New Developments of EDISFAR Programs. Experimental Design in QSAR Practice

Manuel Pastor and Julio Alvarez-Builla*

Department of Organic Chemistry, Alcala University, 28871 Alcalá de Henares, Spain

Received June 21, 1993®

General requirements of experimental design in QSAR studies are revised. Some strategies, incorporated in the computer programs EDISFAR 92, are described and discussed. Through a simulated example, the combined use of factorial and *D*-optimal designs shows how a QSAR model can be obtained from a reduced exploring set of polysubstituted derivatives.

INTRODUCTION

A quantitative structure–activity relationships (QSAR) study is one of the most useful methodologies in drug design. It involves different steps, including experimental design, molecular description, and model analysis and validation. In recent years a big effort has been devoted to the development of new sophisticated molecular description approaches[1,2] and powerful model generation techniques.[3-5] However, much less attention has been paid to series design techniques.

The first step in any QSAR study is the design of the exploring series over which all the rest of the work is to be performed. This series must be regarded as a sample representative of a population of lead analogues. All knowledge about the products behavior comes from this sample. On the other hand, to meet economical requirements implies that such selection must be performed to obtain the maximum information within the minimum of experimental effort. Unplanned series may be (a) nonrepresentative, thus leading to models invalid for products out of the training set, and/or (b) inefficient, with a poor information/size ratio. As a general rule, no QSAR model is better than the training set it comes from.

It has been suggested that classical design methodologies be used in QSAR. Most of them, however, have been developed in other areas and must be adapted to QSAR peculiarities. In our opinion, there are three main factors that limit its direct application:

1. Chemical Field. Design methods do not work with molecules, but with numerical data. It is difficult to reduce chemicals to numerical data and, when done, the data have some peculiarities compared to a usual data matrix (i.e., different scales, binary and scalar parameters, etc.).

2. QSAR Models. Model structure and even the model existence are unknown. The design method must be adapted to an iterating procedure in which designs should be completed, repaired, changed, etc.

3. Researchers. Drug design is a multidisciplinary field. The researchers involved must deal with chemistry, statistics, pharmacology, computing, etc. Often, the important role of appropriate series design in QSAR is not familiar to the researcher, and the design step is neglected out of ignorance.

Previous efforts to develop statistical design methodologies in QSAR practice come from the pioneering work of Kowalski.[6] From therein, different methods, using chemometric tools, have been described,[7] for example cluster analysis, principal component analysis and multidimensional mapping (PCMM), or Wold's multivariate approach. Most of these methods, developed in the 1970s, suffer from various drawbacks in their practical use, as mentioned above. A more recent approach is the expert system SPECTRE.[8] This program includes factorial and composite designs in a very interactive and easy-to-use interface. It also allows the design of polysubstituted derivatives. The system, however, behaves like a black-box to the user, who is not able to control how the program works. Moreover, the program uses a fixed parameter database, which does not look easy to customize to the requirements of special problems.

Some years ago the authors focused the attention on series design in QSAR, adapting some design methodologies and setting them up in computer programs. The latest result of this effort is version 92 of the EDISFAR programs.[9] The objective is obtaining an efficient tool, making series design in the QSAR area affordable for non-design experts, which should be capable of handling design problems as complex as those involving polysubstitution.

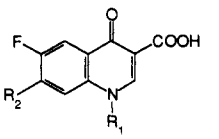
EDISFAR 92 programs are implemented in inexpensive personal computers (MS DOS) and do not require any special equipment. Further details are given in the Appendix.

METHODS

(A) Accommodation to Chemical Field. The target in QSAR experimental design is to obtain a sample (training set) representative of a wide population of products (experimental space, ES). Usually this population corresponds to all possible derivatives of a parent compound whose activity is to be optimized (lead compound).

However, the very first step is reducing ES to a numerical form, which can be handled by mathematical methods. To perform such a transformation, certain molecular properties, which are supposed to be important for activity, had to be represented by numerical variables. For example, in classical extrathermodynamic approach[10] these variables can be *extrathermodynamic parameters* of substituents (such as Hansch π or Hammett σ), experimental data (such as log *P*), or binary variables. The process is named parametrization of products.

Supposing a study is undertaken to obtain the more active **I** derivatives, even if only 10 different substituents were to be varied for each position, ES would include 100 compounds.



I

Then, parametrization would prevent the synthesis of every compound, provided a relationship can be established between

® Abstract published in *Advance ACS Abstracts,* March 15, 1994.

0095-2338/94/1634-0570$04.50/0

Table 1. Parameters Included in the Main EDISFAR Database, MONOBASE

code[a]	name	description	source[b]	no.[c]
Fr	Fr	aliphatic fragmental constant	A	102
π	π	aromatic substituent param	A	217
ES	ES	Taft steric param	A, B	61
L	L	2nd generation STERIMOL, length param	C	286
B1	B1	2nd generation STERIMOL, min width param	C	286
B5	B5	2nd generation STERIMOL, max width param	C	286
l	L	STERIMOL length param	A	131
b1	B1	STERIMOL width param	A	131
b2	B2	STERIMOL width param	A	131
b3	B3	STERIMOL width param	A	131
b4	B4	STERIMOL width param	A	131
MR	MR	molar refractivity	A	386
σ m	σ_m	Hammett const. for meta substitution	A	275
σ p	σ_p	Hammett const for para substitution	A	356
f	$\mathcal{F}$	Swain and Lupton field param	A	259
r	$\mathcal{R}$	Swain and Lupton resonance param	A	259
F o	F_o	Norrington field param for ortho substitution	D	259
F m	F_m	Norrington field param for meta substitution	D	259
F p	F_p	Norrington field param for para substitution	D	259
R o	R_o	Norrington resonance param for ortho substitution	D	259
R m	R_m	Norrington resonance param for meta substitution	D	259
R p	R_p	Norrington resonance param for para substitution	D	259

[a] Label of the parameter in MONOBASE. [b] References from which parameters have been taken, according to the following: A, ref 8; B, ref 9; C, ref 10; D, calculated from values in ref 8, as is explained in ref 11. [c] Number of substituents in MONOBASE for which this parameter is found.

the activity of every product of ES and relevant physicochemical properties of its substituents,[10] as indicated in eq 1,

$$\log \mathrm{Act} = f_h(x_h) + f_e(x_e) + f_s(x_s) + c \qquad (1)$$

where x_h, x_e, and x_s represent, respectively, appropriate hydrophobic, electronic, and steric parameters. Thus, after calculation of the model from an appropriately designed exploring set, the ability to predict the activity of all nonprepared compounds is enhanced. Parametrization should reduce experimental work, so substituent or fragmental parameters are preferable over others obtained from experimental measurements on the whole molecule.

EDISFAR can work with its own substituent parameters database MONOBASE (see Table 1) or with any parameter database defined by the user. The built-in spreadsheet allows the easy customization of databases, making it possible to add or delete substituents and define limit values for any parameter scale.

Once parameters for each substituent have been selected, it would be necessary to work out each possible combination to generate each possible I disubstituted derivative. This is tedious in such a 100 derivatives set, and almost nonaffordable for situations involving more positions and substituents.

The design methods implemented in EDISFAR[15] do not need to work on this data. Once a parameter database is defined for each position involved, EDISFAR performs a partial design (subdesign) for every position. Then such subdesigns are used to generate the polysubstitution global design. Figure 1 shows a block diagram for a simple two-position 2^4 factorial design.

EDISFAR 92 includes three design techniques: factorial, composite central (Box–Wilson), and *D*-optimal. The first two ones are classical methods,[16] largely used in chemistry for process optimization.[17] The data in QSAR, however, are different, as they do not vary in a continuous way and only a few values, represented by possible substituents or molecules,

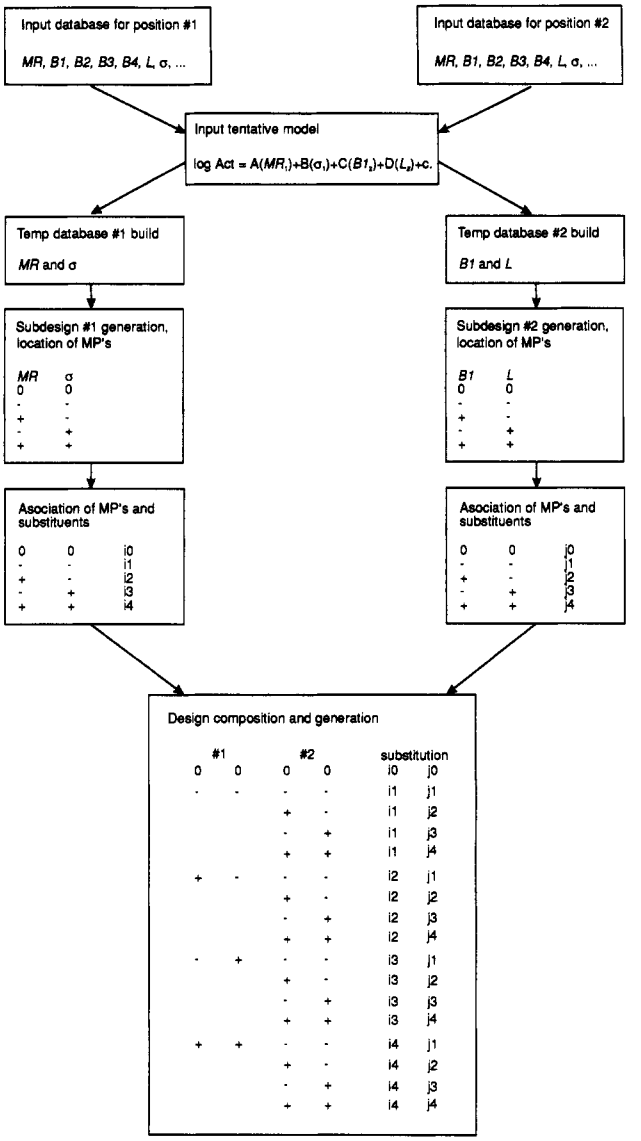


Figure 1. Block diagram of how EDISFAR works out a 2^4 factorial design on a disubstituted prototype.

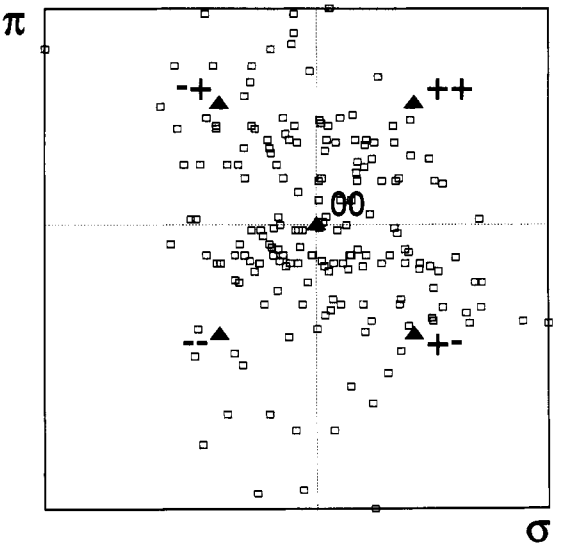


Figure 2. Graphical plot of all ES substituents and the MP given by a 2^2 factorial design.

exist in every scale. Supposing the simple 2^2 factorial design represented in Figure 2, the positions suggested by the method (solid triangles) do not fall over any product present in the

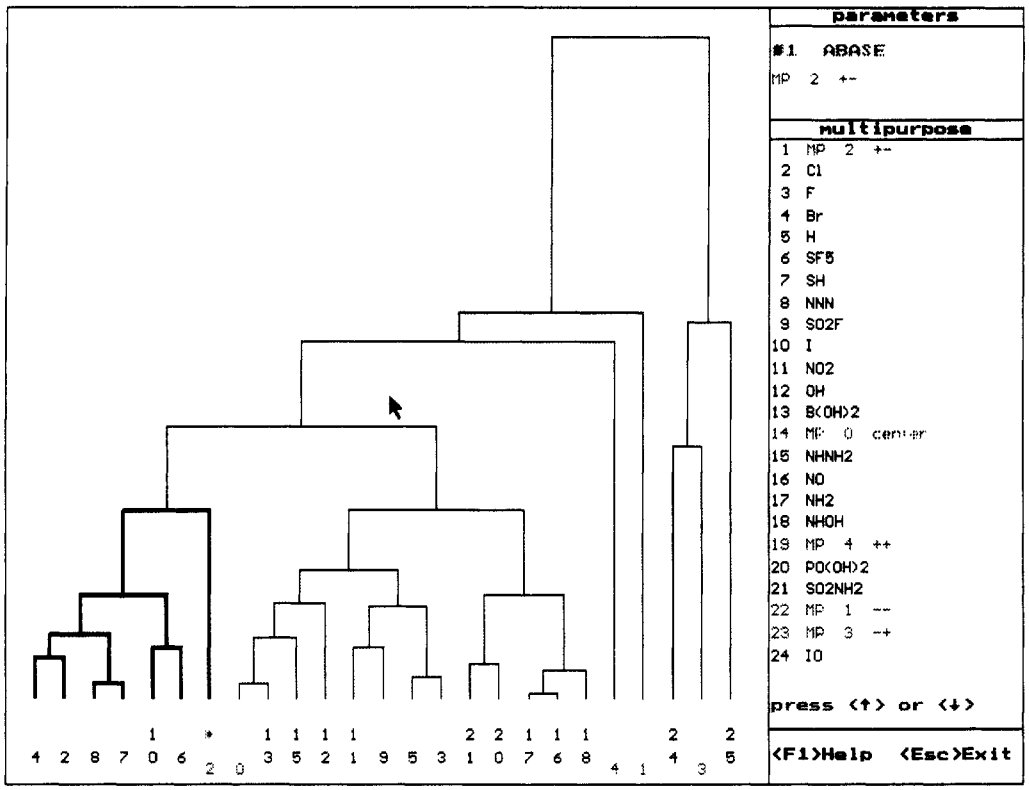


Figure 3. EDISFAR 92 screen showing the results of HCA. Substituents 4, 2, 8, 7, 10, and 6 can be used to represent MP 2 (represented as *2).

ES. These positions, representing ideal experimental conditions, are called *marker points* (MPs).

It is necessary for a criterion to be defined to *associate* the calculated MP with real, existing substituents. On the basis of Austel previous work,[18-20] EDISFAR 92 uses two methods: (a) normalized euclidean distances (NED)[9] and (b) hierarchical clustering analysis (HCA).[6,21] In situations as shown in Figure 2, in which there is a large substituents–MP ratio, NED provides an easy and quick assessment of the most similar substituents. In high-dimension or sparingly populated ES it often become difficult to associate a substituent with MPs. Then, it is recommended that HCA[9] be performed to reveal similarity structures between substituents and MPs. However, resulting dendrograms are difficult to interpret.[18] To make this easier, EDISFAR highlights for each MP substituents included in the bigger cluster containing only this MP and not any other, as is shown in Figure 3.

When there is a large number of parameters describing properties, principal component analysis (PCA) is often used to reduce dimensionality.[22,23] EDISFAR 92 built-in spreadsheet lets the user perform PCA analysis in a very simple way. Simply, the user selects the variables to be included in the analysis, and EDISFAR carries out PCA using the correlation matrix,[24] as indicated in eq 2, where **z** is the scores matrix

$$\mathbf{z} = \mathbf{A}'x^* \tag{2}$$

(matrix of PCs), **A** has columns consisting of the eigenvectors of the correlation matrix, and x^* is the database in standardized (autoscaled) form. The resulting PCs are used for series design and model generation, instead of the original variables. However, there is an important difference; when parameters are used, all of them are considered to equal importance and then all are scaled in the same way, but when PCs are used, the first one is much more important, in terms of variance explained, than the second one, and so on. In this situation EDISFAR can use weighted Euclidean distance (WED)[9] (eq 3), instead of NED, to obtain more realistic results.

$$\mathrm{WED}_{x-s_i} = \left[\sum_{j=1}^{n}\left[\frac{P_{j_x} - P_{j_i}}{P_{j_{\max}} - P_{j_{\min}}}(\mathrm{eig}_j)^{-1/2}\right]^2\right]^{1/2} \tag{3}$$

(B) Accommodation to QSAR Models. At the first stages of development the researcher has no knowledge about which structural characteristics (and thus which parameters) would be related with activity. Initial experimental data must lead to the discovery of such relationships. Bearing in mind that only few of the parameters tested will produce useful relationships, simple design methods such as factorial and fractional factorial designs are initially recommended.

Further developments involving QSAR modeling need improved designs, and it will be desirable to reuse previous work in an efficient way. The whole QSAR process is cyclic; each iteration produces some more information leading to a better QSAR model. That is why design methods must be flexible and powerful enough to work with different kinds of models and use previously obtained information. We suggest starting with factorial or fractional factorial design, to identify the main parameters. To continue, EDISFAR 92 offers two possibilities: composite central designs, and D-optimal designs.

Once an initial factorial design, such as the one represented in Figure 4a, has been performed, and an initial model has been obtained, one may suspect the existence of parabolic interactions that initial design did not consider. One possibility is to complete the design with some axial and central experiments to obtain a composite central or Box–Wilson design[25] (Figure 4b). Depending on the axial spacing of star points (α), the design can meet two possible quality criteria: (a) *orthogonality*, which is necessary to minimize the variance of regression coefficients, and (b) *rotability*, which ensures that variance depends only on the distance from the design center and not on the direction. The α value to obtain orthogonality is expressed by eq 4, where n_f is the number of

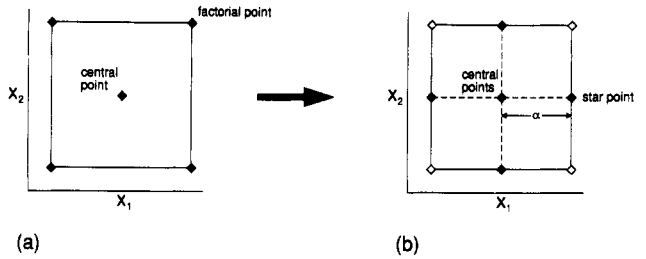


Figure 4. Two sequential design steps: (a) initial factorial design; (b) additional axial and central experiences to complete a composite central design.

$$\alpha^2 = \frac{1}{2}[(n_f + n_a + n_c)n_f - n_f]^{1/2} \quad (4)$$

factorial experiments, n_a is the number of axial experiments, and n_c is the number of experiments in the center of the design. The axial spacing necessary to ensure rotability is expressed in eq 5.

$$\alpha^2 = n_f^{1/2} \quad (5)$$

Both criteria are interesting. If enough central experiments are run, it is possible to make them converge and find an axial spacing which meets both. EDISFAR 92 calculates and offers this value as a default. It asks the user, however, about the number of central experiments to be run. If the user considers the suggested value to be excessive, it is possible to change the α value to meet, at least, orthogonality criteria.

Experiments suggested by this method are considered MP in the same way as above, and substituents can be selected according to NED or clustering criteria.

Often, many of the parameters tested in the initial factorial design do not prove to be useful and then have to be eliminated from the design. Then, the experiments (products) made in relation with these parameters should not be used, and *D*-optimal design would be the best choice.

EDISFAR 92 implements Mitchell's algorithm for *D*-optimal design generation.[26] It is not necessary to apply NED or clustering, because optimal design methods do not give ideal experimental conditions but directly suggest the products to test.

One of the main problems of this method is to define the size of the design. Although the *D* value is by itself a quality criteria, it always increases with the number of experiments. Some authors[27] suggested the use of the normalized information matrix determinant |**M**| as an information/work criteria. In EDISFAR 92, the iterating Mitchell algorithm starts from a given size, but it grows and shrinks during the process, making some trials. A histogram of the highest |**M**| values reached at each tested size is reported at the end of the optimization trials. This is not as exact as calculation of the |**M**| value for each size but gives useful information for evaluating the starting size selection in a much shorter time.

(C) Accommodation to Researchers. Our central idea was to make series design logical and affordable to non-statistic specialists, so a strong effort has been devoted to the user interface. The target was not only to provide an easy-to-use program but also to provide powerful problem planning. In order to make the use of the program easier, we selected a windowing, pop-up menu interface that has recently become a universal standard.[28]

Menus are placed in the menu bar in the logical input order. From left to right there are databases input, model input, and method choice. A key concept in EDISFAR is the model input. It needs no further comment other than that design directly depends on the model it is intended to mimic. In EDISFAR, a hypothetical model must be given to instruct the program about the variables of each database to be used in the design. Models can contain any term usually found in QSAR models; linear, rectangular (interaction), squared, and sum terms. Depending on the design method, information about the given model is used in different ways:

1. *D*-Optimal Design. The model is used to build a model matrix. Bearing in mind that the design would be optimal only for this model, this should include all the terms to be tested. Sum terms involving various positions are not allowed in this method.

2. Factorial and Composite Designs. Variables in any term are included in the factorial scheme. If a variable appears only in a sum term involving two or more positions, the complete design will contain only high- and low-level experiments for the sum variables.[15] Dummy variables can be included without defining a database; the programs will take them into account and work out the proper design.

Model input is independent on the design method. It is possible to change design options or even the design method at any time. Only when all choices have been made is it possible to run the design. If factorial or composite designs were running, a list of the *p* nearest substituents to each MP is presented,[9] being *p* one of the choices of the method. Only one of them has to be included in the final series, but selection is often difficult because it involves knowledge about synthetic accessibility, toxicity, etc. To make this selection, the user should open a selection window and pick up substituents from the list, one at a time. Assessment about the quality of the actual series (orthogonality) can be demanded at any time.

EXAMPLE OF APPLICATION

To illustrate how EDISFAR 92 does work in a QSAR problem, a simulated example of drug optimization is reported. Other possible alternatives, such as the use of data from the literature, were discarded because since EDISFAR work involves series design, it would not be possible to find activity data for each new product suggested. As an example, a quite common situation was chosen. Optimization would be performed on a generic nucleus **II**, on which substituents on

II

two positions (R1 and R2) are to be varied. Most of the usual substituents are synthetically accessible for both positions; symmetrical and nonsymmetrical substitution is equally accessible. There is no previous knowledge of relevant physicochemical properties. The objective is to optimize a hypothetical biological activity, which will be simulated by calculating it with eq 6.

$$-[\log \text{Act}] = 0.5\Sigma\pi_{1,2} + 0.1\text{MR}_1 - 0.01\text{MR}_1^2 - 1\sigma_2 + 10 \quad (6)$$

To introduce some error, as it happens in practice, activity data will be rounded to the next integer value.

(A) Step 1. Initial Series Design and Model Development. For the first step a classical Hansch approach was chosen. Global hydrophobicity and local steric and electronic effects were considered. The first hypothetical model to be tested is

Table 2. Definition of the Experimental Space Used in the Example

param	min	max	mean	std. dev
π	−1.63	2.13	0.32	0.95
MR	5.65	37.88	19.40	7.79
σ_m	−0.15	0.63	0.23	0.19
no. of substituents in this ES: 137				

the one shown in eq 7.

$$-[\log \text{Act}] = A(\Sigma\pi_{1,2}) + B(\text{MR}_1) + C(\text{MR}_2) + D(\sigma_{m1}) + E(\sigma_{m2}) + c \quad (7)$$

No interaction or parabolic term was considered. For this model a fractional factorial scheme 2^{5-2} is good enough. First, we used the EDISFAR spreadsheet to select from MONOBASE a smaller, well-populated ES. The size, limits for each parameter, and other statistical properties are indicated in Table 2.

To make the run it is necessary to instruct EDISFAR about (1) databases to be used for each positions (the database in Table 2 for both positions, in this case), (2) the model to be tested (the model stated above), and (3) the design method (fractional factorial 2^{5-2}; defaults were used for method choices, MPs were placed on the ES centroid with the standard spread of 1.2 standard deviations[9]).

The result of the EDISFAR run was a list of the complete 2^5 scheme. It is possible to choose one of the four fractions, as a function of accessibility, toxicity, etc., of the substituents included. For each MP the user can also choose one of the five substituents presented. Table 3 shows the series obtained, as well as activity data calculated according to eq 6 and some information about the series' orthogonality.

Multiple linear regression (MLR) was performed with this data matrix. Only parameters selected in the design were used. The coefficients for σ_1 and MR_2 were not found to be significant at the 95% confidence interval (T values of 0.73 and −0.11, respectively) and were removed from the equation.

$$-\log \text{Act} = 0.56(\pm0.11)\Sigma\pi_{1,2} - 0.31(\pm0.02)\text{MR}_1 - 1.92(\pm0.96)\sigma_{m2} + 13.03 \quad (8)$$

$$n = 9 \qquad r^2 = 0.985 \qquad F = 109.07$$

Residuals and tentative models suggest the existence of the parabolic MR term. The best equation takes the form of the target model, but the MR_1 coefficient is not significant (T value of 0.38, probability level of 0.72).

$$-[\log \text{Act}] = 0.52(\pm0.06)\Sigma\pi_{1,2} + 0.03(\pm0.09)\text{MR}_1 - 0.008(\pm0.002)\text{MR}_1^2 - 1.53(\pm0.96)\sigma_{m2} + 10.28 \quad (9)$$

$$n = 9 \qquad r^2 = 0.997 \qquad F = 333.62$$

However, the series is not suitable for testing parabolic effects, and it must be completed with some other compounds.

(B) Step 2. Further Development. We decided to perform D-optimal design for the best model found in step 1. All the products in the previous series were forced to be present in the new design (protected experiments), and only five new products were included, selected in a way that maximizes the D determinant. EDISFAR 92 cannot perform D-optimal design for sum terms, so the input equation was as indicated in eq 10.

$$-[\log \text{Act}] = A(\pi_1) + B(\pi_2) + C(\text{MR}_1) + D(\text{MR}_1^2) + E(\sigma_{m2}) + c \quad (10)$$

As stop criteria for the D-optimal search, we used three consecutive provisional optimal occurrences. The search produced the series listed in Table 4. Activity for new products was calculated in a way similar to that described above. The correctness of the design size was confirmed by the histograms of $|\mathbf{M}|$ values.

MLR was performed for this new series, testing the best model (true model), as shown in eq 11.

$$-[\log \text{Act}] = 0.47(\pm0.03)\Sigma\pi_{1,2} + 0.07(\pm0.03)\text{MR}_1 - 0.009(\pm0.0007)\text{MR}_1^2 - 1.35(\pm0.26)\sigma_{m2} + 10.07 \quad (11)$$

$$n = 14 \qquad r^2 = 0.998 \qquad F = 893.55$$

Now all the variables are significant at the 95% confidence interval, and coefficient values are quite similar to the true values.

CONCLUSIONS

In conclusion, EDISFAR 92 is a program which uses classical techniques of sample design. The user can very simply generate exploring series, suitable in complex situations, such as those that appear in practice, including work on polysubstituted lead structures. Starting from parameter databases, which can be easily adapted to every situation, it is possible to use factorial designs in a first stage, which can be implemented by later use of composite central or D-optimal

Table 3. First Training Set

R_1	R_2	$\Sigma\pi_{1,2}$	MR_1	MR_2	σ_1	σ_2	−[log Act]
$CHCHCOOCH_3$	$CHCHCOOCH_3$	0.64	22.56	22.56	0.19	0.19	7
$NCHC_6H_5$	$NHNH_2$	−1.17	33.01	8.44	−0.08	−0.02	2
$OCOCH_3$	$NCHC_6H_5$	−0.93	12.47	33.01	0.39	−0.08	9
$CHNC_6H_5$	$OCOCH_3$	−0.93	33.01	12.47	0.35	0.39	1
$NHNH_2$	$CHNC_6H_5$	−1.17	8.44	33.01	−0.02	0.35	9
Br	C_2H_5	1.88	8.88	10.30	0.39	−0.07	11
$P(C_2H_5)_2$	$P(C_2H_5)$	3.04	30.49	30.49	0.03	0.03	5
C_2H_5	Br	1.88	10.30	8.88	−0.07	0.39	10
$C{=}O(OC_6H_5)$	$C{=}O(OC_6H_5)$	2.92	32.31	32.31	0.37	0.37	4

correlation matrix						eigenvalues	cum variance
	$\Sigma\pi_{1,2}$	MR_1	MR_2	σ_1	σ_2		
$\Sigma\pi_{1,2}$	1.000					1.267	25.35
MR_1	0.098	1.000				1.167	48.68
MR_2	0.098	0.027	1.000			0.969	68.05
σ_1	0.089	0.029	0.187	1.000		0.868	85.41
σ_2	0.054	0.094	0.029	0.138	1.000	0.730	100.00

Table 4. Second Training Set

R_1	R_2	$\Sigma\pi_{1,2}$	MR_1	MR_1^2	σ^2	-[log Act]
$CHCHCOOCH_3$	$CHCHCOOCH_3$	0.64	22.56	508.95	0.19	7
$NCHC_6H_5$	$NHNH_2$	-1.17	33.01	1089.66	-0.02	2
$OCOCH_3$	$NCHC_6H_5$	-0.93	12.47	155.50	-0.08	9
$CHNC_6H_5$	$OCOCH_3$	-0.93	33.01	1089.66	0.39	1
$NHNH_2$	$CHNC_6H_5$	-1.17	8.44	71.23	0.35	9
Br	C_2H_5	1.88	8.88	78.85	-0.07	11
$P(C_2H_5)_2$	$P(C_2H_5)$	3.04	30.49	929.64	0.03	5
C_2H_5	Br	1.88	10.30	106.09	0.39	10
$C{=}O(OC_6H_5)$	$C{=}O(OC_6H_5)$	2.92	32.31	1043.94	0.37	4
NHOH	NHOH	-2.68	7.22	52.13	-0.04	9
CH_3	$SO(CH_3)$	-1.02	5.65	31.92	0.52	9
$NHCSNH_2$	SO_2CH_3	-3.03	22.19	492.40	0.60	5
C_4H_9	C_4H_9	4.26	19.61	384.55	-0.08	10
$NHSO_2C_6H_5$	*a*	2.40	37.88	1434.89	0.49	0

correlation matrix					eigenvalues	cum variance
	$\Sigma\pi_{1,2}$	MR_1	MR_1^2	σ^2		
$\Sigma\pi_{1,2}$	1.000				2.167	54.17
MR_1	0.308	1.000			1.191	83.94
MR_1^2	0.288	0.985	1.000		0.628	99.64
σ_2	0.200	0.159	0.201	1.000	0.014	100.00

[a] 2,5-Dimethyl-1-pyrrolyl.

designs, all producing a QSAR model with the minimum of experimental effort.

ACKNOWLEDGMENT

The authors wish to thank the Comision Interministerial de Ciencia y Tecnología (CICYT) for financial support (Project PB 90/0284) and the Consejería de Educación de la Comunidad de Madrid for a studentship (M.P.).

APPENDIX

EDISFAR 92 is a package of computer programs for IBM PC and compatible computers. The minimum hardware platform requires a 286 or higher microprocessor, VGA graphics card, and a Microsoft compatible mouse. EMS memory is recommended. The programs are fully written in Borland Pascal 7.0 and TurboVision 2.0,[29] using object-oriented programming (OOP) methodology. Further information may be requested directly from the authors (J.A).

REFERENCES AND NOTES

(1) Cramer, R. D., III; Patterson, D. E.; Bunce, J. D. Comparative Molecular Field Analysis (CoMFA). 1. Effect of Shape on Binding of Steroids to Carrier Proteins. *J. Am. Chem. Soc.* **1988**, *110*, 5959–5967.
(2) Alsberg, B.; Esbensen, K. Molecular Structure Description by Segmentation of Atomic Score Spaces (SASS). *Quant. Struct.-Act. Relat.* **1989**, *8*, 218–221.
(3) Crammer, R., III; Bunce, J. D.; Patterson, D. E.; Frank, I. E. Crossvalidation, Bootstrapping, and Partial Least Squares Compared with Multiple Regression in Conventional QSAR Studies. *Quant. Struct.-Act. Relat.* **1988**, *7*, 18–25.
(4) Cruciani, G.; Baroni, M.; Bonelli, D.; Clementi, S.; Ebert, C.; Skagerberg, B. Comparison of Chemometric Models for QSAR. *Quant. Struct.-Act. Relat.* **1990**, *9*, 101–107.
(5) Rose, V. S.; Croall, I. F.; MacFie, H. J. H. An Application of Unsupervised Neural Network Methodology (Kohonen Topology-Preserving Mapping) to QSAR Analysis. *Quant. Struct.-Act. Relat.* **1991**, *10*, 6–15.
(6) Kowalski, B. R.; Bender, C. F. Pattern Recognition. A Powerful Approach to Interpreting Chemical Data. *J. Am. Chem. Soc.* **1972**, *94*, 5632–5639.
(7) Pleiss, M. A.; Unger, S. H. In *Comprehensive Medicinal Chemistry*; Hansch, C., Ed.; Pergamon Press: Oxford, U.K., 1990; Vol. 4, Chapter 21.2, p 561, and references cited therein.
(8) Marsili, M. An Expert System for Chemometrics-Based Optimization in Chemistry. *Tetrahedron Comput. Methodol.* **1988**, *1*, 71–80.
(9) Pastor, M.; Alvarez-Builla, J. The EDISFAR Programs. Rational Drug Series Design. *Quant. Struct.-Act. Relat.* **1991**, *10*, 350–358.
(10) Fujita, T. In *Comprehensive Medicinal Chemistry*; Hansch, C., Ed.; Pergamon Press: Oxford, U.K., 1990; Vol. 4, Chapter 21.1, p 497.
(11) Hansch, C.; Leo, A. *Substituent Constant for Correlation Analysis in Chemistry and Biology*; Wiley: New York, 1976.
(12) Unger, S. H.; Hansch, C. Quantitative Models of Steric Effects. *Prog. Phys. Org. Chem.* **1976**, *12*, 91–118.
(13) Verloop, A. *The STERIMOL Approach to Drug Design*; Dekker: New York, 1978.
(14) Williams, S. G.; Norrington, F. E. Determination of Positional Weighting Factors for the Swain and Lupton Substituent Constants $\mathcal{F}$ and $\mathcal{R}$. *J. Am. Chem. Soc.* **1976**, *98*, 508–516.
(15) Pastor, M.; Alvarez-Builla, J. The EDISFAR Programs. Drug Series Design in Polysubstituted Prototypes. *Quant. Struct.-Act. Relat.*, in press.
(16) Box, G. E. P.; Hunter, W. G.; Hunter, J. S. *Statistic for Experimenters*; Wiley: New York, 1971.
(17) Carlson, R. *Design and Optimization in Organic Synthesis*; Elsevier: Amsterdam, 1992.
(18) Austel, V. A Manual Method for Systematic Drug Design. *Eur. J. Med. Chem.* **1982**, *17*, 9–16.
(19) Austel, V. Selection of Test Compounds from a Basic Set of Chemical Structures. *Eur. J. Med.* **1982**, *17*, 339–347.
(20) Austel, V. 2^n-Factorial Schemes in Drug Design. Extensions Increasing Versatility. *Quant. Struct.-Act. Relat.* **1983**, *2*, 59–65.
(21) Gordon, A. D. A Review of Hierarchical Classification. *J. R. Stat. Soc., Ser. A* **1987**, *150*, 119–137.
(22) Skagerberg, B.; Bonelli, D.; Clementi, S.; Cruciani, G.; Ebert, C. Principal Properties for Aromatic Substituents. A Multivariate Approach for Design in QSAR. *Quant. Struct.-Act. Relat.* **1989**, *8*, 32–38.
(23) Wold, S.; Sjöström, M.; Carlson, T.; Lundstedt, T.; Hellberg, S.; Skagerberg, B.; Wikström, C. Multivariate Design. *Anal. Chem. Acta* **1986**, *191*, 17–32.
(24) Jolliffe, I. T. *Principal Components Analysis*; Springer-Verlag: New York, 1986, p 17.
(25) Box, G. E. P.; Wilson, K. B. On the Experimental Attainment of Optimum Conditions. *J. R. Stat. Soc., Ser. B* **1951**, *13*, 1.
(26) Mitchell, J. T. An Algorithm for the Construction of "D-optimal" Experimental Designs. *Technometrics* **1974**, *16*, 203–210.
(27) Cativiela, C.; García, J. I.; Elguero, J.; Mathieu, D.; Phan Tan Luu, R. Description of Heterocyclic Substituents: A Free-Wilson Approach Using D-Optimal Design. *Quant. Struct.-Act. Relat.* **1987**, *6*, 173–178.
(28) System Application Architecture, Common User Access; *Advanced Interface Design Guide*; International Business Machines: 1989.
(29) Borland Pascal and TurboVision are registered trademarks of Borland International, Inc., Scotts Valley, CA, 1992.