Prediction of UV and ESI−MS Signal Intensities

Harald Mauser,* Olivier Roche, Martin Stahl, and Stephan Müller
F. Hoffmann-La Roche Ltd., Pharmaceuticals Division, CH-4070 Basel, Switzerland

Received November 17, 2004

All major pharmaceutical companies maintain large collections of compounds that are used either for screening against biological targets or as synthetic precursors. The quality assessment of these compounds is typically done by liquid chromatography combined with mass spectroscopy (LC/MS) and UV purity control. To facilitate the analysis of the analytical data, we have built computational models to predict UV and MS signal intensities under experimental LC/MS conditions. The discriminant partial-least-squares technique was used for classifying compounds into those most likely to yield a MS signal and others where the signal is below the detection limit (94% and 88% correct predictions, respectively). In the case of UV prediction, we compared this statistical linear-regression technique to a knowledge-based approach. A combination of both techniques proved to be the most reliable (96/98% correct predictions of UV-active/ UV-inactive compounds). Both models have been incorporated into the automated compound integrity profiling at F. Hoffmann-La Roche.

INTRODUCTION

High throughput screening allows testing of a large number of compounds for biological activity.[1] Over the past decade, this discipline has become a typical starting point for drug discovery programs. As a consequence, all major pharmaceutical companies have been seeking to extend their compound collections, to take full advantage of the steadily growing screening capacity. It soon became obvious that not only the number of compounds but also the distribution of properties[2−7] and, perhaps even more importantly, the integrity of compounds are inevitable requirements for success in lead finding.[8]

Even with high quality control standards for compound synthesis and acquisition, chemicals may degrade over time or they may contain impurities. This is a common source for false positive results obtained when screening compounds in biological assays. Hence, the integrity of the compounds is fundamental for the analysis of the screening data. A combination of liquid chromatography and mass spectrometry (LC/MS) together with UV spectroscopy is the most frequently used technique for purity assessment.[9] The presence of the expected MS signal confirms the identity, whereas the relative amount of impurities can be derived from UV signal intensities. Often, the final purity assessment requires manual analysis of spectroscopic data based on the molecular structure. This detailed analysis can be the time-limiting step in quality control, hampering a higher throughput. A significant speedup would be achieved if one could distinguish between the compounds where the purity assignment is straightforward (automatic assessment) and the difficult cases where the experimental results have to be evaluated manually. Although mass spectrometry and UV detection are well investigated, there was no “in silico” tool available that could be used for this classification. Surprisingly, only a few groups have been working on fast prediction of UV absorbance,[10,11] and we are not aware of any similar tool for predicting MS intensities. We set out to develop tailored, fast in silico filters for assessing ionizabilities as a prerequisite for signals under ESI−MS conditions and for predicting UV intensities of compounds to assist the analytical compound purity assessment (Figure 1a).

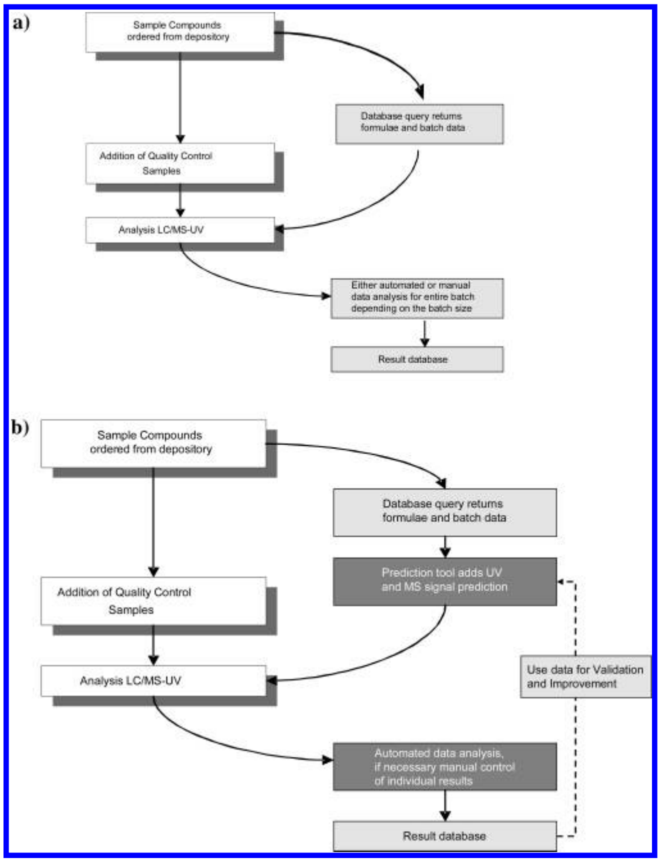

a)
Sample Compounds ordered from depository
Database query returns formulae and batch data
Addition of Quality Control Samples
Analysis LC/MS-UV
Either automated or manual data analysis for entire batch depending on the batch size
Result database
b)
Sample Compounds ordered from depository
Database query returns formulae and batch data
Prediction tool adds UV and MS signal prediction
Addition of Quality Control Samples
Analysis LC/MS-UV
Use data for Validation and Improvement
Automated data analysis, if necessary manual control of individual results
Result database

Figure 1. Workflow of the purity analysis of DMSO samples from the compound depository (a) before and (b) after the prediction tool was integrated.

MATERIALS AND METHODS

Samples. Samples were obtained from the Roche SMART depository, a fully automated inventory system, delivering the samples in 2 μL portions of 5 mM solutions in DMSO.

* Corresponding author phone: +41-(0)61 688 86 04; e-mail: harald.mauser@roche.com.

10.1021/ci0496548 CCC: $30.25
Published on Web 04/27/2005

Data Acquisition. The samples were analyzed on Applied Biosystems API-150 LC/MS instruments fitted with Agilent 1100 UV wavelength range detectors. The MS was operated in positive electrospray mode ("turbo-ionspray"), and the UV chromatogram was acquired in two ranges, $A = 230-300$ nm and $B = 210-230$ nm. A generic fast gradient was used (acquisition time, 4 min; run time, 6 min). The data files were then treated with in-house scripts (Visual Basic or AppleScript). Eventually, the analytical data was stored in a Microsoft Access database. More analytical details will be published in a forthcoming paper.

Data Normalization. The detector response (peak area) of each sample was first corrected for differences in injection volume on the basis of the DMSO peak area in UV-A. Then, a relative response factor was calculated by comparing the compound's signal to the average signal obtained from the quality control samples (diazepam), which were measured in the same series.

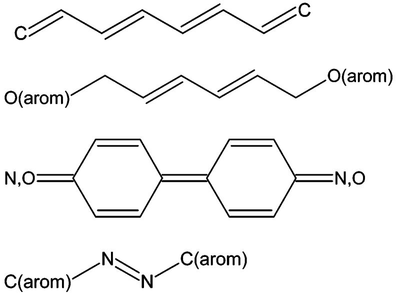
Descriptor Generation. Using the program Corina,[12] all molecules in the datasets were neutralized and desalted; in addition, small complexing ligands were removed. Over 1000 descriptors covering the range from 1D to 3D descriptors were calculated for each molecule. We decided not to include quantum mechanical descriptors (compare ref 11) or complex surface descriptors, as we were aiming toward a fast prediction tool that should be accessible as a web application. For variable selection, we used self-organizing maps[13] (SOMs) to visualize the data representation by various combinations of these descriptors. This analysis was complemented by linear principal component analysis (PCA) to determine possible outliers and data clustering.[14,15] A combination of Ghose−Crippen[16,17] parameters (referred to as ALOGP parameters) with our in-house set of 63 additional topological, electronic, count, and structural descriptors (for a detailed description, see ref 3) was found to produce the best results for both datasets.

Statistical Methods. We constructed discriminant partial-least-squares (PLS) models[15,18] for both UV and positive MS signal predictions using the SIMCA-P+ 9 software.[19] To obtain a balanced training set, 200 compounds (50% positive) were chosen for each model. The number of compounds was based on the number of diverse compounds in the negative sets. The complementary sets of UV-positive or MS-positive compounds were obtained by diverse selection. We employed the Kennard−Stone algorithm[3,20] on the PCA scores (five components) calculated for the corresponding descriptor matrices. The remaining compounds were used as a test set. As the response (*Y* variable), we used the memberships of compounds belonging to either the class of UV/MS-active compounds or the class of UV/MS-inactive compounds. As r^2 and q^2 are less suitable for measuring the quality of binary classification models, we used the Matthews correlation coefficient[21] (eq 1) instead.

$$\mathrm{cc} = \frac{NP - OU}{\sqrt{(N+O)(N+U)(P+O)(P+U)}} \tag{1}$$

where P, N, O, and U are the number of true positive, true negative, false positive, and false negative predictions, respectively. The values for cc can range from −1 to +1. A correlation coefficient of 1 reflects a perfect classification.

Table 1. Selection of Substructural Features (Class 3) with Maximum Absorbance outside the Detectible UV Range Used in the Knowledge-Based UV-Prediction Model

Substructures	Description
	Extended π-systems

Substructure Analysis. In addition to the statistical prediction model, a knowledge-based approach was developed for the UV dataset identifying chromophoric substructures that are responsible for UV absorbance. The substructures themselves were compiled manually and complemented by common motifs extracted from the corresponding dataset. The Daylight SMARTS[22] language was used for defining respective substructures. If at least one substructure out of the list of SMARTS is present in the query molecule, the molecule is flagged. However, substructures cannot be used completely independent from one another. A particular substructure of a molecule may, in principle, be responsible for a positive response, whereas another substructural element could compensate this effect. This is taken into account by considering a second list of unfavorable substructures. If a compound is flagged by the first set and it contains one or more substructures from the second set, a positive response cannot be expected; therefore, the flag is removed.

The advantage of this algorithm is that multiple combinations are possible so that one molecule can be flagged several times in one run, depending on the presence or absence of these predefined substructures. For the prediction of UV signal intensities, we distinguish between more generic SMART representations that reflect a high probability that a given molecule is UV-active and specific substructures that are known to be chromophores (see below). In the second step, these flags are converted into discrete numbers or scores to facilitate the interpretation and to allow a combination of the results of both predictions. Our knowledge-based model for UV prediction gives zero for the absence of an UV signal (intensity below detection limit), corresponding to no flag. The prediction result is set to be "1" if the compound belongs to the generic class of chromophores and to be "2" if a specific chromophore is detected. In addition, we add a warning flag if the signal is predicted to be out of the detectable range. Very large π systems tend to be fluorescent or, in the case of large conjugated alkenes, absorb in the visible wavelength range rather than in the UV. Both effects disturb the measurements by their influence on the reference wavelengths. Thus, the prediction score for these compounds is set to be "−1" (see Table 1).

Validation. To validate our prediction models, we selected an independent set of compounds from the Roche SMART depository, measured the corresponding MS and UV signals under conditions of our standard purity control, and compared the result to the prediction. For selecting the samples, we partitioned the entire library of available compounds into three classes according to their predicted properties: (1) MS and UV signals are expected; (2) no MS signal is expected, but UV-active signal is expected; and (3) MS signal is expected, but no UV signal is expected. From these sets, only compounds that are amenable to the LC/MS analysis could be considered, defined by a molecular weight (MW) between 150 and 800 (true for 98% of all compounds). We selected diverse subsets out of each individual class: 50% belonging to Class 1, 40% to Class 2, and 10% to Class 3. This ratio was chosen because it reflects the composition of our in-house compound collection and was, therefore, considered to be closest to the realistic conditions. In total, 732 compounds were measured.

RESULTS AND DISCUSSION

Data Compilation. All samples were taken from our in-house compound collection. As small molecules and synthetic intermediates are also part of this collection, we did not distinguish between druglike and nondruglike molecules for this study. For example, ca. 20% of the UV dataset (see below) would be rejected by our drug-likeness filters. Diverse and balanced sets would be ideal as a basis for every prediction model. Our first priority was to achieve a high quality of the data, and therefore, we considered only pure compounds that could be clearly identified. Therefore, the entire datasets are biased toward compounds showing both an absorbance in the UV-A range and the corresponding mass signal. The normalized experimental results were transformed into binary data, from which two individual datasets for predicting UV and positive MS signal intensities were extracted. On the basis of the distributions of intensities, we selected compounds having strong UV or MS signals for the respective positive datasets. In contrast, compounds in the respective negative datasets did not show an UV or MS signal. To reduce noise, compounds only showing weak signals were omitted in both datasets.

For the MS data, however, we could not consider compounds that were not showing any signals, neither in the MS nor in the UV, as we could not confirm their identities.

We compiled a dataset of 1248 compounds for the UV prediction; hereof, 964 showed a strong UV signal ("UV-active"), whereas 284 lacked a characteristic UV signal ("UV-inactive").

As a basis for the MS prediction, we compiled a dataset of 338 compounds showing a clear MS signal ("MS active") and 251 compounds not detectable with a positive MS ("MS inactive"). A comparison of both sets is given in Figure 2, showing the distribution over the MW of UV/MS-active and UV/MS-inactive compounds. Interestingly, there seems to be a tendency of UV-active compounds toward lower MWs (see discussion of the most important variables for the PLS model below).

As for the MS-inactive compounds, which required a well-defined UV signal to confirm identity, there is a bias toward MS-inactive compounds containing chromophores. This is reflected in Figure 2b by the shift toward compounds with lower MWs. We do not consider this as problematic, as only very few molecules in our in-house database would fall into the category of being UV- and MS-negative. To ensure that this assumption is correct and to test the reliability of the prediction models, both UV and MS models were subjected to an independent validation.

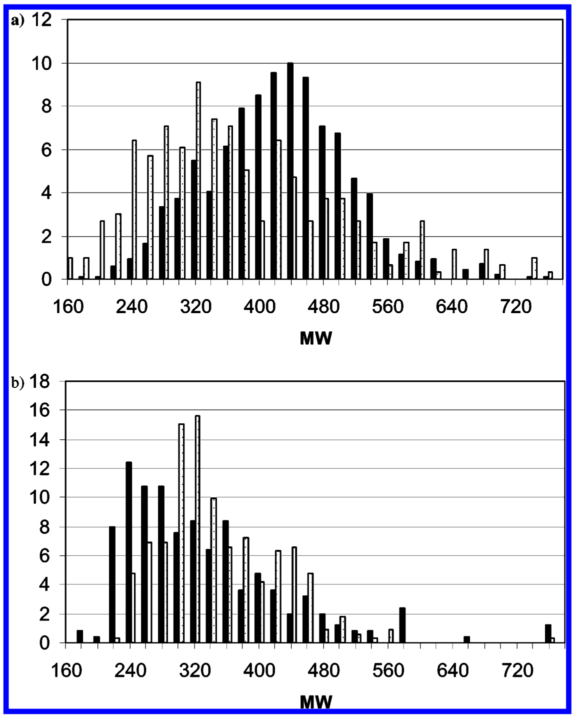


Figure 2. Histograms of the MW of the entire datasets compiled for predicting (a) UV (1248 compounds) and (b) MS (589 compounds) signal intensities. As both data sets are unbalanced, the frequencies of the active and inactive compounds were normalized to 100 in both histograms. Black bars refer to UV/MS-active compounds and white bars to UV/MS-inactive compounds, respectively.

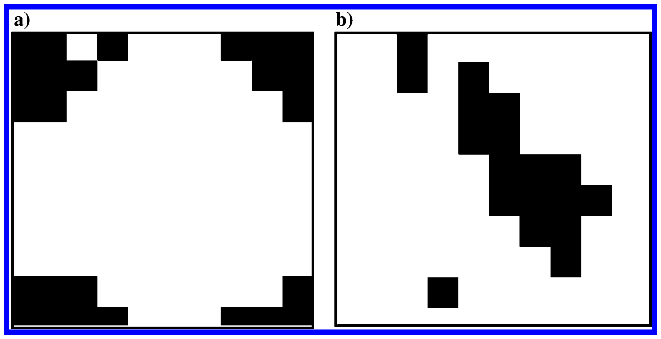


Figure 3. SOM projection as a 10 × 10 grid of the compound distribution in a high-dimensional space spanned by (a) 120 ALOGP parameters and (b) by 63 selected 1-D and 2-D descriptors (see text). Shown is the binary classification of chemical space. Regions in black: no UV signal detected. Regions in white: strong UV absorbance. Note that the maps form a torus.

Analysis of the Dataset and Descriptor Selection. We applied several methods for classification and modeling the data. As a large number of molecular descriptors can be calculated easily, we were interested in only considering those sets of descriptors that are relevant for modeling the data. We employed the SOM approach to evaluate the value of the different sets of descriptors in an analogous manner, as described earlier.[4,23] The high-dimensional descriptor matrix is projected onto two dimensions. The resulting map indicates a good classification, if actives and inactives are clearly separated (indicated by the separation of black and white patches).

For the UV signal intensities, we obtained the best classification using a combination of the 120 ALOGP parameters[16] (ALOGP) with our in-house collection of 63 1-D and 2-D

Table 2. Overview on the Statistics of the Knowledge-Based UV-Prediction Model Compared to the PLS Model

	knowledge-based model[a]	PLS model[a]	consensus model[a]
true positives	95% (920)	88% (848)	96% (922)
false negatives	5% (44)	12% (116)	4% (40)
true negatives	97% (275)	96% (273)	98% (277)
false positives	3% (9)	4% (11)	2% (6)
unclear			1% (3)
Matthew coefficient	0.89	0.76	0.90

[a] The given percentage refers to the entire data set of 1248 compounds (the number of compounds is given in parentheses).

descriptors (see Materials and Methods section for details). The corresponding SOM maps are shown in Figure 3.

We calculated the binary Matthews coefficient (see eq 1) to be cc = 0.76 and cc = 0.62 for the ALOGP and in-house sets, respectively. This raw classification accuracy was obtained with equal weights on the individual descriptor contributions. In an analogous manner, we analyzed the feasibility of different descriptor sets for the prediction of positive MS signal intensities. As for the UV dataset, we found the best combination to be ALOGP parameters and our in-house descriptor set. Because the dataset is highly unbalanced, we complemented the compounds not showing any MS signal with a diverse set of MS-active compounds of the same number. For this subset, Matthews coefficients were calculated to be cc = 0.84 and cc = 0.71 for the ALOGP and the in-house descriptors, respectively.

Detection of Outliers. Employing the PCA technique, the best separation was again achieved by a combination of ALOGP and in-house descriptor sets for both MS and UV datasets. We critically evaluated experimental UV or MS signal intensities by comparing them to structural features for all compounds that were identified as potential outliers. Eventually, all confirmed outliers were removed to avoid models influenced by extreme raw data values.

Linear Prediction Model. As we already found a good separation between actives and nonactives applying linear PCA models, we used discriminant PLS for building prediction models. The advantage of linear models is that the result is much easier to interpret than that of nonlinear models, which we preferred over a potentially better classification when applying nonlinear models.

We selected diverse and balanced training sets out of the data for UV and MS signal intensities. The selection was based on the decorrelated PCA scores. The size of the training set was determined by the number of negatives, which was the smaller fraction of the total dataset in both cases. Here, the importance of variables could be estimated with the variable influence on projection parameters (VIP) score computed by SIMCA-P.[19]

To reduce noise, we applied a moderate variable selection based on VIP scores. We carefully monitored the stability of the model by comparing the ranking of variables in the presence or absence of compounds with extreme values in one or more variables. The best UV model correctly classified 88% of the active and 96% of the inactive compounds (Table 2), yielding a Matthews correlation coefficient of $cc_{training} = 0.90$ ($cc_{test} = 0.76$). (Please note that, in contrast to the training set, the test sets are not balanced; see the Materials and Methods section for details.) The result in parentheses corresponds to the independent test set of remaining compounds that were not used for training the model. In contrast, the MS dataset correctly classified 94% of the MS active and 88% of the MS-inactive compounds (Table 3). The Matthews coefficients were calculated to be $cc_{training} = 0.82$ ($cc_{test} = 0.84$). Interestingly, the value for the test dataset is higher for the MS intensity prediction. This is due to the fact that the model performs better for the MS-positive compounds that are more highly populated in the test set. For both models, three components were found to be significant for representing the data without a tendency to overfit. The corresponding plots are shown in Figure 4; the most relevant descriptors are given in Table 4.

Table 3. Statistics of the PLS Model for Predicting MS Signal Intensities

	statistical model[a]
true positives	94% (318)
false negatives	6% (20)
true negatives	88% (221)
false positives	12% (30)
Matthew coefficient	0.83

[a] The given percentage refers to the entire data set of 589 compounds (the number of compounds is given in parentheses).

In linear prediction models, each descriptor has a well-defined contribution to the predicted activity. Thus, for all compounds, the overall classification score, a value between 0 (=inactive) and 1 (=active) for binary models, can be partitioned into the individual variable contributions. This allows mapping activity onto structural parameters. In the case of our models for UV and MS signal intensities, we would expect that descriptors that are important are those that represent the presence of delocalized π electrons and positively ionizable groups, respectively. Indeed, analyzing the highest scoring variables in Table 4 shows that descriptors linked to aromaticity are very important. In contrast, molecules containing a large proportion of (halogenated) aliphatic C atoms are expected to be UV inactive. This is reflected by the positive or negative contributions, respectively, of these descriptors to the UV activity. In summary, small to medium sized molecules with at least one aromatic ring are very likely to absorb in the detectable UV range. Examples of this category are depicted in Chart 1a, whereas Chart 1b shows representatives of UV-inactive molecules.

For the prediction of MS signal intensities, the first two highest ranking variables are linked to oxygen atoms. Together with the number of negatively charged atoms, they have negative contributions to MS activity. In contrast, the number of positively charged atoms or structural features representing positively ionizable groups, like aliphatic N atoms, have positive contributions to activity. As expected, the accessibility of compounds for positive ionization seems to be most relevant for our definition of MS activity. However, the detrimental importance of keto groups (ranked

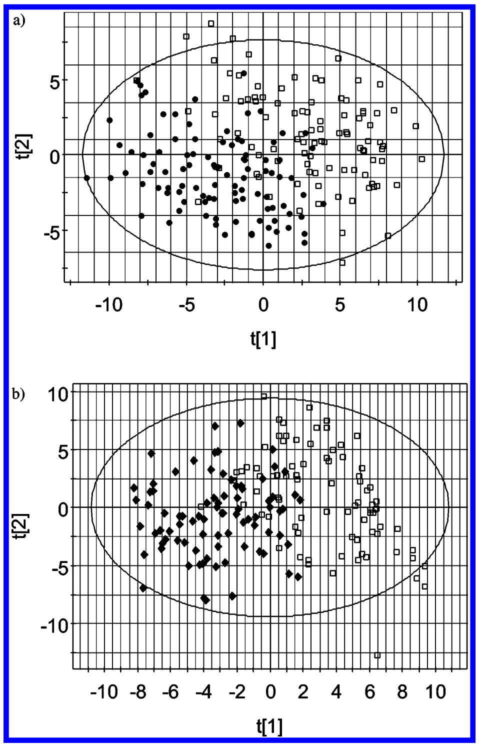


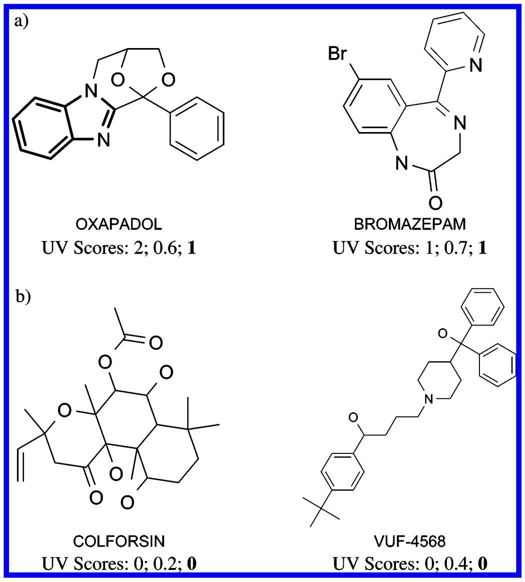
Figure 4. Scoreplots of the first two components obtained by discriminant PLS of the (a) UV and (b) MS datasets. UV-active or MS-active compounds are drawn as diamonds; black squares represent compounds not showing the corresponding UV or MS signal, respectively. The ellipses represent the limits of the 95% confidence region.

Table 4. Ten Most Relevant Descriptors According to VIP Analysis Obtained from Discriminant PLS Models for UV and MS Signal Intensities

rank	variables[a]	VIP
	UV Signal Intensity	
1	total number of aromatic rings	1.83
2	total number of aromatic bonds	1.83
3	total number of aromatic atoms	1.81
4	aromatic density	1.79
5	R− −CX− −R (ALOGP 26)	1.65
6	R− −CH− −R (ALOGP 24)	1.49
7	total number of rings	1.47
8	CR_3X (ALOGP 11)	1.43
9	FC^{sp3} (ALOGP 81)	1.40
10	ClC^{sp3} (ALOGP 86)	1.40
	MS Signal Intensity	
1	total number of oxygen atoms	1.53
2	=O (ALOGP 58)	1.48
3	HC^2_{sp3}, HC^1_{sp2}, HC^0_{sp}(ALOGP 48)	1.38
4	total number of nitrogen atoms	1.31
5	CH_2RX (ALOGP 6)	1.30
6	total number of negatively charged atoms	1.25
7	total number of positively charged atoms	1.24
8	total number of atoms	1.21
9	valence atomic connectivity index rank 1	1.19
10	Al_2NH (ALOGP 67)	1.17

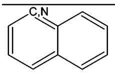
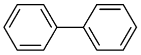
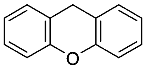
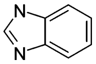
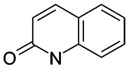
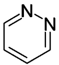
[a] The description and number of the ALOGP parameter refers to ref 17; X represents any heteroatom (O, N, S, P, Se, and halogens); Al represent aliphatic groups; − − represents an aromatic bond as in benzene or delocalized bonds such as the N−O bond in a nitro group; the subscript represents hybridization and the superscript its formal oxidation number. The formal oxidation number of a carbon atom equals the sum of the formal bond orders with electronegative atoms.

Chart 1. Examples of Compounds that Were Correctly Classified by Statistical and Knowledge-Based Models To Be (a) UV-Active and (b) UV-Inactive[a]



[a] The given compound names refer to the names in WDI.[24] The UV scores are given in the following order: knowledge-based, statistical, and consensus models (bold). Bold substructures represent structural moieties recognized by the knowledge-based model (compare Tables 5 and 6).

Table 5. Selection of Known Chromophores Representing Class 1 in the Knowledge-Based Prediction Model for UV-A Absorbance

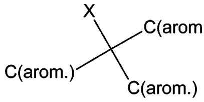
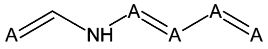
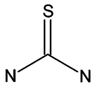
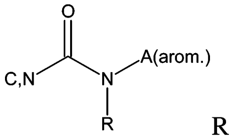
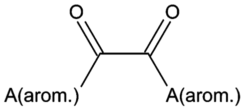
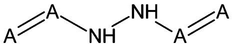
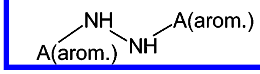
Chromophore	Name	Number of occurrences[a]	Number of false positives[a]
	Naphtyl /Quinoline	53	0
	Biphenyl	81	0
	Xanthene	8	0
	Benzimidazole	57	1
	Hydroquinolinone	72	1
	Pyridazin	4	0

[a] The given number refers to the number of compounds carrying the respective moiety within the total dataset of 1248 compounds (training set and test set).

2 in Table 4, under the MS subheading) might be attributable to one of the following causes: (a) instability of the MH^+ ions toward fragmentation, (b) enhanced probability of cluster formation (Na^+ or solvent adduct formation), or (c) reduction of the proton affinity of adjacent N atoms.

Knowledge-Based Prediction Model. In addition to the statistical PLS model, a knowledge-based approach was developed identifying chromophores and substructural motifs that are responsible for UV-A absorbance. We distinguish between three classes: (1) explicit substructures of known chromophores (Table 5), (2) structural features predominantly found in UV-active compounds (Table 6), and (3) motifs with highly delocalized electrons that lead to absorbance outside the detectable UV range or cause fluorescence, hampering signal detection (Table 1). Using our UV training set, the initial list of substructures (see Materials and

Table 6. Selection of General Substructural Features (Class 2) Used in the Knowledge-Based Prediction Model for UV-A Absorbance

Structural Motifs[a]	Name	Number of UV-active compounds containing this motif[b]	Number of false positives[b]
 X:= O,Cl,Br,I,S	Conjugated systems	45	0
			0
			3
 R: any C not H	Substituted Carbonyls	93	0
			0
 	Delocalized π-system	25	0

[a] A represents any aliphatic heteroatom, and A/C(arom) represents any aromatic atom or aromatic carbon, respectively. [b] The given number refers to the number of compounds carrying the respective moiety within the total dataset of 1248 compounds (training set and test set).

Methods) was refined by optimizing the number of false positives and negatives obtained during substructure recognition. If UV-active compounds having similar or identical substructures were not recognized, we introduced the common motif as a new substructure. This iterative optimization was stopped either if no additional substructure could be extracted or if the new substructure did not enhance the quality of the prediction. We excluded all structures containing metal atoms, as we found their physicochemical properties to differ significantly from purely organic compounds. The obtained model was validated using the test set, analogously to the PLS model described above.

In the final model, we obtained Matthews correlation coefficients of cc = 0.89 for the entire dataset ($cc_{training}$ = 0.94), corresponding to correct predictions of 95% of the UV-active and 97% of the UV-inactive compounds, respectively (Table 2). Examples for correctly classified compounds are shown in Chart 1. Oxapadol was classified as UV-active because, as with the benzimidazol substructure, it contains a known chromophore (score = 2; see Table 5). In contrast, the classification of bromazepam is based on the recognition of general structural features of UV-active molecules (score = 1).

Consensus Model for Predicting UV Signal Intensities. Having developed two independent prediction tools for UV-A absorbance, we could compare the results of the individual predictions and see whether a consensus model would increase the reliability of the prediction. In contrast to the knowledge-based model, the statistical PLS model gives continuous scores that had to be converted into the respective activity class. For calculating the Matthews coefficients as above, we always used a threshold of 0.5 for assigning UV activity: Compounds with a PLS score of 0.5 or higher were assigned to be UV-active. Compounds below 0.5 were predicted to be UV-inactive. However, comparing the predicted PLS scores to the class membership (Figure 5) shows that, in the region between 0.4 and 0.6, no clear classification is possible ("gray zone"). Hence, we abstained from classifying these compounds on the basis of PLS scores and only considered the knowledge-based model to compensate for this deficiency. If we translate this into the individual predictions shown in Chart 1, the individual PLS scores of oxapadol (Chart 1a) and colforsin (Chart 1b) would

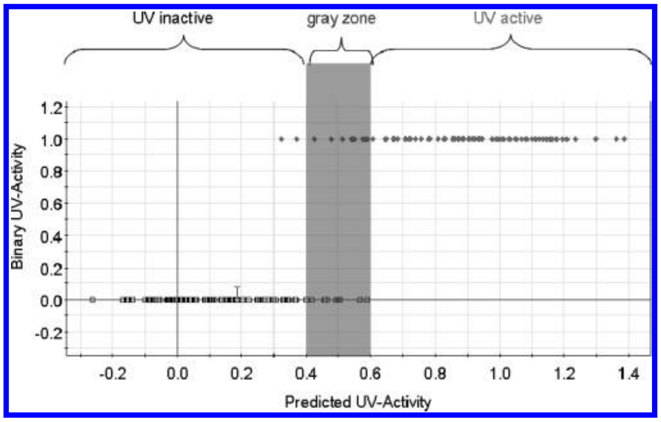
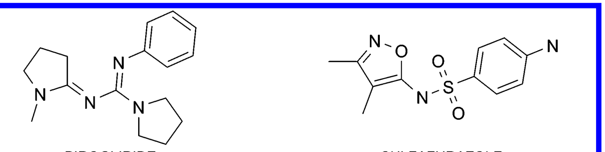


Figure 5. Plot of predicted vs observed class memberships of UV activity for the PLS model. The gray zone represents the region in which a reliable classification is not possible. UV-active compounds are represented by black squares; compounds not showing an UV signal are depicted as diamonds.

Chart 2. Examples of Compounds that Were Incorrectly Classified by Statistical and Knowledge-Based Models[a]



PIROGLIRIDE

UV Scores: 0; 0.4; **0;** Experiment: **UV active**

SULFAFURAZOLE

UV Scores: 0; 0.5; **0;** Experiment: **UV active**

[a] The given compound names refer to the names in WDI.[24] The UV scores are given in the following order: knowledge-based, statistical, and consensus models (bold).

fall into this error-prone gray zone. Here, the independent knowledge-based prediction model assists in correctly classifying both compounds. Also, this approach has its limitations, as demonstrated by the two examples in Chart 2; here, both prediction models failed.

When applied to the entire data set, the consensus model yielded 40 false negatives, corresponding to 96% UV-active compounds classified correctly. Six compounds were predicted to be UV-active out of the set of 284 inactive compounds (2% false positives). We obtained contradictory results for only three compounds (1%). A comparison of the statistics for all models can be found in Table 2.

In total, we calculated the Matthews correlation coefficient for the consensus model to be cc = 0.9 for the classification of the entire dataset (here, the three compounds with unclear results were treated as wrongly classified).

Validating the Prediction Models. For validation, we ran all UV and MS prediction models on the list of available compounds in our in-house library. We then partitioned the compounds according to the predicted classification into three sets of (1) actives in UV and MS, (2) compounds showing an UV signal but which cannot be detected with positive MS, and (3) compounds showing a MS signal but no UV absorbance. We selected a structurally diverse collection out of every set and measured their spectral properties during the standard purity control. A comparison with the classification by the consensus model yielded correct predictions for 93.2% of the UV actives (6.8% false negatives) and for 66.7% of the UV inactives (33.3% false positives; cc = 0.60). The individual models were slightly worse with respect to the classification of UV actives: 92.4% (PLS) and 92.7% (substructure-based). For the MS prediction, we found correct predictions for 88% of the positives (12% false negatives) and 63% of the negatives (37% false positives; cc = 0.4). The relatively low value for the Matthews correlation coefficient is due to the less accurate prediction of MS negative compounds. It has to be considered that multiply charged compounds or compounds forming unstable ions might be missed by the MS-peak detection software, but they were predicted to be MS-positive by our model. As the majority of compounds are MS active, however, this potential deficiency of the model has no big impact on the overall reliability of the prediction.

CONCLUSION

We have developed a novel prediction tool for the prediction positive ESI−MS and UV signal intensities to assist high-throughput compound purity assessments. All models proved to be robust and reliable even for a very diverse collection of compounds such as our additional validation set. We obtained the best results using a consensus scoring scheme for UV prediction integrating two complementary methods, a knowledge-based model and a statistical PLS model. All prediction models are now incorporated into our analytical workflow (Figure 1b) and facilitate the evaluation of experimental data in an automated fashion. During this routine application, we will collect more information that can be directly used to improve the prediction models.

ACKNOWLEDGMENT

We thank Eva-Maria Gutknecht and the Compound Depository team for their excellent logistical support and Christian Albrecht for running the LC/MS analysis. We also thank Wolfgang Guba, Walter Huber, and our colleagues in Molecular Modeling and Analytical Services for fruitful discussion at various stages of the study and during the preparation of this manuscript.

REFERENCES AND NOTES

(1) Landro, J. A.; Taylor, I. C. A.; Stirtan, W. G.; Osterman, D. G.; Kristie, J. HTS in the new millennium. The role of pharmacology and flexibility. *J. Pharmacol. Toxicol. Methods* **2001**, *44*, 273−289.

(2) Willett, P. Chemoinformatics−similarity and diversity in chemical libraries. *Curr. Opin. Biotech.* **2000**, *11*, 85−88.

(3) Zuegge, J.; Fechner, U.; Roche, O.; Parrott, N. J.; Engkvist, O. A fast virtual screening filter for cytochrome P450 3A4 inhibition liability of compound libraries. *Quant. Struct.-Act. Relat.* **2002**, *21*, 249−256.

(4) Roche, O.; Schneider, P.; Zuegge, J.; Guba, W.; Kansy, M. Development of a virtual screening method for identification of “frequent hitters” in compound libraries. *J. Med. Chem.* **2002**, *45*, 137−142.

(5) Leach, A. R.; Bradshaw, J.; Green, D. V. S.; Hann, M. M.; Delany, J. J. Implementation of a system for reagent selection and library enumeration, profiling, and design. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 1161−1172.

(6) Martin, E. J.; Critchlow, R. E. Beyond mere diversity: Tailoring combinatorial libraries for drug discovery. *J. Comb. Chem.* **1999**, *1*, 32−45.

(7) Böhm, H. J.; Stahl, M. Structure-based library design: molecular modelling merges with combinatorial chemistry. *Curr. Opin. Chem. Biol.* **2000**, *4*, 283−286.

(8) Kerns, E. H.; Di, L. Pharmaceutical profiling in drug discovery. *Drug Discovery Today* **2003**, *8*, 316−323.

(9) Lee, M. S.; Kerns, E. H. LC/MS applications in drug development. *Mass Spectrom. Rev.* **1999**, *18*, 187−279.

(10) Molnar, S. P.; King, J. W. Correlation of Ultraviolet Spectra via the Integrated Molecular and Electronic Transformations. *Int. J. Quantum Chem.* **1997**, *65*, 1057−1056.

(11) Fitch, W. L.; McGregor, M.; Katritzky, A. R.; Lomaka, A.; Petrukhin, R. Prediction of Ultraviolet Spectral Absorbance Using Quantitative Structure−Property Relationships. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 830−840.

(12) Gasteiger, J.; Rudolph, C.; Sadowski, J. Automatic Generation of 3D-Atomic Coordinates for Organic Molecules. *Tetrahedron Comput. Methodol.* **1990**, *3*, 537−547.

(13) Kohonen, T. Self-organized formation of topologically correct feature maps. *Biol. Cybernetics* **1982**, *43*, 59−69.

(14) Otto, M. *Chemometrics−Statistics and Computer Application in Analytical Chemistry*; Wiley-VCH: New York, 1999.

(15) Eriksson, L.; Johansson, E.; Kettaneh-Wold, N.; Wold, S. *Multi- and Megavariate Data Analysis−Principles and Applications*; Umetrics AB: Umeå, Sweden, 2001.

(16) Viswanadhan, V. N.; Ghose, A. K.; Revankar, G. R.; Robins, R. K. Atomic physicochemical parameters for three-dimensional structure directed quantitative structure−activity relationships. 4. Additional parameters for hydrophobic and dispersive interactions and their application for an automated superposition of certain naturally occurring nucleoside antibiotics. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 163−172.

(17) Ghose, A. K.; Viswanadhan, V. N.; Wendoloski, J. J. Prediction of Hydrophobic (Lipophilic) Properties of Small Organic Molecules Using Fragmental Methods: An Analysis of ALOGP and CLOGP Methods. *J. Phys. Chem. A* **1998**, *102*, 3762−3772.

(18) Wold, S. Exponentially weighted moving principal component analysis and projection to latent structures. *Chemom. Intell. Lab. Syst.* **1994**, *23*, 149−161.

(19) *SIMCA-P+*, version 9.0; Umetrics AB: Umeå, Sweden.

(20) Kennard, R. W.; Stone, L. A. Computer Aided Design of Experiments. *Technometrics* **1969**, *11*, 137.

(21) Matthews, B. W. Comparison of predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta* **1975**, *405*, 442−451.

(22) *Daylight Toolkit*, version 4.71; Daylight Chemical Information Systems, Inc.: Mission Viejo, CA. http://www.daylight.com.

(23) Roche, O.; Trube, G.; Zuegge, J.; Pflimlin, P.; Alanine, A. A virtual screening method for prediction of the hERG potassium channel liability of compound libraries. *ChemBioChem* **2002**, *3*, 455−459.

(24) World Drug Index, Thompson/Derwent. http://thomsonderwent.com/products/lr/wdi.

CI0496548