Available online at www.sciencedirect.com

ScienceDirect

Mendeleev Commun., 2023, **33**, 419–421

Mendeleev Communications

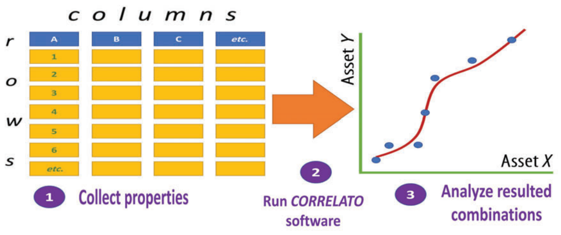
An efficient method of searching for correlations between unlimited datasets to provide forecasting models

Alexander Yu. Tolbin

Institute of Physiologically Active Compounds, Federal Research Center of Problems of Chemical Physics and Medicinal Chemistry, Russian Academy of Sciences, 142432 Chernogolovka, Moscow Region, Russian Federation. E-mail: tolbin@ipac.ac.ru

DOI: 10.1016/j.mencom.2023.04.038

This work presents a highly powerful and very simple statistical method for creating predictive models based on structure–property relationships. The algorithm was implemented in the software CORRELATO, which demo version is presented in Online Supplementary Materials. This algorithm was tested on a small series of phthalocyanine dyes to search for the relationship between optical limiting effect and quantum chemical descriptors responsible for the nonlinear optical properties of these absorbers.



Keywords: correlations, phthalocyanine, J-dimers, optical limiting, forecasting models.

Within the framework of fundamental research, the search for quantitative relationships between the elements of any complicated system allows one to create forecasting models to achieve a specific result without going deep into routine procedures, which is necessary for efficient use of resources. In the field of chemical sciences, in addition to the development of synthetic strategies and the accumulation of data on a number of practically important characteristics of objects, there is a search for correlations between structures of substances and their properties.[1] On the bases of such correlations, it becomes possible not only to predict biological activity using QSAR/QSPR methods (quantitative structure–activity/quantitative structure–property relationship),[2] but also to create entire synthetic procedures parameterized in terms of qualitative or quantitative structure–reactivity relationship (QSRR).[3] In the described and similar approaches, a general complex problem is solved, which has the following trivial representation:

$$\text{Output} = \text{Function of Dataset.} \tag{1}$$

The search for internal relationships (Function) within a lot of different parameters (Dataset) that describe the system allows one to reveal: 1) what influences what and 2) is it possible to predict any property based on a set of molecular descriptors? The solution of the proposed problem becomes possible using statistical methods that involve a set of machine learning algorithms within the framework of artificial intelligence. Thus, we are dealing with the parametric analysis, which is a part of logical statistics exploring the cumulative influence of various parameters of a system on solving a specific problem.[4] The market offers a lot of commercial solutions for statistical data analysis. Particularly, ALTAIR[5] and DATADVANCE[6] allow engineers to comprehensively implement parametric analysis for designing materials for industries such as aerospace, oil and gas, automotive, batteries and fuel cells, *etc.*, as well as to explore their strength[7] and to improve the efficiency of various devices.[8] It should be noted that commercial programs for multiparametric analysis are expensive and sometimes inaccessible to most scientists.

This paper describes a simple, reliable, and high-performance algorithm that allows researchers without special knowledge in the field of mathematical statistics and the logic of creating machine learning algorithms to find relationships between the properties of chemical objects, which can include quantum-chemical descriptors along with spectral response or biological activity. For this purpose, the software CORRELATO[9] was created. The description of this program is given in Online Supplementary Materials.

The developed algorithm is based on the statistical analysis of data summarized in a table. The rows and columns of this table correspond to a set of chemicals and their properties, respectively. Using such a table, at each iteration, two arrays Y and X are formed: $Y = \{Y_n\}_{n=1}^{N}$ and $X = \{X_n\}_{n=1}^{N}$, which are sequences with an equal number of elements N corresponding to the number of rows (or substances in the table). Through a random combination of values at the intersection of the current n^{th} row with selected columns, the following calculations are performed at each iteration:

$$Y_n = \prod_i y_i^{\,a} \ \text{ and } \ X_n = \prod_j x_j^{\,b}, \tag{2}$$

where y_i and x_j are the properties of the system (real values; for example, experimental photocatalytic activity or calculated band gap[10]); i and j are the numbers of columns in the table (integers; for example, 1, 2, 3, *etc.*); and a and b are the degrees that are randomly selected from the limited set of real numbers specified by the researcher. Now it is needed to estimate how the obtained sequences $\{Y_n\}$ and $\{X_n\}$ correlate with each other. The simplest way to do this is to calculate the r-Pearson coefficient:

$$r_j = \frac{\sum_{i=1}^{N}(X_i-\bar{X})(Y_i-\bar{Y})}{[\sum_{i=1}^{N}(X_i-\bar{X})^2(Y_i-\bar{Y})^2]^{1/2}}, \tag{3}$$

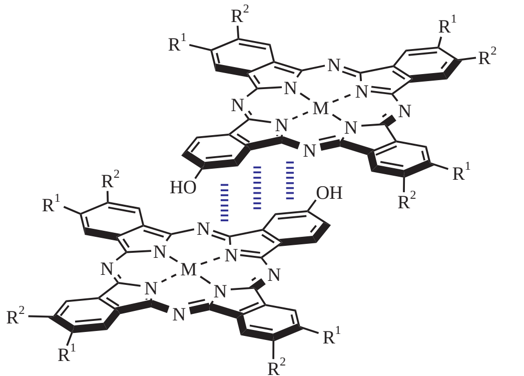
where i is the index for the Y and X arrays; j is the iteration number (the total number of iterations is set by the researcher);

and $\bar{X}$ and $\bar{Y}$ are the arithmetic averages of X and Y arrays, respectively. The Pearson correlation coefficient r can have values from +1 to −1. The stronger the relationship between data sets (the points are near the line), the greater the r value, while the r value close to zero indicates the absence of correlation (the points are scattered on the plane). When implementing the described numerical algorithm, iterations with low ($r < 0.5$) and negative r values are ignored, and among the rest a series is selected having r values corresponding to the task (recommended threshold: $r > 0.85$). As part of the optimization of solutions, it is possible to apply additional conditions, such as selection by the value of the R^2 factor when fitting the values of a user-defined function. As a result, the user has a series of 2D diagrams, and the option for detailed analysis allows one to identify the so-called ‘outliers’ which are the elements that prevent convergence to the specified criteria. These can be erroneous signs (columns) or experimental errors (rows).[9]

The developed algorithm is very fast. For instance, 10 000 iterations for a limited series of five phthalocyanines **1a–e** have taken about 1 minute on a single processor. Dyes **1a–e** belong to stable J-type dimers, which demonstrate the nonlinear attenuation of nanosecond pulsed laser beams.[11]

The nonlinear attenuation of high-intensity light to a safe level underlies the creation of optical limiters represented as the effective means of protection against laser radiation damage.[12] To date, there are no specific approaches to a comprehensive assessment of all parameters of optical limiting in total as well as to the selection of absorbers according to their efficiency in such a process.

Table 1 shows the experimental data (β_{TPA}, E_0, DR, k_A) and quantum chemical parameters (E_g, μ, α, β, γ) calculated by the FF-DFT method (FF = Finite external electric Field).[11] The field value of 0.001 a.u. was used to calculate α, β, and γ with GAMESS (US) software[13].



R¹ and R² = H *vs.* Buᵗ;
M = Mg (**a**), Zn (**b**), Cu (**c**), Ni (**d**), Co (**e**)

1a–e

Figure 1 Chemical structures of phthalocyanine J-dimers **1a–e**. Dashed lines show interactions between the macrocycles.

Table 1 Selected nonlinear optical parameters of the phthalocyanines **1a–e**.[a]

Dye	β_{TPA}/ cm GW^{-1}	E_0/ J cm^{-2}	DR	k_A	μ/D	E_g/ eV	α/Å^3	−lgβ	−lg|γ|
1a	340	0.05	460	7	2.15	1.37	126.6	28.87	34.00
1b	360	0.03	830	7.8	1.94	1.36	126.9	28.64	35.89
1c	228	0.03	930	6.4	1.82	1.38	116.6	28.68	34.85
1d	57	0.65	72	3.2	1.80	1.39	114.8	28.74	35.44
1e	21	0.6	82	1.7	1.82	1.42	114.1	29.05	35.64

[a] β_{TPA} is the two-photon absorption coefficient; E_0 is the limiting threshold; DR is the dynamic range; k_A is the attenuation factor; μ is the dipole moment; E_g is the band gap; α is the linear polarizability; β is the first hyperpolarizability (esu); γ is the second hyperpolarizability (esu).

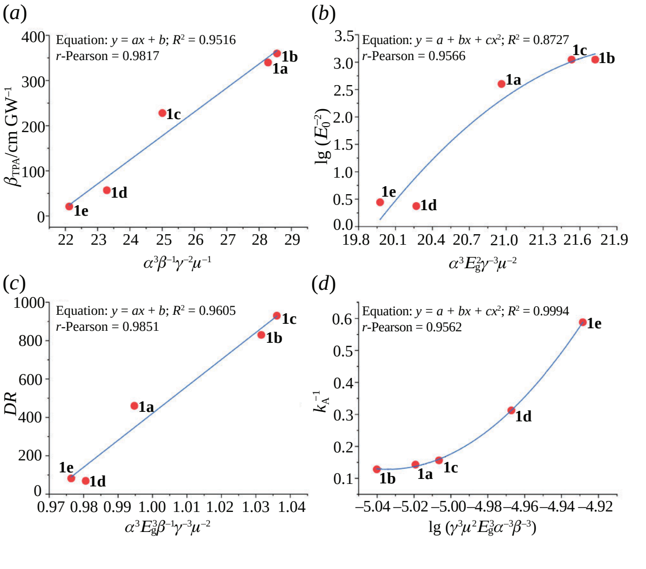


Figure 2 Solving a multiparametric problem [equations (2) and (3)] to find the relationship between the experimental values of optical limiting (Y asset) and quantum chemical descriptors (X asset) for phthalocyanine J-dimers **1a–e**: (*a*) two-photon absorption coefficient; (*b*) limiting threshold; (*c*) dynamic range; (*d*) attenuation factor. Notation changes: β: = −lgβ; γ: = −lg|γ|. The mesurement units of the quantities are given in Table 1.

The task is to find relationships between experimental and theoretical data for predicting the optical limiting effect based on the structure of dyes (*i.e.*, without conducting an experiment). Recently it was shown how information about the excitation of dye molecules with electric fields can be used to predict the effectiveness of the optical limiting response.[14]

The results of solving the multiparametric task based on the data collected in Table 1 are shown in Figure 2.

As can be seen from Figure 2, despite the high Pearson coefficient ($r > 0.95$), the ‘experimental *vs.* theory’ relationship is sometimes better described by a non-linear function. Thus, for E_0 and k_A, polynomials turned out to be the most preferable [Figure 2(*b*),(*d*)]. The fact that such complicated correlations can be found in a very simple way indicates the high importance of the developed algorithm for providing the forecasting models. But even now, using the dependences (Figure 2), one can conclude that nickel **1d** and cobalt **1e** complexes are the least productive absorbers in optical limiting. Magnesium complex **1a** is likely to fill an intermediate shelter when considering only the dynamic range and limiting threshold, respectively. Finally, zinc complex **1b** can be considered the most demanded in terms of all characteristics, and it is expedient to take it as a reference in the future. Continuing these considerations, a universal formula for the efficiency of optical limiting can be derived (Figure 3).

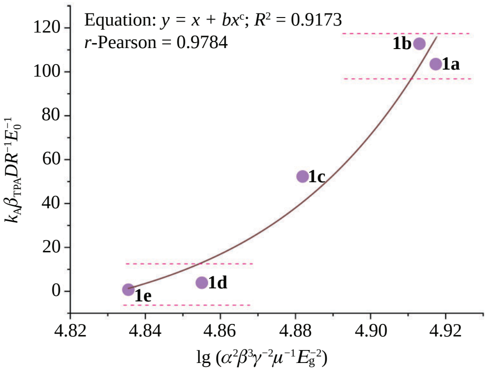


Figure 3 Estimation of the overall efficiency of optical limiting for the phthalocyanine J-dimers **1a–e** based on multiparametric analysis [equations (2) and (3)]. Notation changes: β: = −lgβ; γ: = −lg|γ|. The measurement units of the quantities are given in Table 1.

One of the most successful solutions to the multiparameter problem [equations (2) and (3)] has allowed us to use the following combination of parameters (*Y* axis in Figure 3) as the optical limiting efficiency functional:

$$\sigma = \frac{k_A \beta_{TPA}}{DR\,E_0}. \tag{4}$$

The efficiency of optical limitation increases with an increase in the nonlinearity of the medium (β_{TPA}) and the degree of the laser radiation energy attenuation (k_A), while the parameters *DR* and E_0 are both interconnected with each other and with those variables in the numerator of the fraction (4). In this case, $\sigma \to \infty$. ‘Bad’ absorber dyes exhibit low k_A and β_{TPA} values, along with a narrow dynamic range and a high threshold of activation. This leads to the fact that $\sigma \to 0$.

As can be seen from Figure 3, copper complex **1c** occupies an intermediate position ($\sigma \approx 50$). Magnesium and zinc complexes **1a**,**b** can be considered equal. However, expression (3) does not contain a linear absorption coefficient associated with the sample concentration, since it was the same (1.44 cm^{-1}) for dyes **1a–e** at fixed transmission (0.75) and optical layer thickness (0.2 cm). However, to flexibly adjust the predictive model, it is advisable to consider the mentioned feature.

In conclusion, the simplest algorithm for searching for the correlations between unlimited data sets has been created based on the solution of a multiparametric analysis of the structure–property relationship. As the input data, real values are used that characterize the system according to a set of features chosen by the researcher. An analysis of the obtained correlations has allowed us to derive the analytical form of the functional characterizing the total efficiency of optical limiting to sort the absorbers within the considered series.

The developed algorithm is implemented as a computer program CORRELATO (see Online Supplementary Materials) and does not require special knowledge in the field of mathematical statistics and related disciplines. A practical solution to any problem is possible: the result is limited only by the imagination of the researcher in terms of choosing the properties of a chemical system. Now the development of forecasting models has become available to a wide range of experimenters. The described methodology can be used for a wide range of chemists to solve a key problem *i.e.* the search for the structure of a substance with a given activity or any characteristics.

The development of the analytical algorithm and the creation of software CORRELATO[9] were financially supported by the Russian Science Foundation (grant no. 21-73-20016). The solution of practical multiparametric tasks was carried out within the framework of the State Assignment no. FFSN-2021-0003 (‘Creation of compounds with given physicochemical properties’) at the Joint Supercomputing Center of the Russian Academy of Sciences (www.jscc.ru).

Online Supplementary Materials

Supplementary data associated with this article can be found in the online version at doi: 10.1016/j.mencom.2023.04.038.

References

1 I. Yu. Titov, V. S. Stroylov, P. V. Rusina and I. V. Svitanko, *Russ. Chem. Rev.*, 2021, **90**, 831.

2 M. G. Medvedev, O. V. Stroganov, A. O. Dmitrienko, M. V. Panova, A. A. Lisov, I. V. Svitanko, F. N. Novikov and G. G. Chilov, *Mendeleev Commun.*, 2022, **32**, 735.

3 T. I. Madzhidov, A. Rakhimbekova, V. A. Afonina, T. R. Gimadiev, R. N. Mukhametgaleev, R. I. Nugmanov, I. I. Baskin and A. Varnek, *Mendeleev Commun.*, 2021, **31**, 769.

4 D. Nordstokke, in *Encyclopedia of Quality of Life and Well-Being Research*, ed. A. C. Michalos, Springer, Dordrecht, 2014, pp. 4584–4588.

5 Web reference: https://www.altair.com.

6 Web reference: https://www.datadvance.net.

7 (*a*) K. N. Chitra, R. M. Abhilash, S. S. Chauhan, G. S. Venkatesh and N. D. Shivkumar, *AIP Conf. Proc.*, 2018, **1943**, 020126; (*b*) M. E. García-Sánchez, J. A. Perez-Naitoh, D. E. Ramirez-Arreola, J. R. Robledo-Ortíz, P. Ortega-Gudiño and I. Jiménez-Palomar, *MRS Adv.*, 2016, **1**, 2161; (*c*) O. I. Kazakova, I. Yu. Smolin and I. M. Bezmozgiy, *AIP Conf. Proc.*, 2017, **1909**, 020081.

8 V. Ghungarde, S. Awachar, N. K. Vaidya and T. Jagadeesha, *IOP Conf. Ser.: Mater. Sci. Eng.*, 2019, **624**, 012023.

9 A. Yu. Tolbin, *Establishing Correlations Between Unlimited Datasets – Correlato*, Certificate of state registration of computer program No. 2022613888 (RU), 2022.

10 L. A. Lebedev, M. I. Chebanenko, E. V. Dzhevaga, K. D. Martinson and V. I. Popkov, *Mendeleev Commun.*, 2022, **32**, 317.

11 A. Yu. Tolbin, M. S. Savelyev, A. Yu. Gerasimenko, L. G. Tomilova and N. S. Zefirov, *Phys. Chem. Chem. Phys.*, 2016, **18**, 15964.

12 G. de la Torre, P. Vázquez, F. Agulló-López and T. Torres, *Chem. Rev.*, 2004, **104**, 3723.

13 I. S. Gerasimov, F. Zahariev, S. S. Leang, A. Tesliuk, M. S. Gordon and M. G. Medvedev, *Mendeleev Commun.*, 2021, **31**, 302.

14 A. Yu. Tolbin, M. S. Savelyev, A. Yu. Gerasimenko and V. E. Pushkarev, *ACS Omega*, 2022, **7**, 28658.

Received: 10th October 2022; Com. 22/7017