Prediction of Ultraviolet Spectral Absorbance using Quantitative Structure: Property Relationships

William L. Fitch*, Malcolm McGregor, Affymax Inc., 4001 Miranda Ave., Palo Alto, CA 94304

Alan R. Katritzky*, Andre Lomaka, Ruslan Petrukhin, Center for Heterocyclic Compounds, Department of Chemistry, University of Florida, P.O. Box 11720, Gainesville, FL 32611-7200

Mati Karelson*, Department of Chemistry, University of Tartu, 2 Jakobi Str., Tartu, Estonia

ABSTRACT

High performance liquid chromatography (HPLC) with ultraviolet (UV) spectrophotometric detection is a common method for analyzing reaction products in organic chemistry. This procedure would benefit from a computational model for predicting the relative response of organic molecules. Models are now reported for the prediction of the integrated UV absorbance for a diverse set of organic compounds using a quantitative structure- property relationship (QSPR) approach. A seven-descriptor linear correlation with a squared correlation coefficient (R^2) of 0.815 is reported for a

dataset of 521 compounds. Using the sum of ZINDO oscillator strengths in the integration range as an additional descriptor allowed reduction in the number of descriptors producing a robust model for 460 compounds with 5 descriptors and a squared correlation coefficient 0.857. The descriptors used in the models are discussed with respect to the physical nature of the UV absorption process.

INTRODUCTION

In organic synthesis there have been two traditional methods for performing quantitative analysis. For novel molecules, the method has been to purify the new structure to homogeneity and then weigh the sample. For known samples, a pure reference standard could be obtained which allowed for chromatographic quantitation in comparison to the new batch. In the modern drug discovery laboratory, analysts are asked to quantify the amount of target compound in hundreds of novel samples each day. These molecules are made in sub-milligram amounts and have never been synthesized before. The only information available is the structure of the molecule and properties which can reliably be calculated from the structure. Because the new tools of combinatorial chemistry allow so many compounds to be synthesized in a short time, the old strategy of purify and weigh is no longer satisfactory

The analytical chemistry community has only begun to address this need[1-4]. Approaches to quantitation of unknowns include (i)NMR[5,6]; (ii)HPLC with evaporative light scattering detection[7,8], and probably the most successful solution to this problem, (iii) the recent development and popularization of the combustion-based

chemiluminescent nitrogen detector (CLND) for HPLC[9,10]. Response in this latter detector can be predicted solely from structure because all nitrogens burn to the same analyte. The application of this detector to assessing high throughput parallel synthesis is rapidly increasing in popularity[11].

A specific application of interest to Affymax is the need to quantify compounds synthesized in encoded split-pool libraries for high throughput screening. These experiments are done for the quality control of a split pool encoded library[12,13]. In this technique, sub-nanomole amounts of compound are synthesized. Their structures are confirmed by LC/MS while the LC/UV signal is used to assess purity. It is difficult to assess these samples for amount because none of the standard quantitation techniques (weighing, NMR,ELSD, CLND) has sufficient sensitivity. Better knowledge of released concentration could improve our understanding of hit rates and overall success in bead-based high throughput screening[14].

A second application for generic quantitation is in the impurity profiles of drug substance for regulatory approval. In this process a relatively pure substance is tested by HPLC/UV/MS. All impurities above 0.1% should be identified and quantified. The MS data and knowledge of the process are often sufficient to identify the impurities but quantitation requires the laborious synthesis of a standard pure sample.

We now report attempts to predict response in typical HPLC UV detectors directly from structure. This is a remarkably unexplored goal. Others have derivatized molecules so that a common chromophore can be used for quantitation[15], but only one attempt to do generic UV intensity prediction could be found in the literature[18]. Such predictive capability would be very useful because UV is a nearly universal detector for

drug-like molecules: 85% of the structures in the MDDR (a database of drugs and candidate drugs[16]) contain an aromatic group and most of the others contain amides or other chromophores. In addition, HPLC with UV detection is very widely available as a routine analytical tool in the organic chemistry laboratory. Modern sophisticated diode array UV detectors are easy to use, rugged and reasonably priced.

In the best case, we would like to predict the UV concentration of a compound to within 10% of its true value. But prediction schemes with more error, 20% or even 50% may be useful. Optimally, the procedure should be fast. It would be best if a chemist could draw a structure and have its predicted UV returned in a short time, as for CLOGP[17]. Also, a calculation for the 96 numbers needed for quantitative analysis of a parallel synthesis plate should not take hours.

Theory of UV quantitation

Three typical UV spectra are shown as Fig. 1. Normally, HPLC detection is done at a single wavelength; 220 nm is commonly chosen as the most generic wavelength because of cases like that of Fig 1a. However, the slope of the absorbance spectrum is often very steep at any single wavelength. Moreover a single wavelength poorly captures the magnitude of the absorbance, the feature that is most directly structure-related and hence predictable[18]. For these reasons, we have chosen to integrate the entire spectrum. The practical lower limit to this sum is 220 nm due to HPLC solvent absorbance. The upper limit is arbitrarily set to 360 nm as drugs rarely have absorbance in the visible.

Beer's law states that the light absorption at a single wavelength is proportional to the sample concentration

$$A=\varepsilon bc \quad (1)$$

where ε is the extinction at that wavelength and b is the light path length. This equation holds at each wavelength increment so, by summing a set of these equations, it is apparent that

$$\int A=\int \varepsilon bc \quad (2)$$

Since b and c are wavelength independent, the area under the absorbance curve is proportional to a compound dependent overall extinction number (call it E)

$$\int A=Ebc \quad (3)$$

This extinction can be thought of as the average extinction times the wavelength range. Equation 3 will hold for either single wavelength UV detection or broadband (220-360 nm) detection; so we drop the integral henceforth. This concept of integrating the UV resonances is not novel. It has been used recently to improve signal to noise for trace analysis instrumentation[19].

In HPLC, we need to deal with peaks and changing concentration profiles. It is useful to consider breaking a peak into time segments. For each segment the absorbance will be proportional to the concentration at that moment.

$$A_i=Ebc_i \quad (4)$$

The area of each time segment i will be

$$\text{Segment area}=A_i t_i = Ebc_i t_i \quad (5)$$

The summed areas will be the area of the chromatographic peak

$$\text{area} = Eb\Sigma c_i t_i \quad (6)$$

with Eb outside the sum, since the E and b are constants. Now replace the c_i with the equivalent amount (a_i) divided by the volume (v_i)

$$\text{area} = Eb\Sigma(a_i/v_i)t_i \qquad (7)$$

For a constant flow HPLC method, the volume per unit time is a constant and can be taken out of the sum and included in a new constant term

$$\text{area} = Eb^{*}\Sigma a_i \qquad (8)$$

Finally, the sum of the amounts in each time segment (excluding on-column losses) is the total amount injected. The key equation then is

$$\text{LC peak area} = E\ I\ N \qquad (9)$$

where I is an instrumental factor which includes the UV cell pathlength and the flow rate, and N is the number of moles injected. For cases where the analyst wants to measure injected sample concentration instead of injected amount, the equation is

$$\text{LC peak area} = E\ I\ c \qquad (10)$$

Where the I now includes an injection volume contribution. The E values will come from the prediction scheme. The I values can be measured for each instrument and method using standard compounds.

To predict the characteristics of UV absorption peaks, several quantum chemical models have been developed. The transition frequency can be predicted by calculating the energies of excited electronic states by a configuration interaction (CI) calculation. The intensity of the transitions or oscillator strength can be obtained from the transition dipole moment, which is proportional to the change in the electric charge distribution occurring during excitation. Theoretical oscillator strength is the intensity of this

electronic energy transition; it corresponds to the height of a widthless transition. The experimental oscillator strength[20] is measured by integrating the area under an absorbance band using equation 11.

$$f_{osc}(I) = 4.319\,x\,10-9\int e(\nu)d\nu \qquad (11)$$

One of the most successful models for the calculation of UV spectra is the ZINDO modification of the Intermediate Neglect of Differential Overlap method [21]. Since most of the spectra are taken of molecules in solution, treatment of solvent effects on UV spectra has been an area of active research. The explicit consideration of solvent molecules in the quantum mechanical self-consistent field molecular orbital calculations is usually not feasible. Therefore, various continuum solvent models (CSM) that treat solvent as a simple dielectric continuum have been developed. Traditional CSM is the self consistent reaction field (SCRF) model[22]that has been shown to reproduce the shifts of absorption peaks in aprotic solvents very well[23]. Another continuum solvation model - conductor-like screening model (COSMO) when implemented in the framework of MOPAC reproduces solvatochromic shifts qualitatively in AMI calculations[24]. Specific effects of protic solvents on the UV spectra of uracil and uracil derivatives have been studied using hybrid quantum chemical and molecular mechanics method[25].

Theoretical models have also been developed to approximate the band shape of molecular electronic transitions. Much of the broadening of spectral lines occurs because there are many vibrational and rotational transitions with slightly different energies. Another cause of line broadening is the anisotropic interaction with the medium (solvent). An empirical method for reproducing the band shape from a single geometric

structure has been developed[20] that is significantly faster than molecular dynamics approach and potentially applicable for predicting the appearance of the spectra of large molecular systems. A new recently described parameterization of INDO is equally good for both geometry optimization and spectroscopy [26]. But none of these techniques does a particularly good job in predicting oscillator strengths (absorbance intensities).

A unified treatment of the absorption intensities is further complicated by large differences in the oscillator strengths for transitions of different symmetry. Moreover, additional solvent-induced broadening of the spectral bands arises from the variation of the local environment of the chromophoric solute molecule in the condensed medium. The latter is caused by the thermal motion of the surrounding solvent molecules. At any given instant of time, there is a distribution of differently solvated solute molecules, each of which has characteristic transition energy to the excited state. The resulting distribution of the transition energies leads to the broadening of the spectral band. The theoretical assessment of the solvent-induced spectral broadening has thus to rely on a proper statistical treatment of the solvent distribution around the chromophoric solute molecule, both in the ground and in the excited state of the latter[27].

A QSPR approach has previously been applied to the prediction of absorption wave numbers and molar absorptivities[18]. In this study various structure indices such as the integrated molecular transform and normalized molecular moment indices were used to establish the correlation model. Modeling of molar absorptivity was not successful in this study evidently because absorptivity at a single wavelength (maximum) rather than the integrated area was used.

The aim of the current study is to develop QSPR models for the rationalization and prediction of the ultraviolet integrated absorption for a diverse set of organic compounds at a precision level suitable for application to analytical work. The QSPR method is applied in the framework of the CODESSA program[28]: CODESSA has successfully correlated many properties including boiling points[29a,b], gas chromatographic response factors[29c,d], critical micelle concentrations[29g,h], solubilities[29g,h], polymer glass transition temperatures[29i,j], refractive indices[29k,l], viscosities[29m], and solvent effects on decarboxylation rates[29n]; for reviews see[29o,p].

METHODOLOGY

The availability of UV data has been reviewed recently[30]. The UV spectral data for this study was taken from the Upstream Solutions electronic database (Upstream Solutions, GmbH, Hergiswil, Switzerland; www.upstream.ch). This collection was originally published as the UV-VIS Atlas of Organic Compounds[31]. After transferring the entire collection into ISIS and Oracle databases, a set of 521 small organic compounds and spectra were selected for study.

The prediction scheme is designed to correlate structural descriptors to the integrated UV spectrum from 220-360 nm. An Oracle PL/SQL procedure was written to do the integration. First, the data below 220 and above 360 is ignored. If necessary, interpolation or extrapolation (limited to 5 nm) is used to generate good endpoints. The algorithm then integrates the data between the wavelengths using the trapezoid rule.

$$Area_{220\text{-}360}=\Sigma(\lambda_{i+1}-\lambda_i) \times (\varepsilon_i+\varepsilon_{i+1})/2 \qquad (12)$$

Correlations were produced from HM PRO (the Heuristic method for CODESSA PRO), which has an algorithm consisting of 4 major parts:

1. The 1-parameter descriptors selection. The selection is based on the squared correlation coefficients, Fisher *F*-criteria, and Student *t*-criteria. Highly intercorrelated descriptors and descriptors with insignificant variance are eliminated.
2. Pair-wise selection. This selection is made on the basis of squared correlation coefficient s and Fisher *F*-criteria.
3. Expanding/contracting stage. The unexpanded correlations in the correlation set are expanded by adding previously unselected descriptor. The number of correlations added in this manner can be limited by branching criteria (number of added correlations per each expansion), limiting *F*-criteria (normalized or not), squared intercorrelation coefficients, and standard errors. Correlations with the maximum number of descriptor (given parameter) allowed are not expanded. This stage will be repeated until a stop event occurs, which can be any/all of following:
    a. The correlation set in the memory is overfilled. When the value of fitness function is less than minimal in the set, it is not stored at all after overfilling. In this case, if correlation is inserted into the set, the correlation with the worst value of the fitness function is eliminated. The fitness function is defined as $w = (R^2Fn)/(Ns^2)$ where R^2 is the squared correlation coefficient, F is Fisher criterion, N is the number of descriptors in the model and s is standard deviation.

b. Maximum number of iterations is reached.

c. Time limit is reached.

d. The correlation set does not have any correlation to expand/contract (full search is finished).

4. The output stage. A predetermined number of the "best" correlations is printed out. Iterations for selecting for the printed correlations begin from the "best" correlations. For all correlation in the cycle, a full set of statistical parameters is calculated including intercorrelations of the descriptors (one to all others), cross-validated squared correlation coefficients, etc. The parameters of the method are defined by the set of the selection criteria. For any correlation, a full list of the predecessors in order of calculation can be printed, on the basis of best correlations with subsets of the descriptors until 1-parameter correlation will be printed.

The analysis is subject to the condition that intercorrelation coefficient of a descriptor with respect to all other descriptors in the model remains below a predetermined level (0.5 in the current work).

ZINDO calculations were performed using the MOS-F package from Schrödinger Inc[32]. Each molecular geometry was first optimized with Corina[33] and then submitted to the program which outputs a set of chromophore absorbance frequencies and oscillator strengths.

RESULTS AND DISCUSSION

To this point the argument has been framed in wavelength terms. UV absorbance is caused by energy absorbance by a molecule causing electrons to change energy states. The prediction of extinction will involve analysis of differences in energy states. Therefore it could have been advantageous to consider absorbance in frequency space rather than wavelength space since energy and frequency are directly related and wavelength and frequency are reciprocally related. However, Figure 2 shows a correlation coefficient of 0.97 between integrated area under the wavelength curve and integrated area under the frequency curve for the 521 compounds chosen for this study; no further consideration of frequency is felt necessary.

The spectra in our dataset have been collected in many solvents (typically ethanol or alkanes). The solvent for which we desire information is the solvent composition at the moment of chromatographic elution. The composition of gradients in high throughput LC/UV/MS applications has largely been standardized to acetonitrile/water with pH adjustment using formic acid or TFA[34]. Methods for estimating the protonating power of these solvent mixtures is available[35]. While the energy of absorbance can be quite solvent dependent, the oscillator strength may not be too different in most cases[36]. Refractive index is a key contributor to solvent effects on oscillator strength[36]; but luckily, acetonitrile, water and their mixtures have essentially constant η[37]. Table 1 shows 4 molecules and their relative integrated absorbances in different solvents[38]. (i) The benzyl alcohol spectrum is not much effected by solvent. (ii) Nitroaniline shows a large solvent effect: in water more of its intensity is found below 360 nm compared to in ethanol.

Molecules which have strong absorption near 360 nm will need to be treated specially or excluded. (iii) Crotonaldehyde has all of its absorbance at low wavelength and most of this falls below 220 nm in hexane but above 220 nm in ethanol. Molecules which have all of their absorption below 250 nm will present special solvent effect problems. (iv) Aniline is effected by pH: its protonated form absorbs weakly while the neutral form has a strong absorbance. Thus, molecules must be represented in the correct charged form for their spectra. Many QSPR and quantum calculations do not deal well with formal charges.

Subtle steric effects can have big impacts on UV spectra. For example, Table 2 compares the maximum absorbance of a set of substituted anilines[39]. Crowding causes non-planarity of functional groups, drastically lowering the absorbance in the 2-t-butyl-N,N-dimethylaniline spectrum. This steric issue requires that structural descriptors be calculated from fully energy-minimized structures.

The entire molecule must be considered in predicting UV. It would be simplifying to isolate and add contributions from single chromophores. But chromophores interact, even through multiple sp3 carbons, as shown by Table 3 for a set of benzene analogs. This hyperconjugation effect has been known for a long time[40] but makes the success of a simple additivity scheme (a la CLOGP) highly unlikely for UV.

The first QSPR attempt applying constitutional, topological, geometrical and electrostatic descriptors produced a 7-descriptor correlation equation (Eq. 13) with $R^2 = 0.7063$, $R^2_{cv} = 0.6923$, $F = 175.54$ and $s = 3.5 \cdot 10^5$; see Fig. 3. The statistical parameters for this equation and other best equations with different numbers of descriptors in the range 1-7 are given in Table 4.

$$\begin{aligned} A = {} & (5.46 \pm 0.19) \cdot 10^5 N_b + (1.94 \pm 0.12) \cdot 10^5 N_d + (1.11 \pm 0.10) \cdot 10^5 \Phi + \\ & (1.85 \pm 0.30) \cdot 10^6 V'_M + (2.06 \pm 0.34) \cdot 10^4 PNSA3 - (5.21 \pm 1.00) \cdot 10^6 S_{zx} + \\ & (1.39 \pm 0.30) \cdot 10^5 \overline{{}^0\mathrm{IC}} - (1.29 \pm 0.16) \cdot 10^6 \end{aligned} \quad (13)$$

where N_b is the number of benzene rings, N_d, is the number of double bonds, Φ is the Kier flexibility index, V'_M is the factorized molecular volume, PNSA3 is the atomic charge weighted partial negative surface area, S_{zx} is ZX shadow area, and $\overline{{}^0\mathrm{IC}}$ is zeroth average information content. 28.2 % of structures in the whole dataset (147 out of 521) were predicted to within 20 % of relative error and 56.2 % of structures (293 out of 521) to within 50 % of relative error according to this equation. The number of data points lying outside the range of $\pm 2\sigma$ (95 % confidence limit) from the predicted value was 23.

In the next step, MOPAC SCF calculations were performed for the data set. Inclusion of the quantum chemical descriptors based on the output of MOPAC calculations improved the squared correlation coefficient R^2 to give a 7-descriptor correlation equation (eq. 14) with $R^2 = 0.8152$, $R^2_{cv} = 0.7996$, $F = 322.06$ and $s = 2.8 \cdot 10^5$; see Fig. 4. Statistical parameters for this equation and the best equations with lower number of descriptors are given in Table 5.

$$\begin{aligned} A = {} & (3.28 \pm 0.17) \cdot 10^5 N_b + (37.19 \pm 2.34)\ \gamma + (1.19 \pm 0.10) \cdot 10^5 N_d - \\ & (1.72 \pm 0.17) \cdot 10^5 (\varepsilon_{\mathrm{LUMO}} - \varepsilon_{\mathrm{HOMO}}) + (3.08 \pm 0.32) \cdot 10^4 PNSA3 + (1.61 \pm 0.17) \cdot 10^6 P_c - \\ & (1.81 \pm 0.32) \cdot 10^4 \Delta H^0_f / n_a + (0.25 \pm 2.43) \cdot 10^5 \end{aligned} \quad (14)$$

where γ is gamma polarizability, ($\varepsilon_{\mathrm{LUMO}}$- $\varepsilon_{\mathrm{HOMO}}$) is HOMO - LUMO energy gap, P_c is the average bond order of a C atom, and $\Delta H^0_f / n_a$ is the final heat of formation divided by the number of atoms. 36.5 % of structures in the whole dataset (190 out of 521) were

predicted to within 20 % of relative error and 61.2 % of structures (319 out of 521) to within 50 % of relative error according to this equation. The number of data points lying outside the range of $\pm 2\sigma$ (95 % confidence limit) from the predicted value was 24. Most of these outliers are molecules with very low absorbance in the integration region (molecules with low wavelength absorption maxima).

Most of the descriptors used in the first two equations are well interpreted and in good accordance with the physical picture of UV absorption process. The number of benzene rings and number of double bonds in equations 14 and 15 measure the unsaturation of the molecule. The average bond order of a C atom in Eq. 14 is also related to the unsaturation. Positive correlation coefficients are in good accordance with the fact that increase in the saturation of the molecule causes the shift of absorption maxima to higher frequencies outside the integration range. The negative correlation coefficient of ZX shadow area in Eq. 13 may be related to cavity formation in condensed media and solvent effects affecting UV spectra. Charge distribution-related descriptor PNSA3 in both equations may also be describing solvent effects. The second most significant descriptor in Eq. 14, polarizability, is related to charge migration or displacement during the transition from one electronic state to another. (ε_{LUMO}-ε_{HOMO}) approximates the energy difference between the electronic states, which determines the location of the absorption maximum.

Since the molecules with only low wavelength absorbance were poorly predicted, a dataset of 255 compounds having at least 40 % of their UV absorbance above 250 nm was subjected to study. This selection excludes the spectra such as Figure 1a where only a small overall portion of the UV absorbance falls above 220 nm. For this dataset,

heuristic correlation with 5 descriptors gave a model with $R^2 = 0.7426$, $R^2_{cv} = 0.7177$ and $F = 141.98$. The data is shown as Table 6, equation 15 and Figure 5.

$$\begin{aligned} A = {} & (3.17 \pm 0.33) \cdot 10^1 \gamma + (1.98 \pm 0.22) \cdot 10^5 \, N_b - (2.95 \pm 0.37) - 10^5 (\varepsilon_{\mathrm{LUMO}} - \varepsilon_{\mathrm{HOMO}}) \\ & - (1.10 \pm 0.22) \cdot 10^4 \, PPSA3 + (6.01 \pm 1.23) \cdot 10^3 \, S_{xy} + (3.11 \pm 0.37) \cdot 10^6 \end{aligned} \quad (15)$$

where PPSA3 is the atomic charge weighted partial positive surface area for solvent accessible surfaces. The XY shadow area in this equation can be directly related to molecular cross-section of absorption. The total molecular 2-center exchange energy divided by number of atoms reflects the change in the Fermi correlation energy between the electrons localized on different atoms and is important in determining the spin properties of molecules. In this model, 114 structures (44.7 % of total) were predicted to within 20% of relative error. In addition 185 compounds (72.5 % of the total) were predicted to within 50% of the measured value. Equation 15 is not applicable to end-absorption spectra such as Fig.1a but can be expected to give slightly increased reliability for the integrated intensity for the more common classes of spectra such as 1b and 1c.

Scrutiny of the points in Fig. 5 revealed different populations for the spectra taken in different solvents. So heuristic correlations were performed for the subsets of absorbances measured in nonpolar solvents (75 structures) and ethanol (93 structures). The optimal equations were developed for different numbers of descriptors in the range of 1-5 for nonpolar solvents and in the range 1-9 for polar solvents. Plots of R^2 and R^2_{cv} values against the number of descriptors (not shown), provide guidance regarding the number of descriptors to retain in the models. The models with 5 descriptors ($R^2 = 0.8581$, $R^2_{cv} = 0.8274$, $F = 83.48$) and 7 descriptors ($R^2 = 0.8041$, $R^2_{cv} = 0.7455$, $F = 49.83$) were retained for nonpolar and ethanol subsets, respectively. The descriptors

involved in these correlations are listed in Tables 7 and 8 in order of their statistical significance according to t-test. The respective correlation charts are depicted in Figures 6 and 7. In the nonpolar set 59 % of structures (44 out of 75) and in the ethanol set 60.2 % of structures (56 out of 93) were predicted to within 20 % of relative error.

The separate correlations for the polar and non-polar groups does not reflect exclusively solvent effects. Rather this result points up a weakness in the current descriptors' set. The descriptors which have been tested to date are inadequate to completely model the intensity of UV absorbance. Therefore, the next approach was to consider actually predicting UV absorbance with quantum mechanics and comparing this prediction to the database. The ZINDO program which is included in MOS-F works by identifying chromophores in a molecule, and then calculating the lambda max and estimating the oscillator strength. The output is a set of widthless absorbances as shown in Figure 8a. The software does not address the peak width issue[20] which would be necessary to convert the sticks into peaks as shown in 8b. Nor does the software address the convolution of these peaks to give a true molecular spectrum as shown in Fig 8c. But it was hoped that the UV integration approach might obviate the need to fully understand peak widths.

A limitation of the MOS-F software is that it is only parameterized for uncharged molecules. Therefore a set of 205 neutral compounds were selected. Each was structure optimized with Corina and then submitted to the UV prediction. Each predicted absorbance was artificially broadened with a 10 nm half-width gaussian shape to allow estimation of the percentage which would fall outside of our 220-360 nm range. After this correction, the oscillator strengths were summed and correlated to the integrated UV

absorbance. The result is shown as Fig 9. The correlation coefficient in this case is 0.77 and offered no advantage over the earlier and computationally simpler results.

In order to further improve the quality of the general model (Eq. 14), the ZINDO oscillator strengths were incorporated as external descriptors into a general QSPR treatment along with all other descriptors available from CODESSA. MOS-F calculations were performed with CNDO/S parameterization for a subset of 460 structures from the original dataset of 521 structures. For simplicity, the external descriptors I_{cndo} were calculated as the sums of all oscillator strengths in the range 220-360 nm neglecting the absorbance that would fall outside this range. The "best" 1-5 descriptor models of the integrated absorption in wavelength scale were obtained by using CODESSA HM PRO. No correlations with intercorrelation coefficient less than 0.5 were found for more than 5 descriptors using the correlation set in the upper segment of the fitness function of size 100000. The statistical parameters of the models developed using CNDO/S calculation results as additional descriptors are presented in Table 9.

Comparison of Tables 5 and 9 reveals that after including ZINDO calculation results the number of descriptors for the models with the same predictive power has been reduced significantly. The best model with 5 descriptors has $R^2 = 0.8573$, $R^2_{cv} = 0.8431$ and is given in Equation 16 with the corresponding correlation chart in Figure 10.

$$A = (4.13 \pm 0.19) \cdot 10^5 I_{cndo} + (3.28 \pm 0.21) \cdot 10^1 \gamma - (1.17 \pm 0.16) \cdot 10^5 (\varepsilon_{\mathrm{LUMO}} - \varepsilon_{\mathrm{HOMO}}) + (7.58 + 1.43) \cdot 10^4 N_b + (1.26 + 0.39) \cdot 10^5 P_c + (9.63 \pm 1.63) \cdot 10^5 \quad (16)$$

According to this equation 42.8 % of compounds (197 out of 460) were predicted to within 20 % and 67.6 % of compounds (311 out of 460) to within 50 % of the measured value. The second most significant descriptor in Eq. 16, gamma polarizability γ, that appears also in Eqs. 14 and 15 is defined according to the following equation

$$\mu' = \mu + \alpha \boldsymbol{E} + \frac{1}{2}\beta \boldsymbol{E}^2 + \frac{1}{6}\gamma \boldsymbol{E}^3 \qquad (17)$$

where $\boldsymbol{E}$ is the strength of the applied electrostatic field and μ' is the induced dipole moment. It can be related to the part of transition dipole moment explained by the third order contribution to the response of molecule's dipole moment to the external electrostatic field.

CONCLUSIONS

The prediction of UV spectral intensity should be a useful computational tool in organic chemistry. This capability would add to the armamentarium of spectroscopy and chromatography that analytical chemists bring to the increasing need for data in support of high throughput organic synthesis. When we started this project, no UV spectra were commercially available in electronic form. Our interest led to the release of the Upstream database. We would like to see a similar digitization of the Lang collection[38] and the other great collections[30] from the golden age of UV spectroscopy.

The development of highly significant QSAR or QSPR equations by extraction of molecular descriptors from large descriptor spaces has been successful for the prediction of many physical properties and biological activity of chemical compounds[29]. The present work demonstrates that analogous QSPR equations can be developed for the prediction of UV absorption area. Importantly, the descriptors employed in the best correlation equations are clearly relevant to the physical nature of UV absorption process. These descriptors are related to polarizability, saturation, spin properties, energy difference between the electronic states and solvents effects.

Future refinements in UV prediction will require further improvement in parameterization[26] and better algorithms[20] for calculating oscillator strengths, spectral peak widths and solvent effects.

ACKNOWLEDGEMENT

Steve Muskal and Ray Carhart helped with the Oracle implementation of a spectral database. John McKelvey offered helpful suggestions in the development of this work

REFERENCES AND NOTES

(1) Czarnik, A. W. Combinatorial Chemistry: What's in it for chemists? *Anal. Chem.* **1998***, 70*, 378-386A.

(2) Fitch, W. L. Analytical Methods for Quality Control of Combinatorial Libraries. *Annu. Rep. Comb. Chem. Mol. Diversity* **1997***, 1*, 59-68.

(3) Lewis, K., Phelps, D., Sefler, A. Automated High-Throughput Quantification of Combinatorial Arrays. *Amer. Pharm. Rev.* **2000***, 2000*, 63-68.

(4) Yan, B. *Analytical methods in combinatorial chemistry*; Technomic: Lancaster, PA, 2000; 268 p.

(5) Larive, C. K., Jayawickrama, D., Orfi, L. Quantitative analysis of peptides with NMR spectroscopy. *Appl. Spectros.* **1997***, 10*, 1531-1536.

(6) Gerritz, S. W., Sefler, A. M. 2,5-Dimethylfuran (DMFu): an internal standard for the "traceless" quantitation of unknown samples via 1H NMR. *J. Comb. Chem.* **2000***, 2*, 39-41.

(7) Fang, L., Wan, M., Pennacchio, M., Pan, J. Evaluation of evaporative light-scattering detector for combinatorial library quantitation by reversed phase HPLC. *J. Comb. Chem.* **2000***, 2*, 254-257.

(8) Zambias, R. A., Kassel, D.B. Automated on-line evaporation light scattering detection to quantify isolated fluid sample compounds in microtiter plate format. US Patent No. 6,077,438. 2000

(9) Taylor, E. W., Qian, M.G., Dollinger, G.D. Simultaneous on-line characterization of small organic molecules derived from combinatorial libraries for identity, quantity and purity by reversed phase HPLC with chemiluminescent nitrogen, UV and mass spectrometric detection. *Anal. Chem.* **1998***, 70*, 3339-3347.

(10) Fitch, W. L., Szardenings, A.K., Fujinari, E.M. Chemiluminescent nitrogen detection for HPLC: an important new tool in organic analytical chemistry. *Tetrahedron Lett.* **1997***, 38*, 1689-1692.

(11) Shah, N., Tsutsui, K., Lu, A., Davis, J., Scheuerman, R., Fitch, W.L. A Novel Approach to High-Throughput Quality Control of Parallel Synthesis Libraries. *J. Comb. Chem.* **2000***, 2*, 453-460.

(12) Lewis, K. C., Fitch, W.L., Maclean, D. Characterization of a split/pool combinatorial library. *LC/GC.* **1997***, 16*, 644-649.

(13) Dolle, R. E., Guo, J., O"Brien, L., Jin, Y., Piznik, M., Bowman, K.J., Li, W., Egan, W.J., Cavallaro, C.L., Roughton, A.L., Zhao, Q., Reader, J.C., Orlowski,

M., Jacob-Samuel, B., Caroll, C. A statistical-based approach to assessing the fidelity of combinatorial libraries encoded with electrophoric molecular tags. Development and application of tag decode-assisted single bead LC/MS analysis. *J. Comb. Chem.* **2000***, 2*, 716-731.

(14) Tan, D. S., Burbaum, J.J. Ligand discovery using encoded combinatorial libraries. Curr. Opin. Drug Discovery Dev. **2000***, 3*, 439-453.

(15) Williams, G. M., Carr, R.A.E., Congreve, M.S., Kay, C., McKeown, S.C., Murray, P.J., Scicinski, J.J., Watson, S.P. Analysis of solid-phase reactions: product identification and quantification by use of UV-chromophore-containing dual-linker analytical constructs. *Angew. Chem. Int. Ed.* **2000***, 39*, 3293-3296.

(16) The MDL Drug Data Report (MDDR) is a Commercial Database Available from MDL Information Systems Inc., San Leandro, CA.

(17) Leo, A. Calculating logPoct from structures. *Chem. Rev.* **1993***, 93*, 1281-1306.

(18) Molnar, S. P., King, J. W. Correlation of Ultraviolet Spectra with Structure via the Integrated Molecular and Electronic Transforms. *Int. J. Quantum Chem.* **1997***, 65*, 1047-1056.

(19) Aiello, M., McLaren, R. A sensitive small-volume UV/VIS flow cell and total absorbance detection system for micro-HPLC. *Anal. Chem.* **2001***, 73*, 1387-1392.

(20) Pearl, G. M., Zerner, M. C., Broo, A., McKelvey, J. Method of calculating band shape for molecular electronic spectra. *J. Comput. Chem.* **1998***, 19*, 781-796.

(21) Ridley, J., Zerner, M. An Intermediate Neglect of Differential Overlap Technique for Spectroscopy: Pyrrole and the Azines. *Theoret. Chim. Acta (Berl.)* **1973***, 32*, 111-134.

(22) Karelson, M., Zerner, M. C. Theoretical Treatment of Solvent Effects on Electronic Spectroscopy. *J. Phys. Chem.* **1992***, 96*, 6949-6957.

(23) Karelson, M., Pihlaja, K., Tamm, T., Uri, A., Zerner, M.C. UV-visible spectra of some nitro-substituted porphyrins. *J. Photochem. Photobiol., A* **1995***, 85*, 119-126.

(24) Klamt, A. Calculation of UV/VIS Spectra in Solution. *J. Phys. Chem.* **1996***, 100*, 3349-3353.

(25) Broo, A., Pearl, G., Zerner, M. C. Development of a Hybrid Quantum Chemical and Molecular Mechanics Method with Application to Solvent Effects on the Electronic Spectra of Uracil and Uracil Derivatives. *J. Phys. Chem. A* **1997***, 101*, 2478-2488.

(26) Neto, J. D., Zerner, M. C. New Parametrization Scheme for the Resonance Integrals ($H_{\mu\nu}$) Within the INDO/I Approximation. Main Group Elements. *Int. J. Quantum Chem.* **2001***, 81*, 187-201.

(27) Karelson, M. Theoretical Treatment of Solvent Effects on Electronic and Vibrational Spectra of Compounds in Condensed Media, in: *Handbook of Solvents*, Ed. by G. Wypych, ChemTec Publishing, Toronto – New York, 2001, pp. 639-679.

(28) Katritzky, A. R., Lobanov, V.S., Karelson, M. *CODESSA, Reference Manual*; University of Florida: Gainesville.

(29a) Katritzky, A. R.; Mu, L.; Lobanov, V. S.; Karelson, M. Correlation of Boiling Points with Molecular Structure. 1. A Training Set of 298 Diverse Organics and a Test Set of 9 Simple Inorganics. *J. Phys. Chem.* **1996**, *100*, 10400-10407.

(29b) Katritzky, A. R.; Lobanov, V. S.; Karelson, M. Normal Boiling Points for Organic Compounds: Correlation and Prediction by a Quantitative Structure-Property Relationship. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 28-41.

(29c) Katritzky, A. R.; Ignatchenko, E. S.; Barcock, R. A.; Lobanov, V. S.; Karelson, M. Prediction of gas-chromatographic retention times and response factors using a general quantitative structure-property relationship treatment. *Anal. Chem.* **1994**, *66*, 1799-1907.

(29d) Lučić, B.; Trinajstić, N.; Sild, S.; Karelson, M.; Katritzky, A. R. A New Efficient Approach for Variable Selection Based on Multiregression: Prediction of Gas Chromatographic Retention Times and Response Factors. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 610-621.

(29e) Huibers, P. D. T.; Lobanov, V. S.; Katritzky, A. R.; Shah, D. O.; Karelson, M. Prediction of Critical Micelle Concentration Using a Quantitative Structure-Property Relationship Approach. 1. Nonionic Surfactants. *Langmuir* **1996**, *12*, 1462-1470.

(29f) Huibers, P. D. T.; Lobanov, V. S.; Katritzky, A. R.; Shah, D. O.; Karelson, M. Prediction of Critical Micelle Concentration Using a Quantitative Structure-Property Relationship Approach. 2. Anionic Surfactants. *Journal of Colloid and Interface Science* **1997**, *187*, 113-120.

(29g) Katritzky, A. R.; Mu, L.; Karelson, M. A QSPR Study of the Solubility of Gases and Vapors in Water. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 1162-1168.

(29h) Katritzky, A. R.; Tatham, D. B.; Maran, U. Correlation of the Solubilities of Gases and Vapors in Methanol and Ethanol with Their Molecular Structures. J. Chem. Inf. Comput. Sci. **2001**, *41*, 358-363.

(29i) Katritzky, A. R.; Rachwal, P.; Law, K. W.; Karelson, M.; Lobanov, V. S. Prediction of Polymer Glass Transition Temperatures Using a General Quantitative Structure-Property Relationship Treatment. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 879-884.

(29j) Katritzky, A. R.; Sild, S.; Lobanov, V.; Karelson, M. Quantitative Structure-Property Relationship (QSPR) Correlation of Glass Transition Temperatures of High Molecular Weight Polymers. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 300-304.

(29k) Katritzky, A. R.; Sild, S.; Karelson, M. General Quantitative Structure-Property Realtionship Treatment of the Refractive Index of Organic Compounds. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 840-844.

(29l) Katritzky, A. R.; Sild, S.; Karelson, M. Correlation and Prediction of the Refractive Indices of Polymers by QSPR. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 1171-1176.

(29m) Katritzky, A. R.; Chen, K.; Wang, Y.; Karelson, M.; Lucic, B.; Trinajstic, N.; Suzuki, T.; Schüürmann, G. Prediction of liquid viscosity for organic compounds by a quantitative structure-property relationship. *J. Phys. Org. Chem.* **2000**, *13*, 80-86.

(29n) Katritzky, A. R.; Perumal, S.; Petrukhin, R. A QSPR Treatment of Solvent Effects on the Decarboxylation of 6-Nitrobenzisoxazole-3-carboxylates Employing Molecular Descriptors. *J. Org. Chem.* **2001**, *66*, 4036-4040.

(29o) Katritzky, A. R.; Maran, U.; Lobanov, V. S.; Karelson, M. Structurally Diverse Quantitative Structure-Property Relationship Correlations of Technologically Relevant Physical Properties. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1-18.

(29p) Karelson, M.; Maran, U.; Wang, Y.; Katritzky, A. R. QSPR and QSAR models derived using large molecular descriptor spaces. A review of Codessa applications. *Collect. Czech. Chem. Commun.* **1999**, *64*, 1551-1571.

(30) Fitch, W. L. in. *UV-Visible spectrophotometry of water and wastewater*; Thomas, O., Burgess, C., Ed.; Elsevier, 2001.

(31) Perkampus, H. H. UV-VIS Atlas of Organic Compounds; second ed.; VCH: Weinheim, 1992.

(32) Schrödinger, Inc. World Wide Web, http://www.schrodinger.com/

(33) Molecular Networks GmbH. World Wide Web, http://www.mol-net.de/

(34) Kassel, D. B. Combinatorial chemistry and mass spectrometry in the 21st century drug discovery laboratory. *Chem. Rev.* **2001***, 101*, 255-267.

(35) Espinosa, S., Bosch, E., Roses, M. Retention of ionizable compounds on HPLC. 5. pH Scales and the retention of acids and bases with acetonitrile-water mobile phases. *Anal. Chem.* **2000***, 72*, 5193-5200.

(36) Myers, A. B.; Birge, R. R. The effect of solvent environment on molecular electronic oscillator strengths. *J. Chem. Phys.* **1980***, 73*, 5314-5321.

(37) Aminabhavi, T. M., Gopalakrishna, B. Density, viscosity, refractive index, and speed of sound in aqueous mixtures of N,N-dimethylformamide, dimethyl sulfoxide, N,N-dimethylacetamide, acetonitrile, ethylene glycol, diethylene glyol,

1,4-dioxane, tetrahydrofuran, 2-methoxyethanol, and 2-ethoxyethanol at 298.15 K. *J. Chem. Eng. Data* **1995***, 40*, 856-861.

(38) Lang, L. Absorption spectra in the ultraviolet and visible region.; Academic Press and Robert E. Krieger,: New York, 1961-1982.

(39) Burgers, J., Hoffnagel, M.A., Verkade, P.E., Visser, H., Wepster, B.M. Steric effects on mesomerism. XVII. Some properties of aromatic nitro-, amino- and acylamino compounds with bulky substituents. *Rec. Trav. Chim*. **1958**, *77*, 491-530.

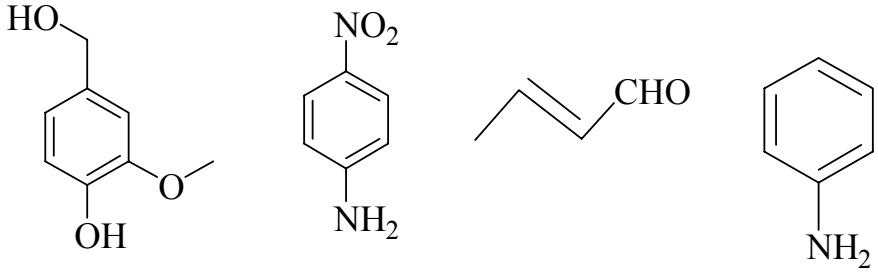
(40) Braude, E. A. Studies in light absorption. Part VIII. Dibenzyl and stilbene derivatives. Interaction between unconjugated chromophores. *J. Chem. Soc.* **1949**, 1902-1909.

LEGENDS TO FIGURES

1. Example UV spectra from the database. Each spectrum was collected in heptane. a) 1,3-pentadiene, b) pyridine –2-carboxaldehyde, c). 2-iodopropane.
2. Correlation of integrated area under the wavelength curve vs integrated area under the frequency curve.
3. Correlation of integrated area and CODESSA non-quantum descriptors; equation 13, table 4.
4. Correlation of integrated area and CODESSA quantum descriptors; equation 14
5. Correlation of integrated area and CODESSA quantum descriptors, excluding spectra with end absorption only ; equation 15

6. Correlation of integrated area and CODESSA quantum descriptors for 75 spectra recorded in nonpolar solvents
7. Correlation of integrated area and CODESSA quantum descriptors for 93 spectra taken in ethanol
8. Simulated UV spectra showing a widthless absorbance as predicted in ZINDO (a); peaks broadened with a fixed peak width (b). and a simulated spectrum truncated at 220 and 360 nm (c)
9. Correlation of integrated area and ZINDO predicted UV absorbance for 205 neutral molecules.
10. Correlation chart for the 3-descriptor model with ZINDO oscillator strengths as additional descriptors.

Table 1. Solvent effects on relative integrated UV extinction from 220-360 nm



Solvent				
EtOH	1.13	.74	3.09	-
CH_3CN	1.14	-	-	-
Hexane	1	1	1	-
H_2O, pH=2	1.08	1.3	-	1
H_2O, pH=8	-	-	-	88

Table 2. Effect of steric crowding on maximum UV absorbance.

Compound	ε_{max}
Aniline	9130
N,N-dimethylaniline	15500
2-(t-butyl)aniline	7850
2-t-butyl-N,N-dimethylaniline	630

Table 3. Effect of conjugation and hyperconjugation on extinction at 220 and 250 nm

Compound	Extinction at	
	220	250
Toluene	1900	69
Biphenyl	4500	16000
Diphenylmethane	8300	380
1,2-Diphenylethane	6300	260

Table 4. Statistical parameters of seven best correlations for the full dataset without quantum chemical descriptors (Eq. 13).

Number of descriptors	R^2	R^2_{cv}	F	s
1	0.3728	0.3626	307.23	$5.04 \cdot 10^5$
2	0.6034	0.5970	392.48	$4.02 \cdot 10^5$
3	0.6343	0.6249	297.71	$3.86 \cdot 10^5$
4	0.6644	0.6543	254.34	$3.70 \cdot 10^5$
5	0.6816	0.6724	219.68	$3.61 \cdot 10^5$
6	0.6945	0.6815	194.01	$3.54 \cdot 10^5$
7	0.7063	0.6923	175.54	$3.47 \cdot 10^5$

Table 5. Statistical parameters of 7 best correlations for the full dataset with quantum chemical descriptors (Eq. 14).

Number of descriptors	R^2	R^2_{cv}	F	s
1	0.4730	0.4639	464.10	$4.62 \cdot 10^5$
2	0.6035	0.5941	392.64	$4.01 \cdot 10^5$
3	0.7342	0.7272	474.24	$3.29 \cdot 10^5$
4	0.7618	0.7467	410.96	$3.12 \cdot 10^5$
5	0.7838	0.7653	371.93	$2.97 \cdot 10^5$
6	0.8037	0.7870	349.39	$2.84 \cdot 10^5$
7	0.8152	0.7996	322.06	$2.75 \cdot 10^5$

Table 6. Statistical parameters of 5 best correlations for the dataset of 255 compounds with at least 40 % of their UV absorbance above 250 nm (Eq. 15).

Number of descriptors	R^2	R^2_{cv}	F	s
1	0.4674	0.4503	219.38	$4.86 \cdot 10^5$
2	0.6061	0.5653	191.60	$4.18 \cdot 10^5$
3	0.7162	0.6277	208.59	$3.56 \cdot 10^5$
4	0.7181	0.6284	157.33	$3.55 \cdot 10^5$
5	0.7426	0.7177	141.98	$3.40 \cdot 10^5$

Table 7. Five-parameter correlation for the nonpolar subset of 75 compounds

$X \pm \Delta X$	t-test	name of the descriptor
$(-2.09 \pm 0.26)\cdot 10^6$	-4.3118	Intercept
$(2.34 \pm 0.28)\cdot 10^5$	8.45	Number of benzene rings
$(2.54 \pm 0.45)\cdot 10^6$	5.6434	Average bond order for atom C
(7.02 ± 1.36)	5.1577	(1/6) X Gamma polarizability
$-(4.72 \pm 0.96)\cdot 10^5$	-4.9219	LUMO energy
$-(2.18 \pm 0.67)\cdot 10^6$	-3.2329	RPCG relative positive charge

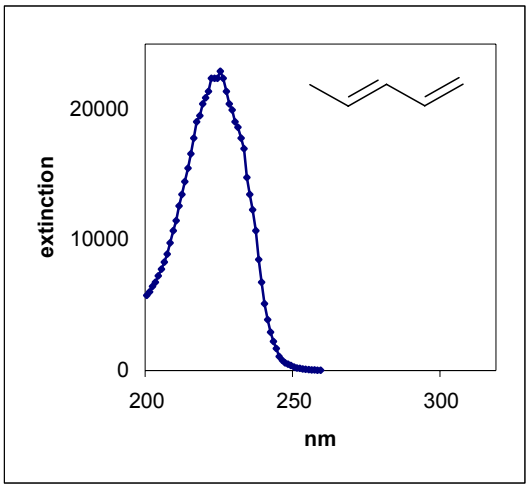
Table 8. Seven-parameter correlation for the ethanol subset of 93 structures.

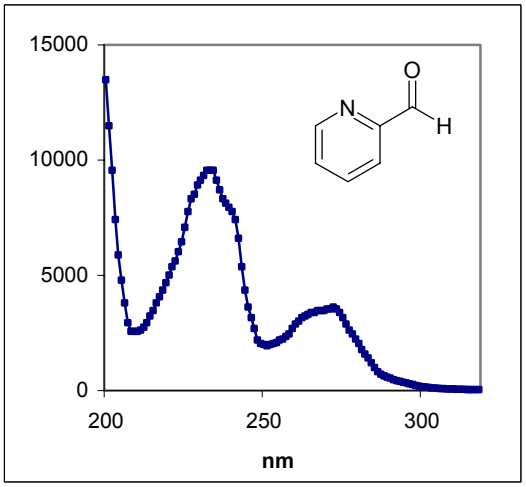
$X \pm \Delta X$	t-test	name of the descriptor
$-(3.03 \pm 1.16)\cdot 10^{6}$	-2.6029	Intercept
$(4.38 \pm 0.59)\cdot 10^{1}$	7.4540	(1/6) X Gamma polarizability
$(2.72 \pm 0.44)\cdot 10^{6}$	6.2394	Average bond order for atom C
$-(3.07 \pm 0.62)\cdot 10^{5}$	-4.9532	HOMO-LUMO energy gap
$(1.57 \pm 0.35)\cdot 10^{5}$	4.5053	Number of rings
$(1.40 \pm 0.32)\cdot 10^{5}$	4.3099	Number of benzene rings
$(4.41 \pm 1.24)\cdot 10^{5}$	3.5299	Minimum atomic state energy for atom H
$-(3.83 \pm 1.20)\cdot 10^{3}$	-3.1930	Complementary information content (order 2)

Table 9. Statistics of best correlations with oscillator strengths as additional descriptors.

Number of descriptors	R^2	R^2_{cv}	F	s
1	0.7058	0.7027	1098.97	$3.53 \cdot 10^5$
2	0.7635	0.7582	737.51	$3.17 \cdot 10^5$
3	0.8317	0.8143	751.13	$2.68 \cdot 10^5$
4	0.8496	0.8351	642.36	$2.54 \cdot 10^5$
5	0.8573	0.8431	545.47	$2.47 \cdot 10^5$

Fig. 1.


20000
10000
0
extinction
200
250
300
nm


15000
10000
5000
0
200
250
300
nm
N
O
H

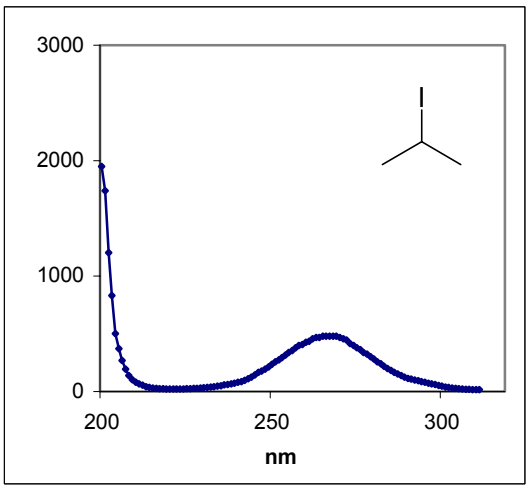

3000
2000
1000
0
200
250
300
nm
I

Fig 2.

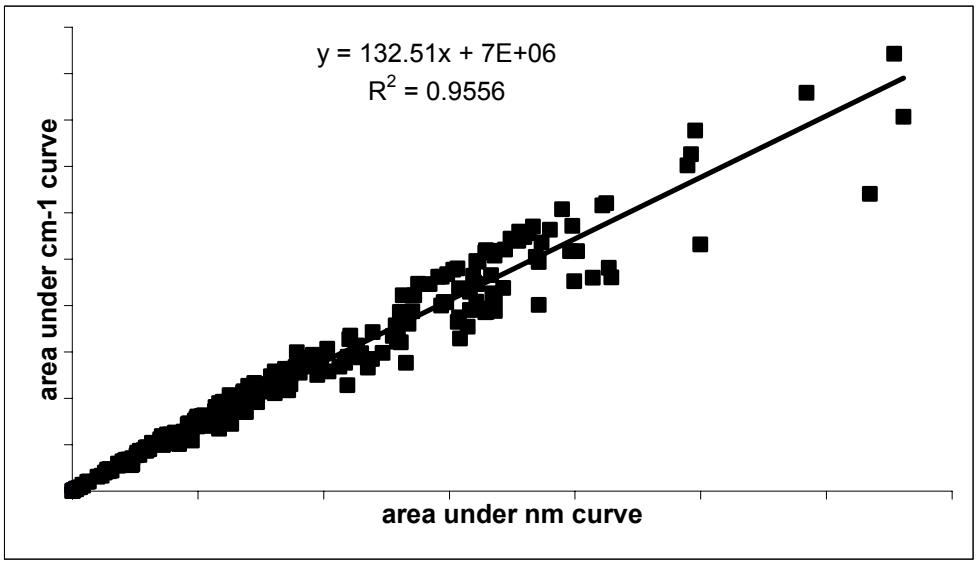

y = 132.51x + 7E+06
R2 = 0.9556
area under cm-1 curve
area under nm curve

Fig. 3. (table 4, eq. 13)

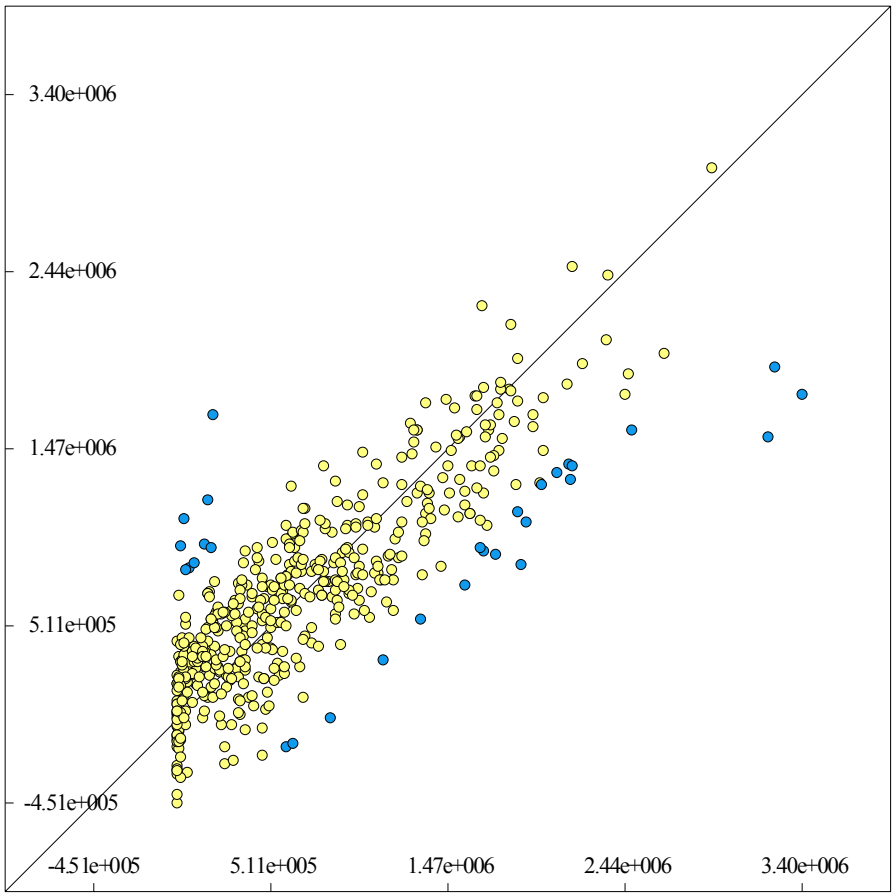


Fig. 4. (table 5, eq. 14)

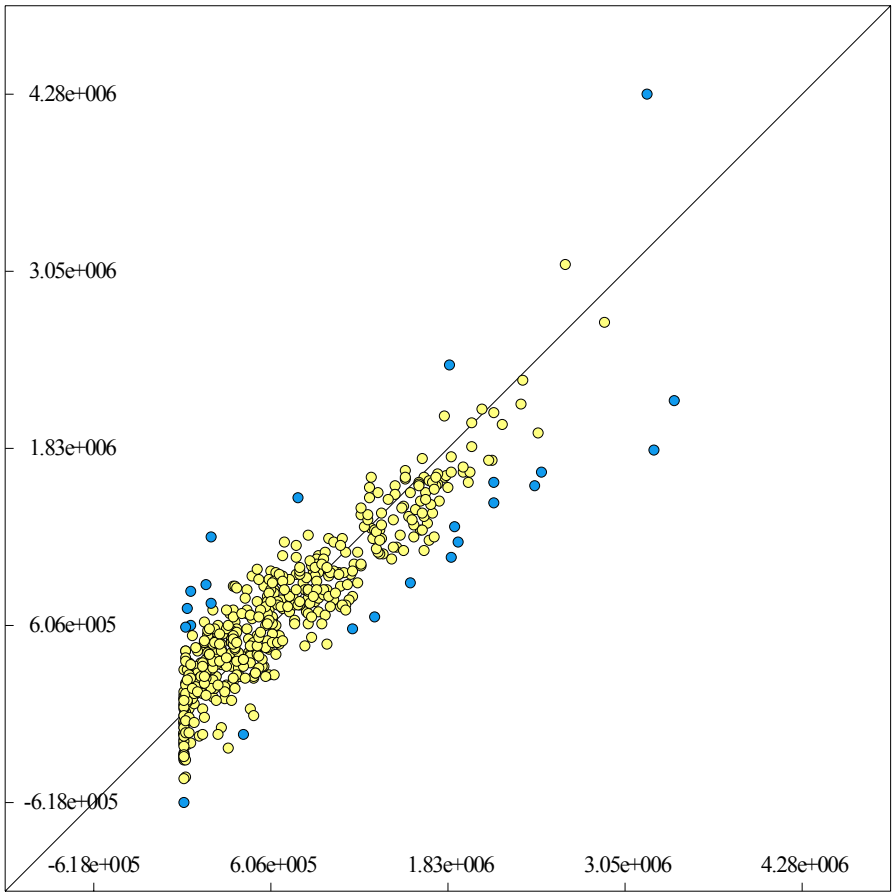


Fig. 5. (table 6, eq. 15)

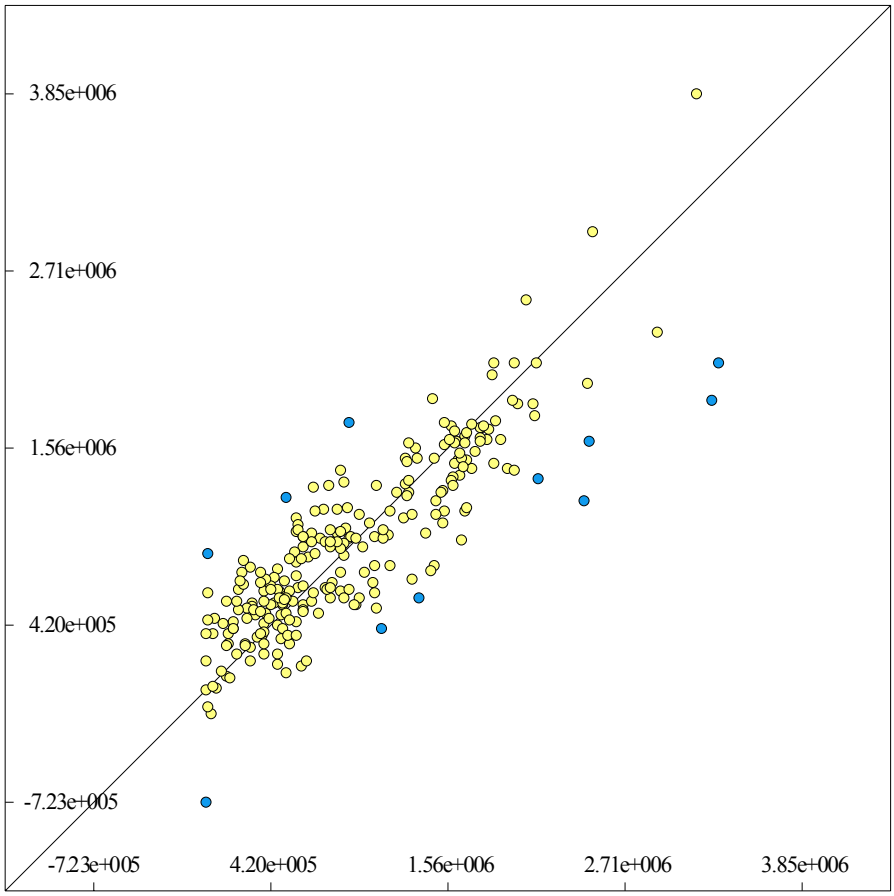


Fig. 6. Observed *vs* predicted chart of 5-parameter equation for the nonpolar subset (Table 7).

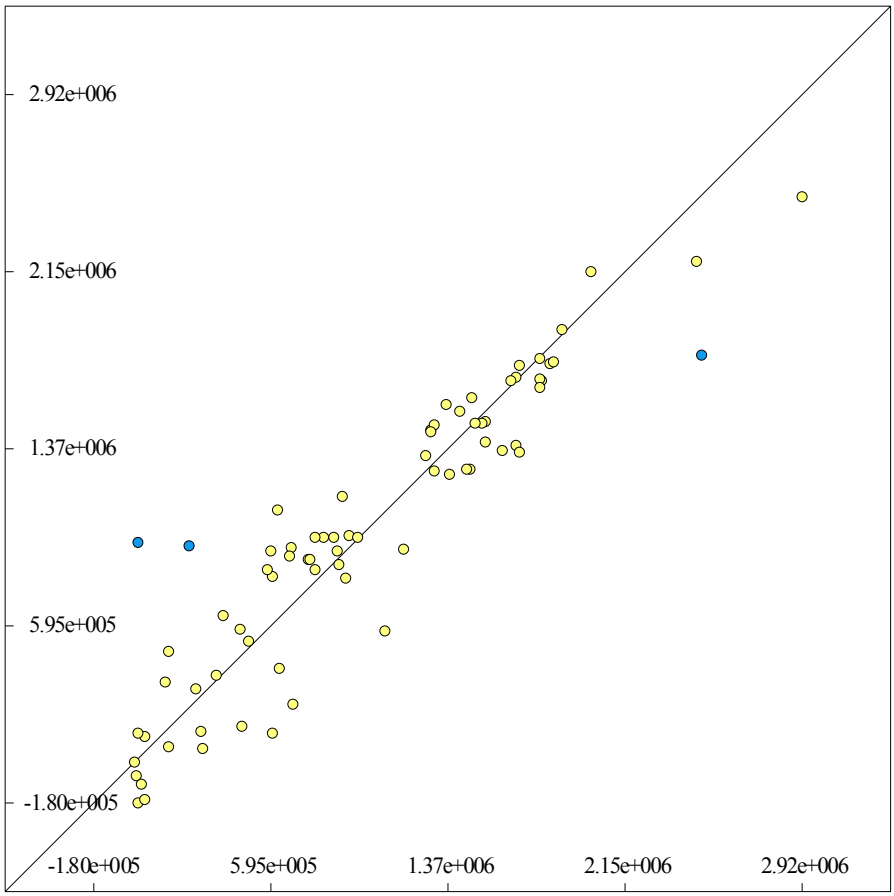


Fig. 7. Observed *vs* predicted chart of 7-parameter equation for the ethanol subset (table 8).

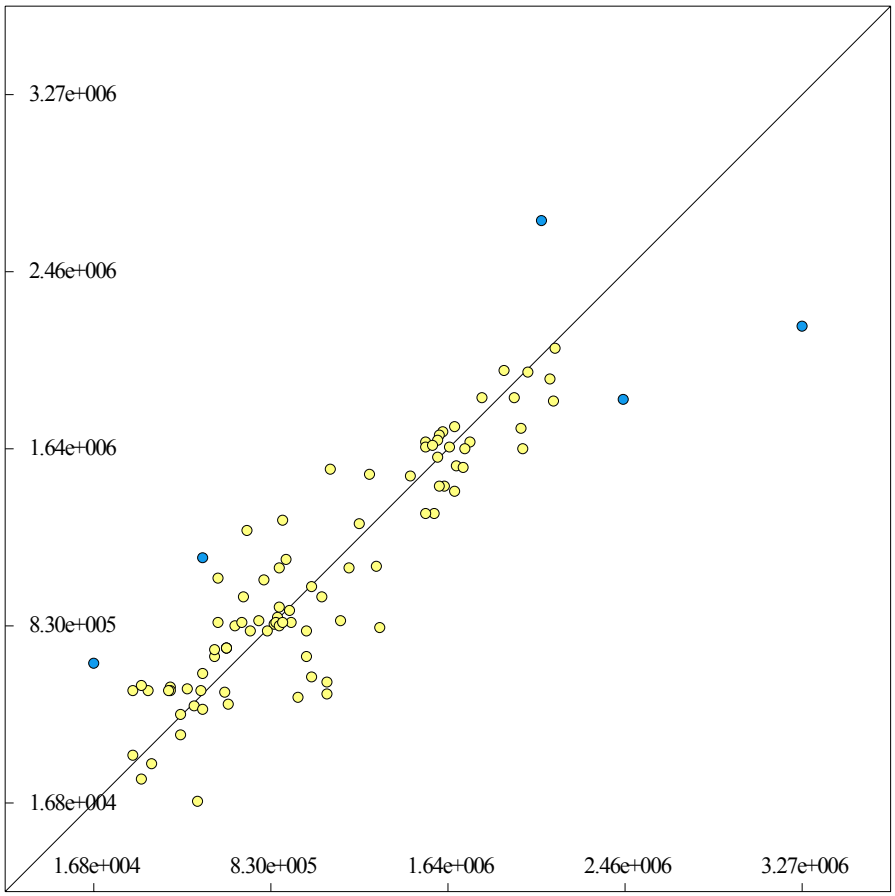
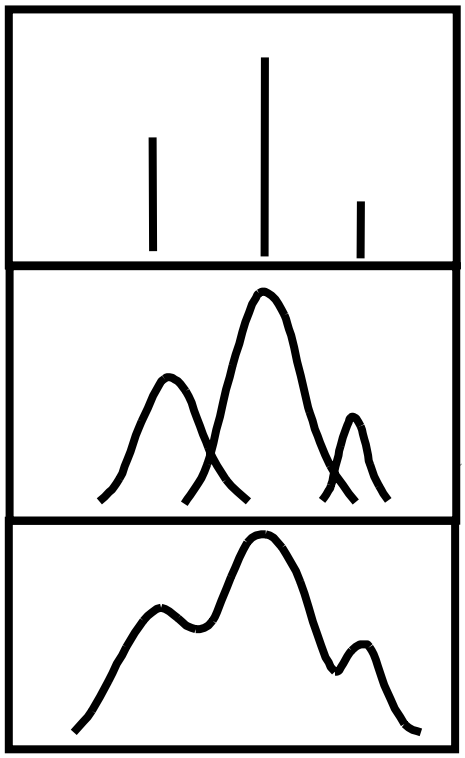


Fig. 8.



a)

b)

c)

Fig. 9.

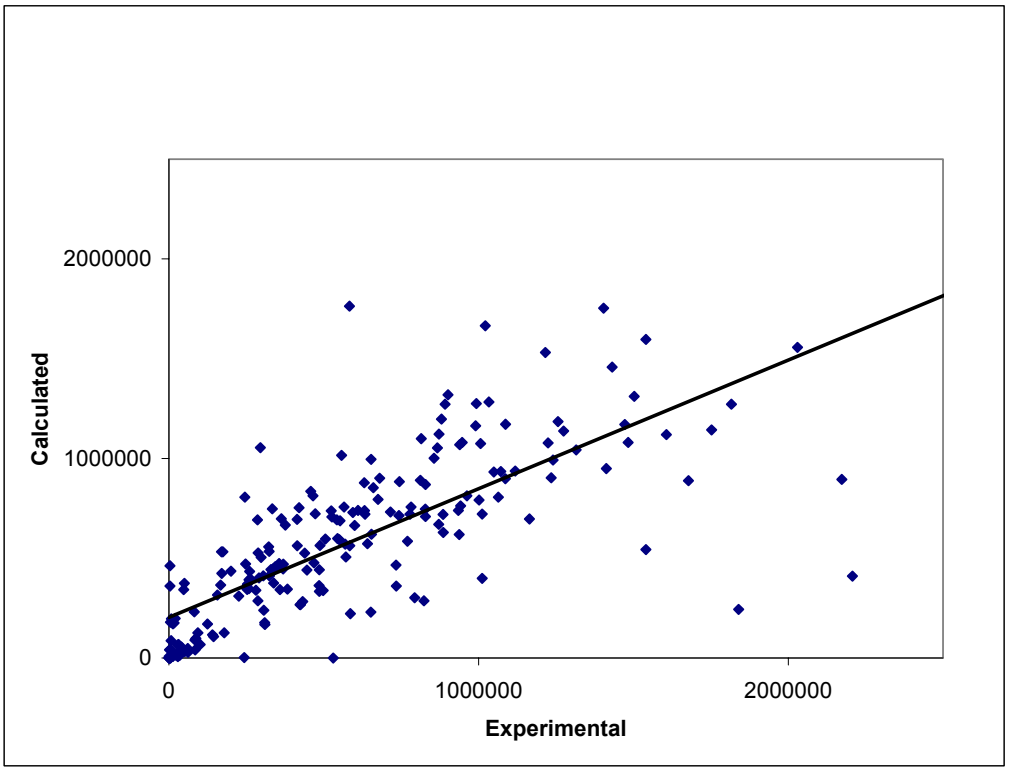

2000000
1000000
0
Calculated
0
1000000
2000000
Experimental

Fig. 10. Correlation chart for the 5-descriptor model with ZINDO oscillator strengths as additional descriptors (Eq. 16, Table 9).