R.D. Cramer, R.J. Jilek, K.M. Andrews, *dbtop*: topomer similarity searching of conventional databases, J. Mol. Graph. Model. 20 (2002) 447–462.
[23] S. Wold, M. Sjöström, Method for analyzing chemical data in terms of similarity and analogy, in: B.R. Kowalski (Ed.), Chemometrics: Theory and Applications, vol. 52, ACS Symposium Series, 1977, pp. 243–282.
[24] Sigma catalog number SC001. http://Sigma-Aldrich.com.
[25] The Merck Index, 12th Ed., Version 12:3, Chapman & Hall/CRCnetBASE Electronic Publishing Division, 2000 (CD-ROM).
[26] J. Gasteiger, M. Marsili, Iterative partial equalization of orbital electronegativity—a rapid access to atomic charges, Tetrahedron 36 (1980) 3219–3228.
[27] J.H. Kelly, N.L. Sussman, A fluorescent cell-based assay for cytochrome P-450 isozyme 1A2 induction and inhibition, J. Biomol. Screen. 5 (2000) 249–253.

[28] D.L. Duewer, R.D. Clark, Rank-order analysis for robust evaluation of multi-response, multi-block comparisons, J. Chemometrics 5 (1991) 503–521.
[29] R.D. Clark, J.J. Parlow, L.H. Brannigan, D.M. Schnur, D.L. Duewer, Applications of scaled rank-sum statistics in herbicide QSAR, in: C. Hansch, T. Fujita (Eds.), Classical and Three-Dimensional QSAR in Agrochemistry, vol. 606, ACS Symposium Series, American Chemical Society, Washington DC, 1995, pp. 264–281.
[30] R.D. Clark, OptiSim: an extended dissimilarity selection method for finding diverse representative subsets, J. Chem. Inf. Comput. Sci. 37 (1997) 1181–1188.
[31] R.D. Clark, US Patent 6,535,819 B1 (2003).
[32] F. Yoshida, J.G. Topliss, QSAR model for drug human oral bioavailability, J. Med. Chem. 43 (2000) 2575–2595.
[33] O.A. Raevsky, S.V. Trepalin, H.P. Trepalina, V.A. Gerasimenko, O.E. Raevskaja, SLIPPER-2001—software for predicting molecular properties on the basis of physicochemical descriptors and structural similarity, J. Chem. Inf. Comput. Sci. 42 (2002) 540–549.
[34] O.A. Raevsky, K.-J. Schaper, P. Artursson, J.W. McFarland, A novel approach for prediction of intestinal absorption of drugs in humans based on hydrogen bond descriptors and structural similarity, Quant. Struct.-Act. Relat. 20 (2002) 402–413.
[35] W.K. Sietsema, The absolute oral bioavailability of selected drugs, Int. J. Clin. Pharmacol. Ther. Toxicol. 27 (1989) 179–211.
[36] D.F. Veber, S.R. Johnson, H.-Y. Cheng, B.R. Smith, K.W. Ward, K.D. Kopple, Molecular properties that influence the oral bioavailability of drug candidates, J. Med. Chem. 45 (2002) 2615–2623.