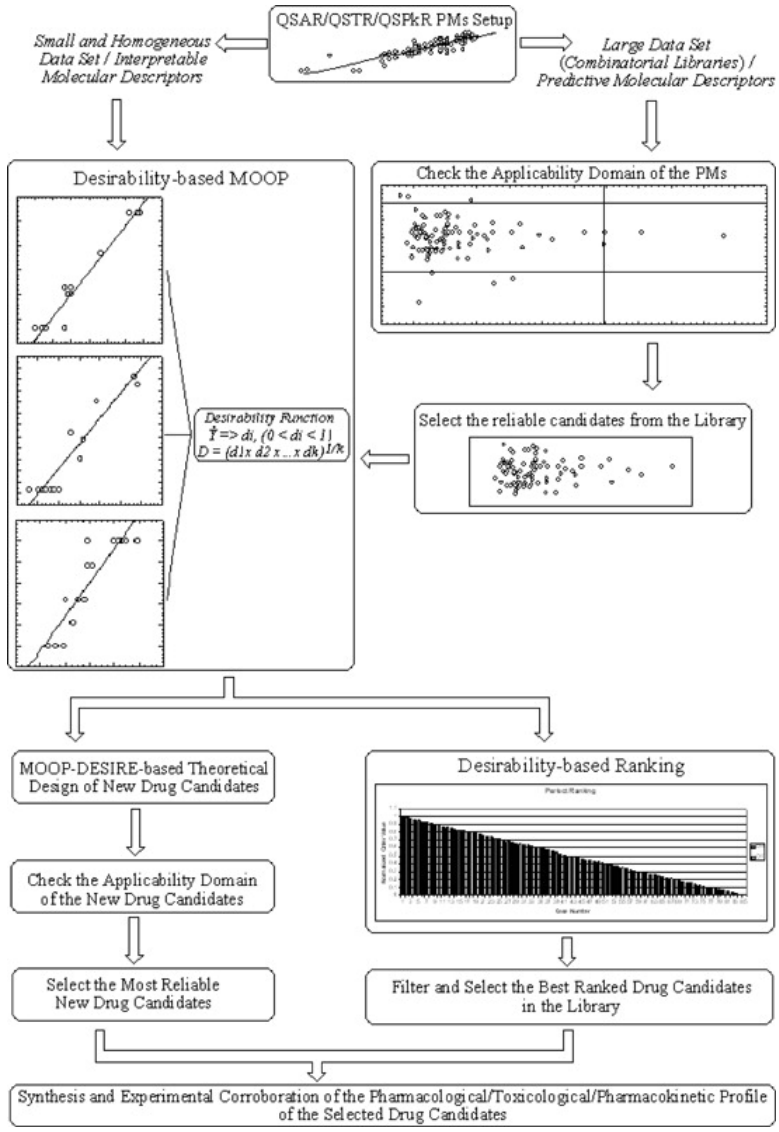
Article

Desirability-Based Methods of Multiobjective Optimization and Ranking for Global QSAR Studies. Filtering Safe and Potent Drug Candidates from Combinatorial Libraries

Maykel Cruz-Monteagudo, Fernanda Borges, M. Nata#lia D. S. Cordeiro, J. Luis Cagide Fajin, Carlos Morell, Reinaldo Molina Ruiz, Yudith Can#izares-Carmenate, and Elena Rosa Dominguez

J. Comb. Chem., **2008**, 10 (6), 897-913• DOI: 10.1021/cc800115y • Publication Date (Web): 15 October 2008

JOURNAL OF combinatorial CHEMISTRY



More About This Article

Additional resources and features associated with this article are available within the HTML version:

- Supporting Information
- Access to high resolution figures
- Links to articles and content related to this article
- Copyright permission to reproduce figures and/or text from this article

Journal of Combinatorial Chemistry is published by the American Chemical Society. 1155 Sixteenth Street N.W., Washington, DC 20036

Desirability-Based Methods of Multiobjective Optimization and Ranking for Global QSAR Studies. Filtering Safe and Potent Drug Candidates from Combinatorial Libraries

Maykel Cruz-Monteagudo,*[,†,‡,§] Fernanda Borges,[†] M. Natália D. S. Cordeiro,[⊥] J. Luis Cagide Fajin,[||] Carlos Morell,[#] Reinaldo Molina Ruiz,[⊥,§] Yudith Cañizares-Carmenate,[‡] and Elena Rosa Dominguez[‡]

Physico-Chemical Molecular Research Unit, Department of Organic Chemistry, Faculty of Pharmacy, REQUIMTE, Department of Chemistry, and CIQ-UP, Faculty of Sciences, University of Porto, 4169-007 Porto, Portugal, and Applied Chemistry Research Center (CEQA), Faculty of Chemistry and Pharmacy, Chemical Bioactive Center (CBQ), and Center of Informatics Studies (CEI), Central University of "Las Villas", Santa Clara, 54830, Cuba

Received July 4, 2008

Up to now, very few applications of multiobjective optimization (MOOP) techniques to quantitative structure−activity relationship (QSAR) studies have been reported in the literature. However, none of them report the optimization of objectives related directly to the final pharmaceutical profile of a drug. In this paper, a MOOP method based on Derringer's desirability function that allows conducting global QSAR studies, simultaneously considering the potency, bioavailability, and safety of a set of drug candidates, is introduced. The results of the desirability-based MOOP (the levels of the predictor variables concurrently producing the best possible compromise between the properties determining an optimal drug candidate) are used for the implementation of a ranking method that is also based on the application of desirability functions. This method allows ranking drug candidates with unknown pharmaceutical properties from combinatorial libraries according to the degree of similarity with the previously determined optimal candidate. Application of this method will make it possible to filter the most promising drug candidates of a library (the best-ranked candidates), which should have the best pharmaceutical profile (the best compromise between potency, safety and bioavailability). In addition, a validation method of the ranking process, as well as a quantitative measure of the quality of a ranking, the ranking quality index (Ψ), is proposed. The usefulness of the desirability-based methods of MOOP and ranking is demonstrated by its application to a library of 95 fluoroquinolones, reporting their gram-negative antibacterial activity and mammalian cell cytotoxicity. Finally, the combined use of the desirability-based methods of MOOP and ranking proposed here seems to be a valuable tool for rational drug discovery and development.

1. Introduction

Development of a successful drug is a complex and lengthy process, and failure at the development stage is caused by multiple factors, such as lack of efficacy, poor bioavailability, and toxicity.[1] Roughly 75% of the total costs during the development of a drug is attributed to poor pharmacokinetics or to toxicity.[2] Improvement of the profile of a candidate drug requires finding the best compromise between various, often competing, objectives. In fact, the ideal drug should have the highest therapeutic efficacy, the highest bioavailability, and the lowest toxicity, which shows the multiobjective nature of the drug discovery and development process. But even when a potent candidate has been identified, the pharmaceutical industry routinely tries to optimize the remaining objectives one at a time, which often results in expensive and time-consuming cycles of trial and error.[3]

In recent years, the drug discovery/development process has been gaining in efficiency and rationality because of the continuous progress and application of chemoinformatics methods.[3] In particular, the quantitative structure−activity relationship (QSAR) paradigm has long been of interest in the drug-design process,[4] redirecting our thinking about structuring medicinal chemistry.[5]

At the same time, the virtual screening (VS)[6,7] of combinatorial libraries has emerged as an adaptive response to the massive throughput synthesis and screening paradigm. In parallel to the development of methods that provide (more) accurate predictions for pharmacological, pharmacokinetic, and toxicological properties for low-number series of com-

* To whom correspondence should be addressed. Phone: +351 222078900. Fax: +351 222003977. E-mail: gmailkelcm@yahoo.es or maikelcm@uclv.edu.cu.

[†] Physico-Chemical Molecular Research Unit, Department of Organic Chemistry, Faculty of Pharmacy, University of Porto.

[‡] Applied Chemistry Research Center (CEQA), Faculty of Chemistry and Pharmacy, Central University of "Las Villas".

[§] Chemical Bioactive Center (CBQ), Central University of "Las Villas".

[⊥] REQUIMTE, Department of Chemistry, Faculty of Sciences, University of Porto.

[||] CIQ-UP, Faculty of Sciences, University of Porto.

[#] Center of Informatics Studies (CEI), Central University of "Las Villas".

10.1021/cc800115y CCC: $40.75
Published on Web 10/15/2008

pounds (tens, hundreds), necessity has forced the computational chemistry community to develop tools that screen against any given target or property, millions or perhaps billions of molecules, virtual or not.[8] VS technologies have thus emerged as a response to the pressure from the combinatorial/high-throughput screening (HTS) community.

Yet standard chemoinformatics approaches usually ignore multiple objectives and optimize each biological property sequentially.[9−20] Nevertheless, some efforts have been made recently toward unified approaches capable of modeling multiple pharmacological, pharmacokinetic, or toxicological properties onto a single QSAR equation.[21−25]

Multiobjective optimization (MOOP) methods introduce a new philosophy to obtain optimality on the basis of compromises among the various objectives. These methods aim at hitting the global optimal solution by optimization of several dependent properties simultaneously. The major benefit of MOOP methods is that local optima, corresponding to one objective can be avoided by taking into account the whole spectra of objectives, thus leading to a more efficient overall process.[26]

Several applications of MOOP methods in the field of drug development have appeared lately, ranging from substructure mining to docking, including inverse quantitative structure property relationship (QSPR) and QSAR.[26] Most of these MOOP applications have been based on the following approaches: weighted-sum-of-objective-functions (WSOF)[27] and pareto-based methods.[26] An excellent review on the subject has been recently published by Nicolaou et al.[26]

Despite the availability of numerous optimization objectives, MOOP techniques have only recently been applied to the building of QSAR models. Actually, very few reports exist of the application of MOOP methods to QSAR,[28−30] and no one reports the simultaneous optimization of competing objectives directly related with the definitive pharmaceutical profile of drugs, such as therapeutic efficacy, bioavailability, and toxicity.

At the same time, ranking of cases is an increasingly important way to describe the result of many data mining and other science and engineering applications.[31] Specifically, in rational drug development, the availability of accurate ranking methods is highly desirable for VS and filtering of promising new drug candidates from combinatorial libraries.[2]

In the present work, we are proposing a MOOP method based on Derringer's desirability function[32] that allows global QSAR studies to be run jointly, considering multiple properties of interest to the drug-design process.[33] The results of the desirability-based MOOP will be used for the implementation of a ranking method also based on the application of desirability functions. In addition, a validation method of the ranking process, as well as a quantitative measure of the quality of a ranking, is proposed. Finally, the usefulness of the desirability-based methods of MOOP and ranking is demonstrated by its application to a library of 95 fluoroquinolones, reporting their gram-negative antibacterial activity and mammalian cell cytotoxicity.

2. Materials and Methods

2.1. Data Set. Our prediction models (PMs), as well as the desirability-based MOOP, were performed using a library of 117 fluoroquinolones published by Suto et al.[34]

The cytotoxicity on Chinese hamster V79 cells expressed as the IC_{50} (μg/mL) and defined as the concentration of compound yielding 50% cell survival compared to untreated control cells. The IC_{50} on Chinese hamster V79 cells is used by Suto et al. as a genetic toxicity end point.[34,35] Gracheck et al.[35] demonstrated that mammalian cell cytotoxicity in Chinese hamster V79 cells was predictive of the in vitro genetic toxicity for the fluoroquinolone class of compounds. In this study, a small group of compounds was evaluated in vitro for their ability to inhibit eukaryotic topoisomerase II activity, their cytotoxicity toward mammalian cells, and their induction of micronuclei, a genetic toxicity end point.[36−40] A strong correlation was seen between the induction of micronuclei in vitro and mammalian cell cytotoxicity ($R^2 = 0.94$).

The compounds were evaluated against five Gram-negative organisms using standard microdilution technique.[41] The data presented represent the geometric mean of the MIC's (μg/mL) for the Gram-negative (*Enterobacter cloacae* MA 2646, *Escherichia coli* Vogel, *Klebsiella pneumonia* MGH-2, *Providencia rettgeri* M 1771, and *Pseudomonas aeruginosa*) bacteria.[34]

Twenty-two out of the 117 compounds reported in ref34 were removed from the data because these values were inaccurately reported (less than, greater than, or greater than or equal to values were reported). The use of inaccurate values reduces significantly the goodness of fit of a multiple linear regression (MLR) model. On the other hand, the values of IC_{50} and MIC of the 95 compounds used as training were transformed ($1/1+ IC_{50}$ or MIC) to obtain the best fit with the predictive variables. The chemical structure and the values of IC_{50} and MIC of the 117 fluoroquinolones are shown in the Supporting Information (see Table SI1).

2.2. Computational Methods. The structures of all compounds were first drawn with the aid of ChemDraw software package,[42] and reasonable starting geometries were obtained by resorting to the MM2 molecular mechanics force field.[43,44] Molecular structures were then fully optimized with the PM3 semiempirical Hamiltonian,[42] implemented in the MOPAC 6.0 program.[45] Here, it should be remarked that the final molecular structures pertain only to the compounds' global minimum energy conformations, and indeed, further molecular simulations or docking studies would be desirable to reach reliable conclusions about conformational requirements and ligand−receptor interactions. But the point of any QSAR model is to have a set of readily calculated descriptors, and such an approach would require much more extensive calculations.

Subsequently, the optimized structures were brought into the DRAGON software package[46] for computation of a total of 1481 molecular descriptors.[47] As part of the necessary variable reduction, descriptors having constant or near-constant values, as well as highly pair-correlated ($|R| > 0.95$) values, were excluded. Table 1 summarizes the DRAGON molecular descriptors used in this work.

Table 1. DRAGON Molecular Descriptors

0D descriptors		1D descriptors	
class	no.	class	no.
constitutional descriptors	47	functional groups	121
		atom-centered fragments	120
		empirical descriptors	3
		properties	3

2D descriptors		3D descriptors	
class	no.	class	no.
topological descriptors	262	charge descriptors	14
molecular walk counts	21	aromaticity indices	4
BCUT descriptors	64	Randic molecular profiles	41
Galvez topological charge indices	21	geometrical descriptors	58
2D autocorrelations	96	RDF descriptors	150
		3D-MoRSE descriptors	160
		WHIM descriptors	99
		GETAWAY descriptors	197

The task of selecting the descriptors that will be more suitable to model the activity of interest is complicated because there are no absolute criteria for such selection. Herein, an optimization technique, the genetic algorithm (GA), was applied for variable selection[48−51] by using the BuildQSAR software package.[52,53] GA evolves a group of random initial models with fitness scores and searches for chromosomes with better fitness functions through natural selection and Darwinian evolution (mutation and crossover). Table 2 depicts the DRAGON molecular descriptors selected by the GA method, which were finally applied to model the antibacterial and cytotoxic properties of the flouroquinolones library used in this study.

For the modeling technique, we opted for a regression-based approach; in this case, the regression coefficients and statistical parameters were obtained by multiple linear regression (MLR) analysis by means of the STATISTICA software package.[54] For each PM, the goodness of fit was assessed by examining the determination coefficient (R^2), the adjusted determination coefficient (Adj.R^2), the standard deviation (s), Fisher's statistics (F), as well as the ratio between the number of compounds (N), and the number of adjustable parameters (p') in the model, known as the ρ statistics. The stability and predictive ability of the models was approached by means of internal cross-validation (CV), specifically by the leave-one-out (LOO) technique.[55] Basically, LOO consists of forming N subsets from the entire data set, each missing one point, which in turn is used to validate a new model that is trained with the corresponding subset. The quality of the new models (cross validation R^2/Q_{LOO}^2) gives an estimated measure of the predictive ability of the full model.

We have also checked the validity of the preadopted parametric assumptions, another important aspect in the application of linear multivariate statistical-based approaches.[56] These include the linearity of the modeled property and the homoscedasticity (or homogeneity of variance), as well as the normal distribution of the residuals and nonmulticollinearity between the descriptors.[57]

Finally, the applicability domain of the final PMs was identified by a leverage plot, that is, a plot of the standardized residuals versus leverages for each training compound.[55,58] The leverage (h_i) of a compound in the original variable space measures its influence on the model and is calculated as

$$h_i = \mathbf{t}_i(\mathbf{T}^T\mathbf{T})^{-1}\mathbf{t}_i^T \quad (1)$$

where $\mathbf{t}_i$ is the descriptor vector of that compound and $\mathbf{T}$ is the model matrix derived from the training set descriptor values. In addition, the warning leverage h^* is defined as

$$h^* = 3 \times p'/N \quad (2)$$

Leverage values can be calculated for both training compounds and new compounds. A leverage higher than the warning leverage h^* means that the compound predicted response can be extrapolated from the model, and thus, the predicted value must be used with great care. On the other hand, a standardized residual value greater than two indicates that the value of the dependent variable for the compound is significantly separated from the remainder training data, and hence, such predictions must be considered with much caution too. In this work, only predicted data for new compounds belonging to the applicability domain of the training set can be considered reliable.

2.3. Desirability Functions Specifications. In the present work, the optimization of the overall desirability was carried on by the "Use general function optimization" option[62–64] of the general regression module of STATISTICA.[54] This process was carried out on a Windows platform in approximately 16 h. Two desirability functions, one for each response, were fitted. Specifically, the cytotoxicity over mammalian cells ought to be minimized (eq 6). This property is expressed here through the IC_{50} value. According to the meaning, this value should be maximized in such a way that the compound with the highest IC_{50} value should be the most desirable ($d_i = 1$). Because of the transformation applied ($1/1+IC_{50}$), this value actually have to be minimized (the same for the antibacterial activity). For estimation of d_i, the lower value $L_i = T_i$ was set to $1/1+IC_{50} = 0.002 = (IC_{50} = 380\ \mu g/mL)$, coinciding with the least cytotoxic compound used for training, and the upper value U_i was set to 0.1/8 μg/mL (the most cytotoxic compound). In contrast, the antibacterial activity against gram-negative microorganisms must be maximized where $L_i = (1/1+MIC = 0.038) = (MIC = 25\ \mu g/mL)$ and $U_i = T_i = (1/1+MIC = 0.99/MIC = 0.01\ \mu g/mL)$ (eq 5). Furthermore, the spline method[59,60] was used for fitting the desirability function, and the current level of each independent variable was set equal to its optimal value. As to the s and t parameters, these were fixed at 1.00 by assuming that the desirability functions increase linearly toward T_i on the two responses.

2.4. Multiobjective Optimization Based on the Desirability Estimation of Several Interrelated Responses. Improvement of the profile of a molecule for the drug discovery and development process requires the simultaneous optimization of several different objectives. The ideal drug should have the highest therapeutic efficacy and bioavailability, as well as the lowest toxicity. Because of the conflicting relationship among the aforementioned properties, such a drug is almost unattainable, and if possible, it is an extremely difficult, expensive, and time-consuming task.

Table 2. DRAGON Molecular Descriptors Selected by the GA Method That Were Used on the Desirability-Based MOOP Process

symbol	definition	class	type	property
MATS3e	Moran autocorrelation lag 3/weighted by atomic Sanderson electronegativities	2D autocorrelations	2D	IC_{50}
GATS5p	Geary autocorrelation lag 5/weighted by atomic polarizabilities	2D autocorrelations	2D	IC_{50}
JGI6	Mean topological charge index of order 6	Galvez topological charge indices	2D	IC_{50}
D/Dr06	distance/detour ring index of order 6	topological descriptors	2D	MIC
BELp1	lowest eigenvalue *n*. One of Burden matrix/weighted by atomic polarizabilities	BCUT descriptors	2D	MIC
H4m	H autocorrelation of lag 4/weighted by atomic masses	GETAWAY descriptors	3D	IC_{50} and MIC
HATS3m	Leverage-weighted autocorrelation of lag 3/weighted by atomic masses	GETAWAY descriptors	3D	MIC
HATS3e	Leverage-weighted autocorrelation of lag 3/weighted by atomic Sanderson electronegativities	GETAWAY descriptors	3D	MIC
H6v	H autocorrelation of lag 6/weighted by atomic van der Waals volumes	GETAWAY descriptors	3D	IC_{50}
R4e+	R maximal autocorrelation of lag 4/weighted by atomic Sanderson electronegativities	GETAWAY descriptors	3D	IC_{50}
R5p	R autocorrelation of lag 5/weighted by atomic polarizabilities	GETAWAY descriptors	3D	IC_{50}
Mor24v	3D-MoRSE signal 24/weighted by atomic van der Waals volumes	3D-MoRSE descriptors	3D	IC_{50}
Mor05m	3D-MoRSE signal 05/weighted by atomic masses	3D-MoRSE descriptors	3D	MIC
Mor14v	3D-MoRSE signal 14/weighted by atomic van der Waals volumes	3D-MoRSE descriptors	3D	MIC
RDF020e	radial distribution function 2.0 /weighted by atomic Sanderson electronegativities	RDF descriptors	3D	MIC
RDF050e	radial distribution function 5.0/weighted by atomic Sanderson electronegativities	RDF descriptors	3D	MIC
FDI	folding degree index	geometrical descriptors	3D	IC_{50}
G(F···F)	sum of geometrical distances between F···F	geometrical descriptors	3D	IC_{50} and MIC

However, finding the best compromise between such objectives is an accessible and more realistic target (see Figure 1).

In this work, we are proposing a multiobjective optimization technique based on the desirability estimation of several interrelated responses (MOOP-DESIRE) as a tool to perform global QSAR studies, considering simultaneously the pharmacological, pharmacokinetic, and toxicological profiles of a set of drug candidates. The MOOP-DESIRE methodology is intended to find the most desirable solution that optimizes a multiobjective problem by using the Derringer's desirability function,[32] specifically addressed to confer rationality to the drug development process. The MOOP method introduced in this work is based on the compromise of potency, safety, and bioavailability. Because other parameters would be also comprised in their future application, the current MOOP is named to identify the possible content. Therefore, this specific application is named MOOP-DESIRE$_{(PHARM-TOX)}$ in allusion to the pharmaceutical and toxicological properties simultaneously optimized.

The process of simultaneous optimization of multiple properties of a drug candidate can be described as follows.

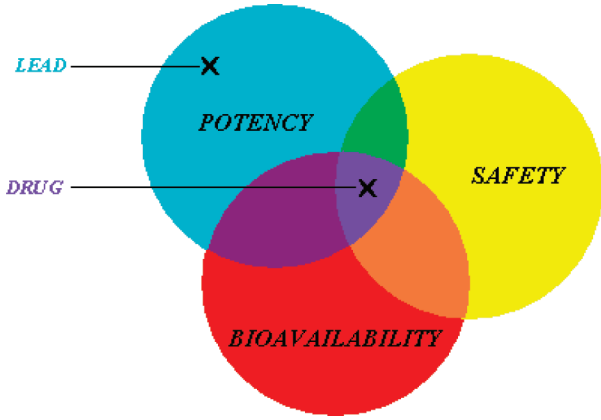


Figure 1. Graphic representation of the compromise between therapeutic efficacy (potency), bioavailability (ADME properties), and toxicity (safety) required to reach a successful drug.

From now on, the terms "response variable" and "independent variables" should be understood as any property to be optimized and any set of molecular descriptors used to model each property, respectively.

2.4.1. Prediction Model Setup. Each response variable (Y_i) is related to the n independent variables (X_n) by an unknown functional relationship, often (but not necessarily) approximated by a linear function. Each predicted response (Y_i) is then estimated by a least-squares regression technique.

In some cases, the developed prediction model for some responses may share the same independent variables of other responses' prediction models but with different coefficients. In this atypical case, attaining the best compromise among the responses turns out to be simpler. Actually, because of the multiplicity of factors involved in the "drugability" of a molecule, one should not expect that the same subset of independent variables can optimally explain both different types of biological properties (especially conflicting properties like potency and toxicity). However, in the latter case, there is still a way to maximize the desirability of both biological properties, that is, to setup a global prediction model where the predicted values of each response are fitted to a linear function using the whole subset of independent variables employed in modeling the k original responses. Here, the independent variables used in computing the predicted values for the original responses will remain the same. Independent variables not used in computing the predicted values for the original responses will be set to zero.

2.4.2. Desirability Function Selection and Evaluation. For each predicted response Y_i, a desirability function d_i assigns values between 0 and 1 to the possible values of Y_i. This transformed response d_i, can have many different shapes. Regardless of the shape, $d_i = 0$ represents a completely undesirable value of Y_i, and $d_i = 1$ represents a completely desirable or ideal response value. The individual desirabilities are then combined using the geometric mean, which gives the overall desirability D

$$D = (d_1 \times d_2 \times \ldots \times d_k)^{\frac{1}{k}} \quad (3)$$

with k denoting the number of responses.

This single value of D gives the overall assessment of the desirability of the combined response levels. Clearly, the range of D will fall in the interval [0, l] and will increase as the balance of the properties becomes more favorable. Notice that if for any response $d_i = 0$, then the overall desirability is zero. Thus, the desirability maximum will be at the levels of the independent variables that simultaneously produce the maximum desirability, given the original models used for predicting each original response.

Depending on whether a particular response is to be maximized, minimized, or assigned a target value, different desirability functions can be used. Here, we used the desirability functions proposed by Derringer and Suich.[32]

Let L_i, U_i, and T_i be the lower, upper, and target values, respectively, that are desired for the response Y_i, with $L_i \leq T_i \leq U_i$.

If a response is of the *target* best kind, then its individual desirability function is defined as

$$d_i = \begin{cases} \left[\dfrac{\hat{Y}_i - L_i}{T_i - L_i}\right]^s & \text{if } L_i \leq \hat{Y}_i \leq T_i \\ \left[\dfrac{\hat{Y}_i - U_i}{T_i - U_i}\right]^t & \text{if } T_i < \hat{Y}_i \leq U_i \\ 0 & \text{if } \hat{Y}_i < L_i \text{ or } \hat{Y}_i > U_i \end{cases} \quad (4)$$

If a response is to be maximized instead, its individual desirability function is defined as

$$d_i = \begin{cases} 0 & \text{if } \hat{Y}_i \leq L_i \\ \left[\dfrac{\hat{Y}_i - L_i}{T_i - L_i}\right]^s & \text{if } L_i < \hat{Y}_i < T_i \\ 1 & \text{if } \hat{Y}_i \geq T_i = U_i \end{cases} \quad (5)$$

In this case, T_i is interpreted as a large enough value for the response, which can be U_i.

Finally, if one wants to minimize a response, one might use

$$d_i = \begin{cases} 1 & \text{if } \hat{Y}_i \leq T_i = L_i \\ \left[\dfrac{\hat{Y}_i - U_i}{T_i - U_i}\right]^s & \text{if } U_i < \hat{Y}_i < T_i \\ 0 & \text{if } \hat{Y}_i \geq U_i \end{cases} \quad (6)$$

Here, T_i denotes a small enough value for the response, which can be L_i. Moreover, the exponents s and t determine how important is to hit the target value T_i. For $s = t = 1$, the desirability function increases linearly toward T_i. Large values for s and t should be selected if it is very desirable that the value of Y_i be close to T_i or increase rapidly above L_i. On the other hand, small values of s and t should be chosen if almost any value of Y_i above L_i and below U_i are acceptable or if having values of Y_i considerably above L_i are not of critical importance.[32]

In this way, one may predict the overall desirability for each drug candidate determined by k responses, which in turn are at the same time determined by a specific set of independent variables. However, as the Derringer's desirability function is built using the estimated responses Y_i, there is no way to know how reliable the predicted D value of each candidate is.

To overcome this shortcoming, we propose a statistical parameter, the *overall desirability's determination coefficient* (R_D^2), which measures the effect of the set of independent variables X_n in reduction of the uncertainty when predicting the D values.

If the response variable is estimated as a continuous function of the independent variables X_n, the individual desirabilities d_i, are continuous functions of the estimated Y_i values (eqs 4−6), and the overall desirability D is a continuous function of the d_i values s (eq. 3), then D is also a continuous function of the X_n. Therefore, R_D^2 can be computed in analogy with the so-called determination coefficient R^2. Specifically, R_D^2 is computed by using the observed D_{Y_i} (calculated from Y_i) and the predicted D_{Y_i} (calculated from Y_i) overall desirability values instead of using directly the measured (Y_i) and predicted (Y_i) response values.

$$R_D^2 = 1 - \frac{\text{SSE}}{\text{SSTO}} = 1 - \frac{\sum (D_{Y_i} - D_{\hat{Y}_i})^2}{\sum (D_{Y_i} - \bar{D}_{Y_i})^2} \quad (7)$$

where D_{Y_i} and D_{Y_i} have been defined previously $\bar{D}_{Y_i}$ is the mean value of D for the Y_i responses of each case included in the data set, SSTO is the total sum of squares, and SSE is the sum of squares due to error.

Similar to R^2, the *adjusted overall desirability's determination coefficient* (Adj.R_D^2) can be computed as shown below.

$$\text{Adj.}R_D^2 = 1 - \frac{\text{SSE}}{\text{SSTO}} = 1 - \frac{\dfrac{\sum (D_{Y_i} - D_{\hat{Y}_i})^2}{N-2}}{\dfrac{\sum (D_{Y_i} - \bar{D}_{Y_i})^2}{N-1}} \quad (8)$$

Like this, both R_D^2 and Adj.R_D^2 have the same properties of R^2 and Adj.R^2. Thus, both will fall in the range [0, 1], and the larger R_D^2 /Adj.R_D^2 is, the lower is the uncertainty in predicting D by using a specific set of independent variables X_n.[61]

Since R_D^2 and Adj.R_D^2 measure the goodness of fit rather than the predictive ability of a certain PM, it is advisable to use an analogue of the leave one out cross-validation determination coefficient (Q_{LOO}^2) to establish the reliability of the method in predicting D. For this, the *overall desirability's LOO-CV determination coefficient* (Q_D^2) can be defined in a manner analogous to that of R_D^2

$$Q_D^2 = 1 - \frac{\text{SSE}_{\text{LOO-CV}}}{\text{SSTO}} = 1 - \frac{\sum (D_{Y_i} - D_{\hat{Y}_i}(\text{LOO-CV}))^2}{\sum (D_{Y_i} - \bar{D}_{Y_i})^2} \quad (9)$$

where $\text{SSE}_{\text{LOO-CV}}$ and D_{Y_i}(LOO-CV) are the leave one out cross validation square sum of residuals and the predicted overall desirability by LOO-CV, respectively.

Table 3. Example of Ordered Lists

O_T	1	2	3	4	5	6	7	8	9	10
O_R	a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8	a_9	a_{10}
	3	6	2	4	5	8	1	7	10	9
O_W	10	9	8	7	6	5	4	3	2	1

In this way, we can have a measure of how reliable will be the simultaneous optimization of the k responses over the independent variables domain.

2.4.3. Multiobjective Optimization. As seen before, the desirability function condenses a multivariate optimization problem into a univariate one. Thus, the overall desirability D can be maximized over the independent variables domain. To accomplish this, one can use the "Response/Desirability Profiler" option of any of the modules of regression or discriminant analysis implemented in STATISTICA.[54] The overall desirability D is optimized with the "Use general function optimization" option, which is, the simplex method of function optimization,[62−64] or the "Optimum desirability at exact grid points" option, which performs exhaustive searches for the optimum desirability at exact grid points. The first option is usually faster, but the default option is the later one, except when the number of predicted values that must be computed to perform the exhaustive grid search exceeds 200 000, in which case the "Use general function optimization" option becomes the default.

The final result is to find the optimal levels (or an optimal range) of the independent variables that optimize simultaneously the k responses determining the final quality of the product. In this way, the best possible compromise between the k responses is found, and consequently, the highest overall desirability for the final compound is reached (i.e., the more enviable drug candidate).

2.5. Desirability-Based Ranking Algorithm. Case-based reasoning (CBR) is mainly based on the assumption that problems (cases; compounds in this work) with similar descriptions (features; molecular descriptors determining the chemical structure in this work) should have similar solutions (the goal of the study; the biological properties involved in the final pharmaceutical profile of the drug candidate in this work).[65] Consequently, by adaptation of previously successful solutions to similar problems, it is possible (at least theoretically) to find the solution of a case only based on its description (that is, to infer the properties of a compound based on their chemical structure from a previous knowledge of the properties of a compound structurally similar).

On the basis of this reasoning paradigm, we are proposing a ranking algorithm based on quantitative parameters estimated from the description of the cases. Specifically, by the application of this algorithm, it will be possible to rank drug candidates (included on the model's applicability domains) with unknown pharmaceutical profiles (like those coming from combinatorial libraries) according to their similarity with the optimal drug candidate determined by the simultaneous multiobjective optimization process previously described.

Δ_i is the parameter used here to describe the similarity between a case i and the optimal case as a function of the subset of descriptive variables used for the multiobjective optimization process, which is defined as

$$\Delta_i = \sum_{X=1}^{m} \delta_{i,X} \cdot w_X \tag{10}$$

where $\delta_{i,X}$ is the Euclidean distance between the case i and the optimal case, considering the parameters X, and w_X represents the weight or influence of the variable X over the global desirability D of the case i.

The Euclidean distance of a case i to a case j considering several features or variables is defined as

$$E = \left[\sum (X_i - X_j)^2\right]^{1/2} \tag{11}$$

Here, we decided to determine the degree of similarity between a case i and the optimal case by considering one by one every single variable X instead of considering simultaneously all the X variables describing a case. By doing this, it is possible to confer a higher degree of freedom to the process of finding the optimal set of weighs associated to the respective variables X. At the same time, this process allows us to infer the relative influence of every variable X over the global desirability D of a case i.

In a case like this one, where only one feature or variable is considered at a time, the Euclidean distance between two cases coincide with the absolute value of the difference between their respective levels of that feature. Thus, $\delta_{i,X}$ is defined as

$$\delta_{i,X} = |X_i - X_{OPT}| \tag{12}$$

where X_i and X_{OPT} are the values of the parameter X for the case i and the optimal case, respectively.

The Δ_i values are normalized by means of the application of the Derringer desirability functions[32] to bring them to the same scale as D_i. In this manner, it is possible to minimize the difference between the values of Δ_i and D_i for every case. Specifically, the respective values of Δ_i are minimized by means of eq 6 in such a way that the lower values (indicative of a higher similarity with respect to the optimal case) will take the values more close to 1 and vice versa. Here, L_i correspond to the lowest value of Δ_i (Δ_{iMIN}) and $U_i = \Delta_{iMAX}$.

Next, the optimal set of weighs w_X minimizing the difference between the values of D_i and the normalized values of Δ_i for every case is found by a least-squares nonlinear data-fitting process. The weights were obtained through a nonlinear curve-fitting using the large-scale optimization algorithm,[66,67] implemented in the "lsqcurvefit" function of MATLAB program, version 7.2.[68] This process was carried out over a windows platform at a very low computational cost. A copy of the function employed is available in the Supporting Information.

After we minimized the differences between D_i and the normalized values of Δ_i, we achieved the highest possible degree of concordance between the description (expressed through the normalized values of Δ_i which encode the information related to the molecular structure expressed as a function of the molecular descriptors employed) and the solution of the cases (determined by the respective values of D_i, which represents the combination of the k properties involved on the final quality of the drug candidate). Thus, according to the CBR paradigm, it will be possible to rank, according to Δ_i, new and pharmaceutically unknown drug

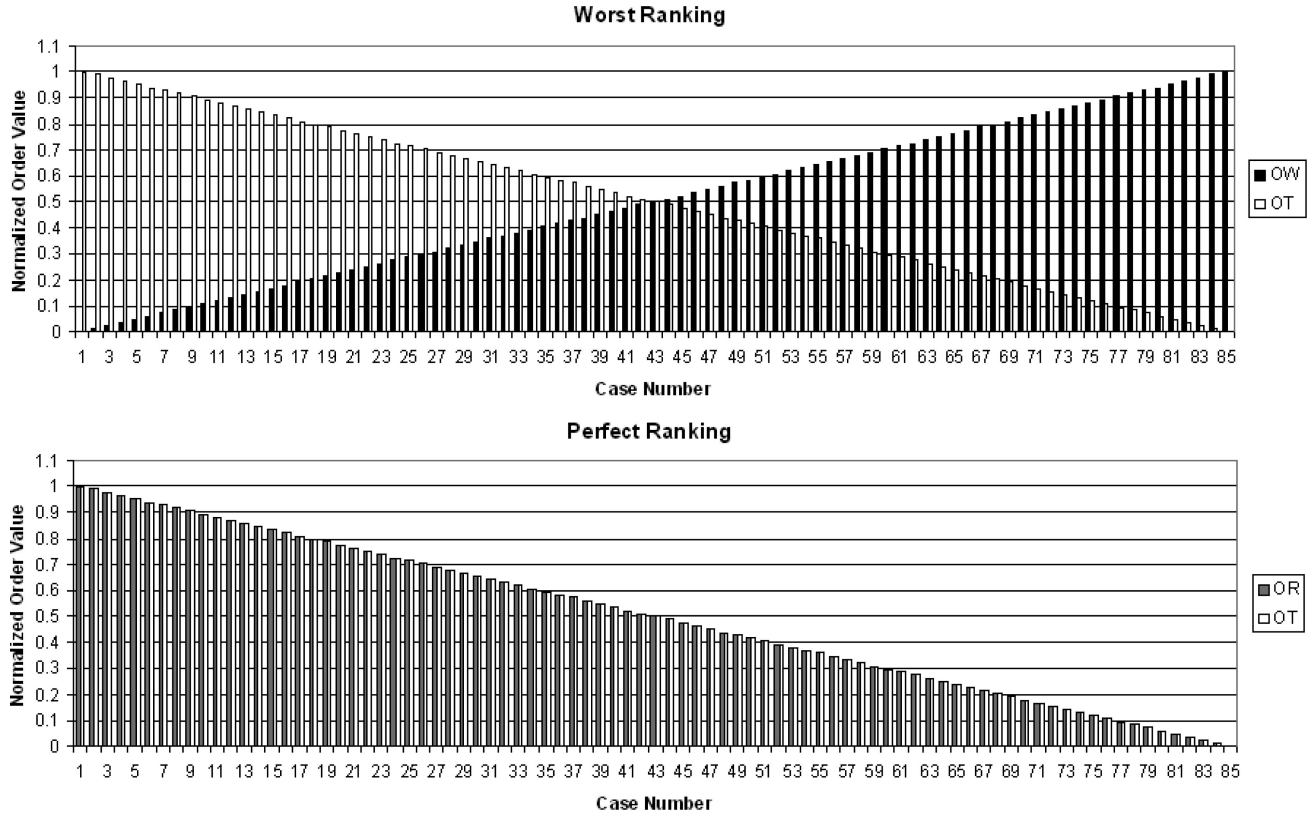


Figure 2. Worst (top) and perfect (bottom) ranking.

Table 4. Regression Coefficients and Statistical Parameters for the MLR Models

antibacterial activity MLR model (MIC = 1/(1 + MIC))

1/1 + MIC = 27.127(±3.925) − 1.573(±0.170)·H4M − 13.504(±1.969)·BELp1 +0.071(±0.012)·RDF020e − 0.130(±0.024)·Mor05m − 0.006(±0.001)·G(F···F) +5.670(±1.097)·HATS3m + 0.002(±0.000)·D/Dr06 − 0.234(±0.064)·Mor14v +1.449(±0.423)·HATS3e + 0.011(±0.003)·RDF050e

N	*R*	R^2	Adj.R^2	*S*	Q^2	SPRESS	ρ	F	p
95	0.883	0.779	0.753	0.096	0.725	0.107	8.636	29.601	0.0000

cytotoxicity MLR model (IC_{50}= 1/(1 + IC_{50}))

1/1 + IC_{50} = −0.966(±0.146) + 0.611(±0.053)·R5p − 0.135(±0.012)·GATS5p −0.147(±0.018)·H4m + 1.239(±0.156)·FDI + 0.002(±0.000)·G(F···F) + 0.114(±0.019)·Mor24v − 0.162(±0.039)·H6v + 0.183(±0.045)·MATS3e − 0.329(±0.086)·R4e$^+$ − 1.152(±0.397)·JGI6

N	*R*	R^2	Adj.R^2	*S*	Q^2	SPRESS	ρ	F	p
95	0.867	0.750	0.721	0.014	0.686	0.016	8.636	25.313	0.0002

candidates for which just their molecular structure is known (like those coming from combinatorial libraries). In this way, it will be possible to filter and identify the most promising drug candidates, which will logically be placed first on the order list (the candidates with the lowest values of Δ_i and consequently the most similar ones with the optimal drug candidate determined by the desirability-based MOOP process) and to discard the candidates ordered last.

2.6. Ranking Algorithm Validation and Estimation of the Ranking Quality Index (Ψ). Even though the CBR suggests that the nonlinear data-fitting process employed to find the optimal set of weighs can lead to an adequate ranking of the cases, it is not possible to know the quality of the ranking achieved through this process. Considering the above-mentioned, we are proposing a method for the validation of the ranking obtained by the use of the optimal set of weighs. In addition, we propose a quantitative criterion of the quality of a ranking. Specifically, in this work we use the same data set used for the desirability-based MOOP process.

We will use some simple notations to represent ordering throughout this paper. Without loss of generality, for *n* cases to be ordered, we use the actual ordering position of each case as the label to represent this case in the ordered list. For example, suppose that the label of the actual highest ranked case is *n*, the label of the actual second highest ranked case is $n - 1$, etc. We assume the examples are ordered incrementally from left to right. Then the *true-order list* is $O_T = 1, 2, 3, ..., n$. For any ordered list generated by a ranking algorithm, it is a permutation of O_T. We use O_R to denote the ordered list generated by the ranking algorithm *R*. O_R can be written as $a_1, a_2, ..., a_I$, where a_i is the actual ordering position of the case that is ranked *i*th in O_R (see Table 3).

The ranking validation includes the following steps:

1. Order the cases in the library according to *D* in a decreasing fashion (starting with the case exhibiting the highest value of *D*) and label each case as described above ((1, 2, 3, ..., *n*). This ordering corresponds to the true-order list (O_T).
2. Invert O_T. This new ordering corresponds to the worst-order list (O_W).
3. Order incrementally the cases in the library according to Δ_i (starting with the case exhibiting the lowest value of Δ_i) and label each case as described above ($a_1, a_2, ..., a_n$). This

Table 5. Observed and Predicted Values of the Optimized Properties and Their Respective Individual and Overall Desirability Values for the Compounds Used on the Desirability-Based MOOP Process

compound ID	1/1 + MIC	predicted 1/1 + MIC	*d*(MIC)	predicted *d*(MIC)	1/1 + IC_{50}	predicted 1/1 + IC_{50}	*d*(IC_{50})	predicted *d*(IC_{50})	$D_{MIC-IC_{50}}$	predicted $D_{MIC-IC_{50}}$
004-4-ciprofloxacin	0.909	0.908	0.915	0.914	0.003	−0.010	0.994	1.000	0.954	0.956
006-6-tosufloxacin	0.917	0.931	0.924	0.938	0.008	−0.006	0.941	1.000	0.932	0.968
007-7-PD117558	0.917	0.693	0.924	0.688	0.083	0.052	0.170	0.489	0.396	0.580
008-8	0.833	0.607	0.835	0.598	0.006	0.017	0.957	0.848	0.894	0.712
010-10	0.355	0.281	0.333	0.255	0.017	0.021	0.847	0.801	0.531	0.452
012-13	0.193	0.555	0.163	0.543	0.004	−0.002	0.978	1.000	0.400	0.737
014-15	0.641	0.576	0.633	0.565	0.003	−0.011	0.988	1.000	0.791	0.751
015-16	0.685	0.764	0.680	0.763	0.006	0.020	0.957	0.814	0.806	0.788
016-17	0.556	0.644	0.544	0.636	0.005	0.007	0.967	0.945	0.725	0.776
018-19	0.893	0.891	0.898	0.896	0.003	0.003	0.987	0.993	0.941	0.943
019-20	0.885	0.947	0.890	0.955	0.003	0.006	0.988	0.962	0.937	0.959
020-21	0.962	0.891	0.970	0.896	0.032	0.031	0.691	0.701	0.819	0.793
021-22	0.769	0.872	0.768	0.876	0.006	0.009	0.957	0.926	0.857	0.901
022-23A	0.833	0.807	0.835	0.808	0.026	0.021	0.759	0.805	0.796	0.806
023-23B	0.909	0.795	0.915	0.796	0.008	0.026	0.936	0.759	0.925	0.777
024-23C	0.909	0.936	0.915	0.944	0.007	0.015	0.953	0.865	0.934	0.904
025-23D	0.769	0.780	0.768	0.780	0.007	0.027	0.953	0.743	0.855	0.761
026-23E	0.074	0.304	0.038	0.279	0.004	−0.007	0.984	1.000	0.193	0.529
027-23F	0.794	0.905	0.794	0.911	0.017	0.021	0.847	0.805	0.820	0.856
028-24A	0.917	0.973	0.924	0.982	0.014	0.014	0.881	0.882	0.902	0.930
029-24C	0.971	0.865	0.980	0.869	0.037	0.013	0.642	0.889	0.793	0.879
030-24D	0.935	0.893	0.942	0.898	0.022	0.015	0.799	0.866	0.867	0.882
031-24E	0.833	0.683	0.835	0.677	0.003	0.001	0.987	1.000	0.908	0.823
032-24F	0.971	0.944	0.980	0.951	0.010	0.017	0.917	0.844	0.948	0.896
033-25A	0.833	0.827	0.835	0.829	0.012	0.026	0.896	0.753	0.865	0.790
034-25B	0.952	1.016	0.960	1.000	0.083	0.075	0.170	0.258	0.404	0.508
036-25D	0.901	0.879	0.906	0.884	0.091	0.045	0.093	0.558	0.290	0.702
037-25E	0.658	0.618	0.651	0.609	0.026	0.021	0.759	0.802	0.703	0.699
038-25F	0.877	0.848	0.882	0.851	0.019	0.041	0.824	0.598	0.852	0.713
040-26D	0.794	0.745	0.794	0.743	0.043	0.028	0.577	0.731	0.677	0.737
041-26E	0.625	0.600	0.617	0.590	0.008	0.005	0.936	0.969	0.760	0.756
042-26F	0.826	0.795	0.828	0.796	0.006	0.019	0.957	0.828	0.890	0.811
043-27A	0.658	0.773	0.651	0.772	0.042	0.037	0.595	0.647	0.623	0.707
044-27B	0.885	0.843	0.890	0.845	0.111	0.112	0.000	0.000	0.000	0.000
045-27C	0.935	0.989	0.942	0.999	0.111	0.088	0.000	0.122	0.000	0.349
046-27D	0.794	0.855	0.794	0.858	0.026	0.052	0.759	0.487	0.776	0.647
047-27E	0.500	0.598	0.485	0.588	0.009	0.026	0.928	0.756	0.671	0.667
048-27F	0.741	0.717	0.738	0.714	0.038	0.036	0.628	0.658	0.681	0.685
049-28A	0.714	0.687	0.710	0.681	0.005	0.018	0.971	0.832	0.830	0.753
050-28B	0.813	0.823	0.814	0.824	0.111	0.085	0.000	0.156	0.000	0.359
051-28C	0.794	0.659	0.794	0.652	0.042	0.064	0.595	0.367	0.687	0.489
052-28D	0.658	0.729	0.651	0.726	0.008	0.032	0.936	0.689	0.781	0.707
054-28F	0.625	0.702	0.617	0.698	0.017	0.013	0.847	0.891	0.723	0.789
055-29B	0.935	1.009	0.942	1.000	0.021	0.032	0.808	0.698	0.872	0.835
056-29C	0.935	1.016	0.942	1.000	0.023	0.039	0.788	0.626	0.862	0.791
057-29D	0.935	0.883	0.942	0.888	0.012	0.025	0.897	0.761	0.919	0.822
058-29E	0.870	0.664	0.873	0.658	0.006	−0.002	0.957	1.000	0.914	0.811
059-29F	0.917	0.919	0.924	0.925	0.008	0.013	0.936	0.886	0.930	0.905
061-30B	0.952	0.938	0.960	0.946	0.007	0.021	0.953	0.804	0.957	0.872
062-30C	0.813	0.824	0.814	0.826	0.007	0.024	0.948	0.776	0.879	0.800
063-30D	0.746	0.744	0.744	0.742	0.002	0.002	1.000	0.996	0.863	0.860
064-30E	0.524	0.637	0.510	0.629	0.002	−0.007	1.000	1.000	0.714	0.793
065-30F	0.855	0.784	0.858	0.783	0.004	−0.019	0.980	1.000	0.917	0.885
066-31A	0.794	0.808	0.794	0.809	0.004	0.006	0.976	0.962	0.880	0.882
067-31B	0.833	0.888	0.835	0.893	0.042	0.044	0.595	0.576	0.705	0.717
068-31C	0.926	0.898	0.933	0.904	0.053	0.042	0.483	0.595	0.671	0.733
070-31E	0.794	0.674	0.794	0.668	0.048	0.013	0.534	0.885	0.651	0.769
071-31F	0.813	0.771	0.814	0.770	0.010	0.023	0.919	0.790	0.865	0.780
073-32B	0.885	0.951	0.890	0.959	0.019	0.035	0.831	0.660	0.860	0.796
074-32C	0.935	0.869	0.942	0.873	0.040	0.044	0.612	0.576	0.759	0.709
075-32D	0.813	0.834	0.814	0.836	0.014	0.020	0.875	0.815	0.844	0.826
077-32F	0.714	0.787	0.710	0.787	0.010	0.010	0.919	0.914	0.808	0.848
078-33B	0.813	0.739	0.814	0.736	0.011	0.010	0.907	0.914	0.859	0.820
079-34B	0.658	0.628	0.651	0.620	0.010	0.023	0.918	0.789	0.773	0.699
080-35B	0.741	0.799	0.738	0.799	0.003	−0.002	0.985	1.000	0.853	0.894
081-36B	0.556	0.525	0.544	0.511	0.005	0.002	0.967	1.000	0.725	0.715
082-37B	0.488	0.562	0.472	0.550	0.008	0.014	0.943	0.879	0.667	0.695
083-38A	0.794	0.826	0.794	0.827	0.026	0.013	0.759	0.889	0.776	0.857
084-38B	0.685	0.723	0.680	0.720	0.004	0.012	0.980	0.896	0.816	0.803
085-39A	0.500	0.376	0.485	0.355	0.009	0.002	0.928	1.000	0.671	0.596
086-39B	0.326	0.296	0.302	0.271	0.053	0.054	0.483	0.472	0.382	0.358
088-41A	0.926	0.934	0.933	0.941	0.022	0.039	0.799	0.619	0.863	0.763

Table 5. Continued

compound ID	1/1 + MIC	predicted 1/1 + MIC	*d*(MIC)	predicted *d*(MIC)	1/1 + IC_{50}	predicted 1/1 + IC_{50}	*d*(IC_{50})	predicted *d*(IC_{50})	$D_{MIC-IC_{50}}$	predicted $D_{MIC-IC_{50}}$
090-42A	0.685	0.634	0.680	0.626	0.005	0.017	0.974	0.849	0.814	0.729
092-48	0.685	0.673	0.680	0.667	0.014	0.013	0.875	0.890	0.771	0.770
093-49	0.654	0.844	0.647	0.847	0.004	0.001	0.981	1.000	0.797	0.920
094-50	0.833	0.873	0.835	0.877	0.031	0.034	0.702	0.678	0.766	0.771
095-51	0.962	0.936	0.970	0.943	0.018	0.010	0.835	0.914	0.900	0.929
096-52	0.917	0.910	0.924	0.916	0.067	0.053	0.340	0.482	0.561	0.664
098-54	0.962	0.913	0.970	0.920	0.014	0.002	0.881	0.995	0.924	0.957
100-56	0.926	0.807	0.933	0.808	0.010	0.003	0.919	0.991	0.926	0.895
101-57	0.038	0.294	0.000	0.269	0.005	0.004	0.967	0.982	0.022	0.514
102-58	0.990	0.926	1.000	0.933	0.063	0.043	0.383	0.584	0.619	0.738
103-59	0.926	0.960	0.933	0.968	0.017	0.029	0.850	0.725	0.891	0.838
104-60	0.901	0.917	0.906	0.923	0.010	0.016	0.919	0.858	0.913	0.890
105-61	0.524	0.498	0.510	0.483	0.003	0.017	0.985	0.850	0.709	0.641
106-62	0.980	0.877	0.990	0.881	0.083	0.078	0.170	0.226	0.410	0.446
107-63	0.971	0.973	0.980	0.982	0.023	0.030	0.788	0.718	0.879	0.840
110-70	0.488	0.460	0.472	0.443	0.015	0.010	0.870	0.916	0.641	0.637
111-71	0.524	0.593	0.510	0.583	0.003	0.015	0.985	0.869	0.709	0.712
112-72	0.741	0.619	0.738	0.610	0.016	0.009	0.856	0.929	0.795	0.753
113-73	0.625	0.570	0.617	0.559	0.023	0.025	0.783	0.769	0.695	0.655
114-74	0.641	0.661	0.633	0.655	0.021	0.015	0.803	0.868	0.713	0.754
115-75	0.592	0.619	0.582	0.611	0.019	0.027	0.831	0.745	0.695	0.675
117-77	0.781	0.820	0.781	0.821	0.100	0.082	0.000	0.188	0.000	0.393
118-78	0.625	0.623	0.617	0.615	0.004	0.004	0.983	0.977	0.778	0.775
	$R_{D(MIC-IC_{50})}^2 = 0.702$					Adj.$R_{D(MIC-IC_{50})}^2 = 0.698$				

ordering corresponds to the order generated by the ranking algorithm R (O_R).

4. Normalize (through eq 6) the values (labels) assigned to each case in steps 1−3 where $L_i = T_i = 1$ and $U_i =$ the number of cases included in the library (n). In this way, we obtained the respective normalized order values for the true ($^{OT}d_i$) and worst ($^{OW}d_i$) order lists, as well as the order generated by the ranking algorithm R ($^{OR}d_i$).

5. Use the respective normalized order values to determine the difference between O_R and O_T ($^{OT-OR}\delta_i$)

$$^{OT-OR}\delta_i = |^{OT}d_i - {}^{OR}d_i| \tag{13}$$

and between O_W and O_T ($^{OT-OW}\delta_i$)

$$^{OT-OW}\delta_i = |^{OT}d_i - {}^{OW}d_i| \tag{14}$$

The ideal difference is 0 for all the cases and corresponds to a perfect ranking. Figure 2 illustrates both worst and perfect rankings, respectively.

6. Estimate the quality of the order generated by the ranking algorithm R (O_R) by means of the ranking quality index (Ψ), which can be defined as the absolute value of the mean of $^{OT-OR}\delta_i$, for the n cases included in the library to be ranked

$$\Psi = \frac{\left|\sum_{i=1}^{n} {}^{OT-OR}\delta_i\right|}{n} \tag{15}$$

Ψ is in the range [0, 0.5], being $\Psi = 0$ if a ranking is perfect and $\Psi \cong 0.5$ for the worst ranking. The closer Ψ is to 0 for a certain ranking, the higher the quality of this ranking. In contrast, values of Ψ near 0.5 indicate a low ranking quality. Because the value of Ψ associated with the worst ranking is dependent on the size of the library to be ranked, this value is not exactly, but is approximately, equal to 0.5. At the same time, a range [0, 1] rather than [0, 0.5] is a more clear indicator of the quality of a ranking. Considering both of the previous questions, a correction factor (F) is applied to Ψ

$$F = \frac{2}{\Psi^{OW}} \tag{16}$$

where Ψ^{OW} is the quality index for the worst ranking. F is used here to obtain a more representative indicator Ψ of the quality of a ranking and at the same time to include Ψ in the range [0, 1], where Ψ^{OW} is exactly equal to 1. In this way, we obtain the corrected ranking quality index (Ψ^*)

$$\Psi^* = \frac{\left|\sum_{i=1}^{n} {}^{OT-OR}\delta_i\right|}{n} \cdot F = \frac{\left|\sum_{i=1}^{n} {}^{OT-OR}\delta_i\right|}{n} \cdot \frac{2}{\Psi^{WR}} \tag{17}$$

Finally, it is possible to express Ψ^* as the percentage of ranking quality ($R_{\%}$)

$$R_{\%} = (1 - \Psi^*) \cdot 100 \tag{18}$$

3. Results and Discussion

3.1. MOOP-DESIRE$_{(PHARM-TOX)}$-Based Optimization. To test the utility of the MOOP-DESIRE methodology for the simultaneous optimization of multiple properties, it was applied to a library of 95 fluoroquinolones reported by Suto et al. with the aim of simultaneously optimizing their antibacterial activity over gram-negative microorganisms (MIC) and their cytotoxic effects over mammalian cells (IC_{50}).

Following the strategy outlined previously, we began by seeking the best linear models relating each property to the DRAGON molecular descriptors. One should emphasize here that the reliability of the final results of the optimization process strongly depends on the quality of the initial set of PMs.

One MLR-based PM containing 10 variables previously selected by GA was developed for both properties. The

Table 6. Predicted Values of the Optimized Properties and Their Respective Individual and Overall Desirability Values Obtained after the LOO-CV Experiment for the Compounds Used on the Desirability-Based MOOP Process

compound ID	LOO-CV predicted 1/1 + MIC	LOO-CV predicted *d*(MIC)	LOO-CV predicted 1/1 + IC_{50}	LOO-CV predicted *d*(IC_{50})	LOO-CV predicted $D_{MIC-IC_{50}}$
004-4-ciprofloxacin	0.908	0.914	−0.011	1.000	0.956
006-6-tosufloxacin	0.935	0.943	−0.008	1.000	0.971
007-7-PD117558	0.683	0.678	0.051	0.505	0.585
008-8	0.600	0.590	0.018	0.837	0.703
010-10	0.261	0.234	0.022	0.793	0.431
012-13	0.578	0.568	−0.002	1.000	0.753
014-15	0.568	0.557	−0.014	1.000	0.746
015-16	0.772	0.771	0.022	0.800	0.786
016-17	0.651	0.644	0.008	0.942	0.779
018-19	0.891	0.896	0.003	0.994	0.944
019-20	0.952	0.960	0.006	0.959	0.960
020-21	0.887	0.892	0.031	0.702	0.791
021-22	0.877	0.882	0.009	0.925	0.903
022-23A	0.800	0.801	0.021	0.809	0.805
023-23B	0.780	0.779	0.027	0.747	0.763
024-23C	0.939	0.947	0.016	0.857	0.901
025-23D	0.781	0.780	0.029	0.726	0.753
026-23E	0.363	0.342	−0.008	1.000	0.585
027-23F	0.920	0.927	0.021	0.803	0.863
028-24A	0.977	0.987	0.014	0.882	0.933
029-24C	0.858	0.861	0.011	0.909	0.884
030-24D	0.891	0.896	0.015	0.870	0.883
031-24E	0.674	0.668	0.000	1.000	0.817
032-24F	0.942	0.949	0.018	0.841	0.893
033-25A	0.827	0.829	0.027	0.742	0.784
034-25B	1.024	1.000	0.074	0.265	0.515
036-25D	0.878	0.882	0.040	0.616	0.737
037-25E	0.616	0.607	0.021	0.806	0.699
038-25F	0.846	0.849	0.042	0.588	0.706
040-26D	0.740	0.737	0.026	0.750	0.743
041-26E	0.596	0.587	0.004	0.975	0.756
042-26F	0.792	0.792	0.020	0.817	0.804
043-27A	0.782	0.782	0.035	0.668	0.722
044-27B	0.840	0.842	0.112	0.000	0.000
045-27C	0.996	1.000	0.084	0.162	0.402
046-27D	0.861	0.865	0.057	0.434	0.613
047-27E	0.606	0.596	0.028	0.733	0.661
048-27F	0.716	0.712	0.035	0.662	0.686
049-28A	0.685	0.679	0.021	0.808	0.741
050-28B	0.823	0.825	0.080	0.205	0.412
051-28C	0.643	0.636	0.067	0.334	0.461
052-28D	0.735	0.733	0.035	0.665	0.698
054-28F	0.708	0.703	0.012	0.899	0.795
055-29B	1.013	1.000	0.032	0.692	0.832
056-29C	1.023	1.000	0.040	0.616	0.785
057-29D	0.877	0.881	0.026	0.755	0.816
058-29E	0.630	0.622	−0.004	1.000	0.789
059-29F	0.919	0.925	0.013	0.883	0.904
061-30B	0.936	0.943	0.022	0.794	0.865
062-30C	0.826	0.827	0.026	0.760	0.793
063-30D	0.744	0.741	0.002	0.996	0.859
064-30E	0.655	0.648	−0.008	1.000	0.805
065-30F	0.775	0.775	−0.021	1.000	0.880
066-31A	0.810	0.811	0.006	0.960	0.882
067-31B	0.891	0.896	0.044	0.574	0.717
068-31C	0.895	0.900	0.040	0.607	0.739
070-31E	0.663	0.656	0.009	0.929	0.781
071-31F	0.767	0.766	0.024	0.779	0.772
073-32B	0.957	0.965	0.036	0.654	0.794
074-32C	0.857	0.861	0.044	0.573	0.702
075-32D	0.836	0.839	0.021	0.810	0.824
077-32F	0.793	0.793	0.010	0.914	0.851
078-33B	0.723	0.720	0.010	0.915	0.812
079-34B	0.607	0.598	0.030	0.715	0.654
080-35B	0.815	0.816	−0.002	1.000	0.904
081-36B	0.520	0.506	0.001	1.000	0.711
082-37B	0.577	0.566	0.015	0.867	0.701
083-38A	0.827	0.829	0.011	0.904	0.865
084-38B	0.730	0.726	0.014	0.877	0.798
085-39A	0.362	0.340	0.000	1.000	0.583
086-39B	0.280	0.254	0.054	0.466	0.344
088-41A	0.935	0.942	0.041	0.606	0.755
090-42A	0.630	0.622	0.018	0.839	0.723

Table 6. Contiuned

compound ID	LOO-CV predicted 1/1 + MIC	LOO-CV predicted *d*(MIC)	LOO-CV predicted 1/1 + IC_{50}	LOO-CV predicted *d*(IC_{50})	LOO-CV predicted $D_{MIC-IC_{50}}$
092-48	0.669	0.662	0.013	0.891	0.768
093-49	0.861	0.865	0.001	1.000	0.930
094-50	0.894	0.899	0.034	0.673	0.778
095-51	0.933	0.940	0.010	0.918	0.929
096-52	0.909	0.915	0.051	0.497	0.674
098-54	0.911	0.917	0.002	1.000	0.957
100-56	0.799	0.799	0.001	1.000	0.894
101-57	0.349	0.327	0.003	0.986	0.568
102-58	0.922	0.928	0.041	0.602	0.748
103-59	0.964	0.973	0.030	0.713	0.833
104-60	0.920	0.927	0.017	0.848	0.886
105-61	0.492	0.477	0.019	0.830	0.629
106-62	0.865	0.868	0.077	0.232	0.449
107-63	0.973	0.983	0.031	0.707	0.833
110-70	0.447	0.430	0.010	0.922	0.629
111-71	0.601	0.591	0.016	0.861	0.713
112-72	0.606	0.597	0.008	0.936	0.747
113-73	0.566	0.555	0.025	0.767	0.653
114-74	0.664	0.657	0.014	0.881	0.761
115-75	0.622	0.613	0.029	0.724	0.666
117-77	0.824	0.826	0.077	0.235	0.440
118-78	0.621	0.613	0.005	0.972	0.772
			$Q_{D(MIC-IC_{50})}{}^2 = 0.629$		

Table 7. Results of the Desirability-Based MOOP Process

predictors optimum level		
JGI6 = 0.058539124	R4e+ =0.215402953	RDF020e = 6.533512527
MATS3e = 0.097921819	R5p = 0.560622	RDF050e = 21.75996
GATS5p = 2.71639566	G(F···F) = −5.395274574	Mor05m = −6.618889553
FDI = 0.996478400	H4m = 0.836178947	Mor14v = −0.049636561
Mor24v = 0.095266	D/Dr06 = 202.3135	HATS3m = 0.049289
H6v = 0.266748712	BELp1 = 2.022804936	HATS3e = 0.242572857

Table 8. Optimal Set of Weights

variable	w_i	relative importance (%)	variable	w_i	relative importance (%)
JGI6	23.323	17.561	H4m	1.573	6.019
MATS3e	−1.259	4.517	D/Dr06	−0.001	5.184
GATS5p	1.190	5.817	BELp1	11.365	11.215
FDI	−9.772	0.000	RDF020e	0.026	5.199
Mor24v	3.710	7.153	RDF050e	−0.019	5.175
H6v	4.903	7.787	Mor05m	0.013	5.192
R4e+	−1.053	4.626	Mor14v	0.560	5.482
R5p	−6.980	1.481	HATS3m	−9.248	0.278
G(F..F)	0.052	5.213	HATS3e	−5.811	2.101

resulting best-fit models are given in Table 4, together with the statistical regression parameters. The computed DRAGON molecular descriptors (GA selected and included on the respective MLR models) for the 95 training compounds are shown in the Supporting Information (see Table SI2).

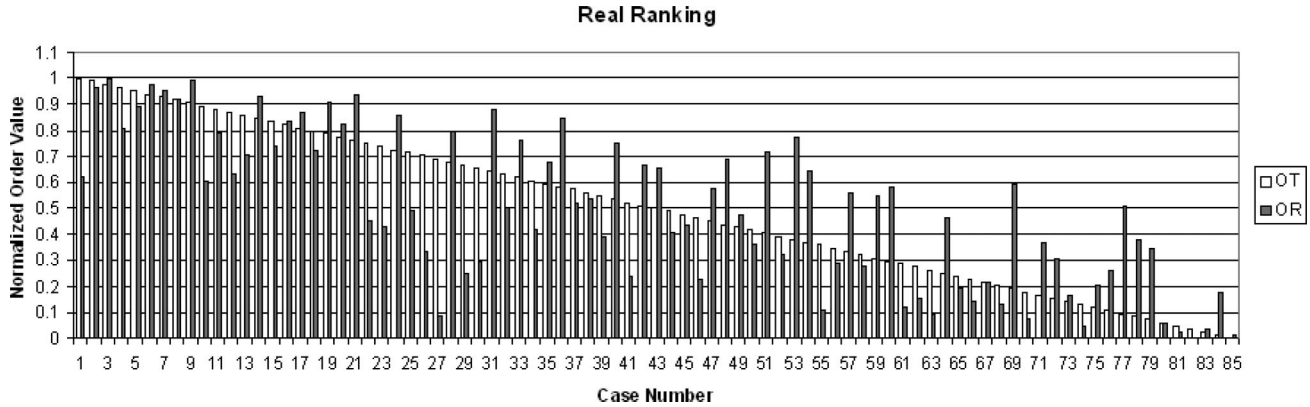


Figure 3. Δ_i-Based ranking of the fluoroquinolone library.

Table 9. Δ_i, ${}^D\Delta_i$, and D_i Values of the Library of Compounds Used for Ranking

compound ID	Δ_i	${}^D\Delta_i$	predicted $D_{\text{MIC-IC}_{50}}$	compound ID	Δ_i	${}^D\Delta_i$	predicted $D_{\text{MIC-IC}_{50}}$
004-4-ciprofloxacin	0.305	0.993	0.956	064-30E	1.221	0.766	0.793
006-6-tosufloxacin	0.330	0.987	0.968	065-30F	0.718	0.891	0.885
010-10	2.764	0.382	0.452	066-31A	0.359	0.980	0.882
014-15	0.801	0.870	0.751	067-31B	1.241	0.761	0.717
015-16	0.927	0.839	0.788	068-31C	0.871	0.853	0.733
016-17	1.416	0.717	0.776	070-31E	0.947	0.834	0.769
018-19	0.463	0.954	0.943	071-31F	0.765	0.879	0.780
019-20	0.510	0.943	0.959	073-32B	1.130	0.788	0.796
020-21	1.274	0.753	0.793	074-32C	1.123	0.790	0.709
021-22	0.919	0.841	0.901	075-32D	0.970	0.828	0.826
022-23A	0.528	0.938	0.806	077-32F	0.708	0.893	0.848
023-23B	1.132	0.788	0.777	078-33B	1.205	0.770	0.820
024-23C	0.411	0.967	0.904	079-34B	2.903	0.348	0.699
025-23D	1.040	0.811	0.761	080-35B	0.988	0.824	0.894
027-23F	0.680	0.900	0.856	081-36B	1.729	0.640	0.715
028-24A	0.730	0.888	0.930	082-37B	1.703	0.646	0.695
029-24C	0.576	0.926	0.879	083-38A	1.046	0.809	0.857
030-24D	0.829	0.863	0.882	084-38B	1.589	0.674	0.803
031-24E	1.060	0.806	0.823	085-39A	2.044	0.561	0.596
032-24F	0.701	0.895	0.896	086-39B	4.303	0.000	0.358
033-25A	1.004	0.820	0.790	088-41A	1.117	0.792	0.763
034-25B	1.713	0.644	0.508	090-42A	1.214	0.768	0.729
037-25E	1.425	0.715	0.699	092-48	0.745	0.884	0.770
038-25F	0.859	0.856	0.713	093-49	0.486	0.949	0.920
040-26D	1.658	0.657	0.737	094-50	1.120	0.791	0.771
041-26E	1.904	0.596	0.756	095-51	0.672	0.902	0.929
042-26F	0.631	0.912	0.811	096-52	1.279	0.751	0.664
043-27A	1.723	0.641	0.707	098-54	0.444	0.959	0.957
044-27B	2.595	0.424	0.000	100-56	0.746	0.884	0.895
046-27D	1.405	0.720	0.647	102-58	1.183	0.775	0.738
047-27E	1.572	0.679	0.667	103-59	0.656	0.906	0.838
048-27F	1.359	0.731	0.685	104-60	0.680	0.900	0.890
049-28A	1.912	0.594	0.753	105-61	0.825	0.864	0.641
052-28D	1.509	0.694	0.707	106-62	2.219	0.518	0.446
054-28F	1.784	0.626	0.789	107-63	1.159	0.781	0.840
055-29B	1.132	0.788	0.835	110-70	1.630	0.664	0.637
056-29C	1.012	0.818	0.791	111-71	1.050	0.808	0.712
057-29D	1.061	0.806	0.822	112-72	1.142	0.785	0.753
058-29E	0.279	1.000	0.811	113-73	1.205	0.770	0.655
059-29F	0.711	0.893	0.905	114-74	1.631	0.664	0.754
061-30B	1.191	0.773	0.872	115-75	1.495	0.698	0.675
062-30C	1.278	0.752	0.800	118-78	0.739	0.886	0.775
063-30D	0.945	0.834	0.860				

As can be noticed, the models are good in both statistical significance and predictive ability (see Table 4). Good overall quality of the models is revealed by the large F and small p values, satisfactory ρ values ($\rho = 5$), and R^2 and Adj.R^2 (goodness of fit) values ranging from 0.75 to 0.779 and 0.721 to 0.753, respectively; as well as $Q_{\text{LOO}}{}^2$ (predictivity) values between 0.686 and 0.725.

The next step is to find out if the basic assumptions of MLR analysis are fulfilled. No violations of such assumptions were found that could compromise the reliability of the resulting predictions. A deeper discussion about the fulfilling of the parametric assumptions for the MLR models is included in the Supporting Information (check Table SI4).

Figure 4. Ranking attained for the 10% of the library of compounds.

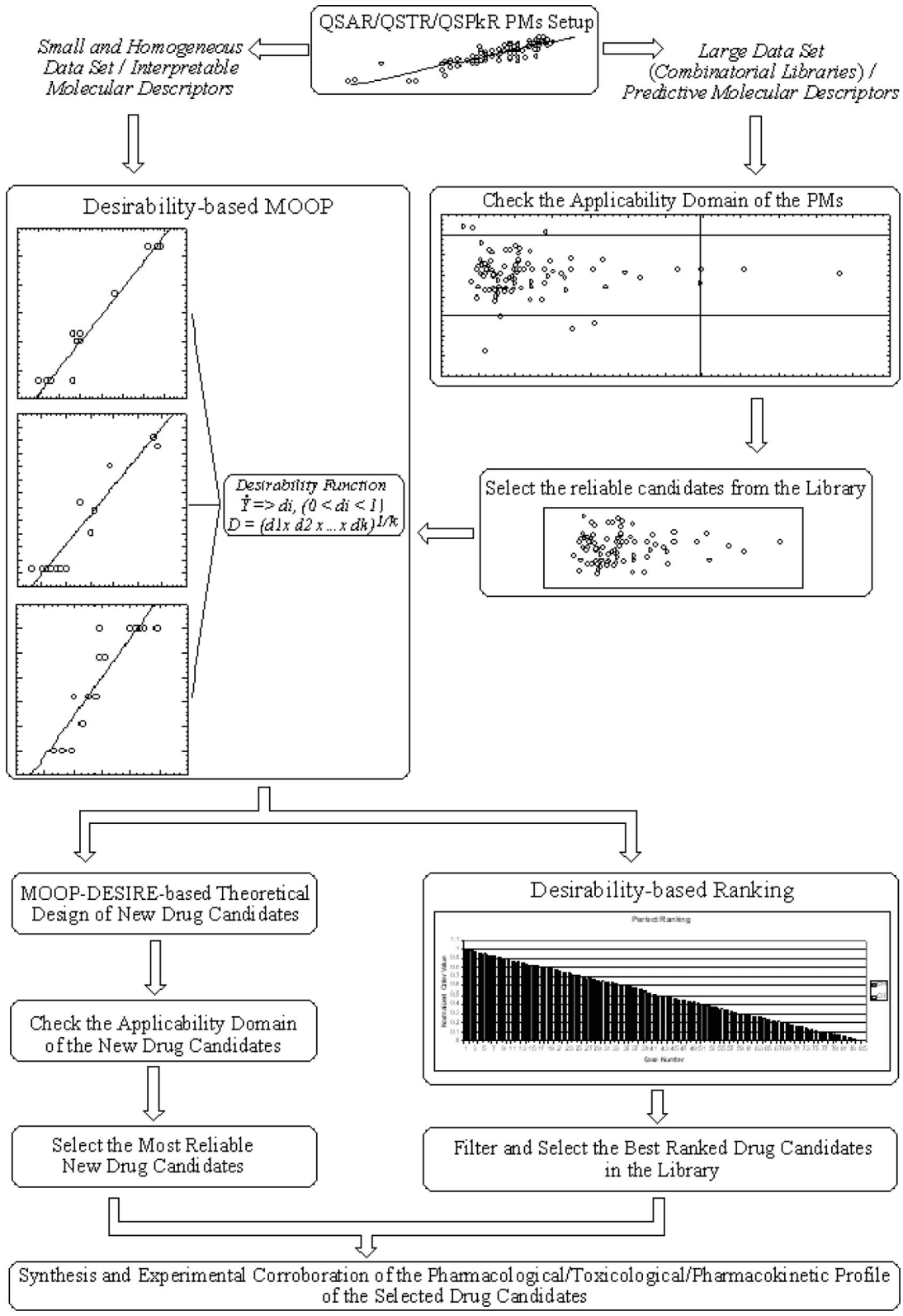


Figure 5. MOOP-DESIRE-based rational drug discovery and development.

Another aspect to consider in PMs development is to establish their applicability domain. The leverage values (h) and standardized residuals (Std. Res.) related to three PMs for the 95 training compounds are shown in Table SI3 (Supporting Information), whereas Figure SI1 (Supporting Information) shows the corresponding leverage plots. From these plots, the applicability domain is established inside a squared area within ± 2 standard deviations and a leverage threshold h^* of 0.347. (Notice that each model was fitted using 95 training compounds and included 11 adjustable parameters: 10 DRAGON descriptors plus the intercept.)

So far, we have demonstrated the satisfactory accuracy and the acceptable predictive ability of the developed PMs. We may now thus proceed with an adequate level of

Table 10. Residual Analysis of the Original and Desirability-Transformed Responses Employed for the MLR Modeling, MOOP, and Estimation of Weights Used for Ranking Based on a Nonlinear Curve-Fitting Algorithm

	residuals										
	MLR modeling					MOOP					ranking
	FIT		LOO-CV			FIT			LOO-CV		
compound ID	1/1 + MIC	1/1 + IC_{50}	1/1 + MIC	1/1 + IC_{50}	d_{MIC}	$d_{IC_{50}}$	$D_{MIC\text{-}IC_{50}}$	d_{MIC}	$d_{IC_{50}}$	$D_{MIC\text{-}IC_{50}}$	$(D - {}^{D}\Delta_i)$
004-4	0.001	0.013	0.001	0.014	0.001	−0.006	−0.002	0.001	−0.006	−0.002	−0.037
006-6	−0.014	0.014	−0.018	0.016	−0.014	−0.059	−0.036	−0.019	−0.059	−0.039	−0.019
007-7	0.224	0.031	0.234	0.032	0.236	−0.319	−0.184	0.246	−0.335	−0.189	
008-8	0.226	−0.011	0.233	−0.012	0.237	0.109	0.182	0.245	0.120	0.191	
010-10	0.074	−0.004	0.094	−0.005	0.078	0.046	0.079	0.099	0.054	0.100	0.07
012-13	−0.362	0.006	−0.385	0.006	−0.380	−0.022	−0.337	−0.405	−0.022	−0.353	
014-15	0.065	0.014	0.073	0.017	0.068	−0.012	0.040	0.076	−0.012	0.045	−0.119
015-16	−0.079	−0.014	−0.087	−0.016	−0.083	0.143	0.018	−0.091	0.157	0.020	−0.051
016-17	−0.088	−0.002	−0.095	−0.003	−0.092	0.022	−0.051	−0.100	0.025	−0.054	0.059
018-19	0.002	0.000	0.002	0.000	0.002	−0.006	−0.002	0.002	−0.007	−0.003	−0.011
019-20	−0.062	−0.003	−0.067	−0.003	−0.065	0.026	−0.022	−0.070	0.029	−0.023	0.016
020-21	0.071	0.001	0.075	0.001	0.074	−0.010	0.026	0.078	−0.011	0.028	0.04
021-22	−0.103	−0.003	−0.108	−0.003	−0.108	0.031	−0.044	−0.114	0.032	−0.046	0.06
022-23A	0.026	0.005	0.033	0.005	0.027	−0.046	−0.010	0.034	−0.050	−0.009	−0.132
023-23B	0.114	−0.018	0.129	−0.019	0.119	0.177	0.148	0.136	0.189	0.162	−0.011
024-23C	−0.027	−0.008	−0.030	−0.009	−0.029	0.088	0.030	−0.032	0.096	0.033	−0.063
025-23D	−0.011	−0.020	−0.012	−0.022	−0.012	0.210	0.094	−0.012	0.227	0.102	−0.05
026-23E	−0.230	0.011	−0.289	0.012	−0.241	−0.016	−0.336	−0.304	−0.016	−0.392	
027-23F	−0.111	−0.004	−0.126	−0.004	−0.117	0.042	−0.036	−0.133	0.044	−0.043	−0.044
028-24A	−0.056	0.000	−0.060	0.000	−0.058	−0.001	−0.028	−0.063	−0.001	−0.031	0.042
029-24C	0.106	0.024	0.113	0.026	0.111	−0.247	−0.086	0.119	−0.267	−0.091	−0.047
030-24D	0.042	0.007	0.044	0.007	0.044	−0.067	−0.015	0.046	−0.071	−0.016	0.019
031-24E	0.150	0.002	0.159	0.003	0.158	−0.013	0.085	0.167	−0.013	0.091	0.017
032-24F	0.027	−0.007	0.029	−0.008	0.029	0.073	0.052	0.031	0.076	0.055	0.001
033-25A	0.006	−0.014	0.006	−0.015	0.006	0.143	0.075	0.006	0.154	0.081	−0.03
034-25B	−0.064	0.008	−0.072	0.009	−0.040	−0.088	−0.104	−0.040	−0.095	−0.111	−0.136
036-25D	0.022	0.046	0.023	0.051	0.022	−0.465	−0.412	0.024	−0.523	−0.447	
037-25E	0.040	0.005	0.042	0.005	0.042	−0.043	0.004	0.044	−0.047	0.004	−0.016
038-25F	0.029	−0.022	0.031	−0.023	0.031	0.226	0.139	0.033	0.236	0.146	−0.143
040-26D	0.049	0.015	0.054	0.017	0.051	−0.154	−0.060	0.057	−0.173	−0.066	0.08
041-26E	0.025	0.003	0.029	0.004	0.027	−0.033	0.004	0.030	−0.039	0.004	0.16
042-26F	0.031	−0.013	0.034	−0.014	0.032	0.129	0.079	0.036	0.140	0.086	−0.101
043-27A	−0.115	0.005	−0.124	0.007	−0.121	−0.052	−0.084	−0.131	−0.073	−0.099	0.066
044-27B	0.042	−0.001	0.045	−0.001	0.045	0.000	0.000	0.048	0.000	0.000	−0.424
045-27C	−0.054	0.023	−0.061	0.027	−0.057	−0.122	−0.349	−0.058	−0.162	−0.402	
046-27D	−0.061	−0.026	−0.067	−0.031	−0.064	0.272	0.129	−0.071	0.325	0.163	−0.073
047-27E	−0.098	−0.017	−0.106	−0.019	−0.103	0.172	0.004	−0.111	0.195	0.010	−0.012
048-27F	0.024	0.002	0.025	0.003	0.024	−0.030	−0.004	0.026	−0.034	−0.005	−0.046
049-28A	0.027	−0.013	0.029	−0.016	0.029	0.139	0.077	0.031	0.163	0.089	0.159
050-28B	−0.010	0.026	−0.010	0.031	−0.010	−0.156	−0.359	−0.011	−0.205	−0.412	
051-28C	0.135	−0.022	0.151	−0.025	0.142	0.228	0.198	0.158	0.261	0.226	
052-28D	−0.071	−0.024	−0.077	−0.027	−0.075	0.247	0.074	−0.082	0.271	0.083	0.013
054-28F	−0.077	0.004	−0.083	0.005	−0.081	−0.044	−0.066	−0.086	−0.052	−0.072	0.163
055-29B	−0.074	−0.011	−0.078	−0.011	−0.058	0.110	0.037	−0.058	0.116	0.040	0.047
056-29C	−0.081	−0.016	−0.088	−0.017	−0.058	0.162	0.071	−0.058	0.172	0.077	−0.027
057-29D	0.052	−0.013	0.058	−0.014	0.054	0.136	0.097	0.061	0.142	0.103	0.016
058-29E	0.206	0.008	0.240	0.010	0.215	−0.043	0.103	0.251	−0.043	0.125	−0.189
059-29F	−0.002	−0.005	−0.002	−0.005	−0.001	0.050	0.025	−0.001	0.053	0.026	0.012
061-30B	0.014	−0.014	0.016	−0.015	0.014	0.149	0.085	0.017	0.159	0.092	0.099
062-30C	−0.011	−0.017	−0.013	−0.019	−0.012	0.172	0.079	−0.013	0.188	0.086	0.048
063-30D	0.002	0.000	0.002	0.000	0.002	0.004	0.003	0.003	0.004	0.004	0.026
064-30E	−0.113	0.009	−0.131	0.010	−0.119	0.000	−0.079	−0.138	0.000	−0.091	0.027
065-30F	0.071	0.023	0.080	0.025	0.075	−0.020	0.032	0.083	−0.020	0.037	−0.006
066-31A	−0.014	−0.002	−0.016	−0.002	−0.015	0.014	−0.002	−0.017	0.016	−0.002	−0.098
067-31B	−0.055	−0.002	−0.058	−0.002	−0.058	0.019	−0.012	−0.061	0.021	−0.012	−0.044
068-31C	0.028	0.011	0.031	0.013	0.029	−0.112	−0.062	0.033	−0.124	−0.068	−0.12
070-31E	0.120	0.035	0.131	0.039	0.126	−0.351	−0.118	0.138	−0.395	−0.130	−0.065
071-31F	0.042	−0.013	0.046	−0.014	0.044	0.129	0.085	0.048	0.140	0.093	−0.099
073-32B	−0.066	−0.016	−0.072	−0.017	−0.069	0.171	0.064	−0.075	0.177	0.066	0.008
074-32C	0.066	−0.004	0.078	−0.004	0.069	0.036	0.050	0.081	0.039	0.057	−0.081
075-32D	−0.021	−0.006	−0.023	−0.007	−0.022	0.060	0.018	−0.025	0.065	0.020	−0.002
077-32F	−0.073	0.000	−0.079	0.000	−0.077	0.005	−0.040	−0.083	0.005	−0.043	−0.045
078-33B	0.074	0.001	0.090	0.001	0.078	−0.007	0.039	0.094	−0.008	0.047	0.05
079-34B	0.030	−0.013	0.051	−0.020	0.031	0.129	0.074	0.053	0.203	0.119	0.351
080-35B	−0.058	0.005	−0.074	0.005	−0.061	−0.015	−0.041	−0.078	−0.015	−0.051	0.07
081-36B	0.031	0.003	0.036	0.004	0.033	−0.033	0.010	0.038	−0.033	0.014	0.075
082-37B	−0.074	−0.006	−0.089	−0.007	−0.078	0.064	−0.028	−0.094	0.076	−0.034	0.049
083-38A	−0.032	0.013	−0.033	0.015	−0.033	−0.130	−0.081	−0.035	−0.145	−0.089	0.048

Table 10. Continued

	residuals										
	MLR modeling					MOOP					ranking
	FIT		LOO-CV			FIT			LOO-CV		
compound ID	1/1 + MIC	1/1 + IC_{50}	1/1 + MIC	1/1 + IC_{50}	d_{MIC}	$d_{IC_{50}}$	$D_{MIC\text{-}IC_{50}}$	d_{MIC}	$d_{IC_{50}}$	$D_{MIC\text{-}IC_{50}}$	$(D - {}^{D}\Delta_i)$
084-38B	−0.038	−0.008	−0.045	−0.010	−0.040	0.084	0.013	−0.046	0.103	0.018	0.129
085-39A	0.124	0.007	0.138	0.009	0.130	−0.072	0.075	0.145	−0.072	0.088	0.035
086-39B	0.030	−0.001	0.046	−0.001	0.031	0.011	0.024	0.048	0.017	0.038	0.358
088-41A	−0.008	−0.017	−0.009	−0.019	−0.008	0.180	0.100	−0.009	0.193	0.108	−0.029
090-42A	0.051	−0.012	0.055	−0.013	0.054	0.125	0.085	0.058	0.135	0.091	−0.039
092-48	0.012	0.001	0.016	0.001	0.013	−0.015	0.001	0.018	−0.016	0.003	−0.114
093-49	−0.190	0.003	−0.207	0.003	−0.200	−0.019	−0.123	−0.218	−0.019	−0.133	−0.029
094-50	−0.040	−0.003	−0.061	−0.003	−0.042	0.024	−0.005	−0.064	0.029	−0.012	−0.02
095-51	0.026	0.008	0.029	0.008	0.027	−0.079	−0.029	0.030	−0.083	−0.029	0.027
096-52	0.007	0.014	0.008	0.016	0.008	−0.142	−0.103	0.009	−0.157	−0.113	−0.087
098-54	0.049	0.012	0.051	0.012	0.050	−0.114	−0.033	0.053	−0.119	−0.033	−0.002
100-56	0.119	0.007	0.127	0.009	0.125	−0.072	0.031	0.134	−0.081	0.032	0.011
101-57	−0.256	0.001	−0.311	0.002	−0.269	−0.015	−0.492	−0.327	−0.019	−0.546	
102-58	0.064	0.020	0.068	0.022	0.067	−0.201	−0.119	0.072	−0.219	−0.129	−0.037
103-59	−0.034	−0.012	−0.038	−0.013	−0.035	0.125	0.053	−0.040	0.137	0.058	−0.068
104-60	−0.016	−0.006	−0.019	−0.007	−0.017	0.061	0.023	−0.021	0.071	0.027	−0.01
105-61	0.026	−0.014	0.032	−0.016	0.027	0.135	0.068	0.033	0.155	0.080	−0.223
106-62	0.103	0.005	0.115	0.006	0.109	−0.056	−0.036	0.122	−0.062	−0.039	−0.072
107-63	−0.002	−0.007	−0.002	−0.008	−0.002	0.070	0.039	−0.003	0.081	0.046	0.059
110-70	0.028	0.005	0.041	0.005	0.029	−0.046	0.004	0.042	−0.052	0.012	−0.027
111-71	−0.069	−0.012	−0.077	−0.013	−0.073	0.116	−0.003	−0.081	0.124	−0.004	−0.096
112-72	0.122	0.007	0.135	0.008	0.128	−0.073	0.042	0.141	−0.080	0.048	−0.032
113-73	0.055	−0.002	0.059	−0.002	0.058	0.014	0.040	0.062	0.016	0.042	−0.115
114-74	−0.020	0.006	−0.023	0.007	−0.022	−0.065	−0.041	−0.024	−0.078	−0.048	0.09
115-75	−0.027	−0.008	−0.030	−0.010	−0.029	0.086	0.020	−0.031	0.107	0.029	−0.023
117-77	−0.039	0.018	−0.043	0.023	−0.040	−0.188	−0.393	−0.045	−0.235	−0.440	
118-78	0.002	0.000	0.004	−0.001	0.002	0.006	0.003	0.004	0.011	0.006	−0.111
residual mean	**0.00006**	**0.00001**	**−0.0003**	**0.00006**	**0.00080**	**0.01150**	**−0.01513**	**0.00070**	**0.01260**	**−0.01579**	**−0.00921**

confidence to the simultaneous optimization of the antibacterial and cytotoxic properties for the set of compounds.

First, the predicted values for each property were used to fit a model containing all the independent variables applied in modeling the original properties. In so doing, one is able to discriminate opposite objectives like efficacy (antibacterial activity) and toxicity (cytotoxicity) with partial overlap of the descriptors set used to built the PMs. (Notice that both PMs share H4m and G(F···F); see Table 4.)

Once the models have been set up, the desirability functions for each property (d_i) might be specified. To obtain candidate(s) with high antibacterial potency (MIC = 1/1 + MIC) and low cytotoxicity (IC_{50} = 1/1 + IC_{50}), 1/1 + MIC should be maximized (eq 5), and 1/1 + IC_{50} should be minimized (eq 6). In addition, the individual d_i values for the antibacterial and cytotoxicity properties were determined by setting the L_i, U_i, and T_i values, as described previously. Then, the two d_i values were combined into the single overall desirability D by means of eq 3.

The expected and predicted desirability values attributable to each response plus the overall desirability for the training set are depicted in Table 5. In addition, the LOO-CV predicted values and the desirability values for each response, along with the overall desirability values are shown in Table 6. As can be seen, the overall desirability function exhibits good statistical quality as indicated by the R_D^2 and Adj.R_D^2 values (∼0.7). Moreover, a Q_D^2 value of 0.63 provides an adequate level of reliability on the method in predicting D.

Finally, the optimization of the overall desirability was carried out to obtain the levels of the descriptors included in the PMs that simultaneously produce the most desirable combination of the properties. The results of the desirability-based MOOP process are detailed in Table 7. Here are shown the levels of the predictive variables required to reach a highly desirable ($D_{MIC-IC_{50}}$ = 1) fluoroquinolone-like candidate with the best possible compromise between antibacterial and cytotoxicity properties.

3.2. MOOP-DESIRE$_{(PHARM\text{-}TOX)}$-Based Ranking and Filtering. Once found, the levels of the predictive variables required to reach a highly desirable fluoroquinolone-like candidate are used as a pattern to rank the library of flouroquinolones. Previously, 10 compounds were removed from the initial library because of their outlier nature to avoid their negative influence in the ulterior data-fitting process.

Through a nonlinear curve-fitting process implemented in MATLAB, we found the optimal set of weighs w_i required to minimize the differences between descriptions (Δ_i) and solutions (D_i) in the library of compounds to rank.

Next, Δ_i is used as a ranking criterion to obtain an ordered list of the flouroquinolones. The list start with the compound most similar to the optimal fluoroquinolone-like candidate previously determined by the process of simultaneous optimization of antibacterial and cytotoxicity properties (see the levels of the predictive variables found for the optimal candidate in Table 7). The computed values of D_i, Δ_i, and the normalized values of Δ_i (${}^{D}\Delta_i$) of the library of compounds used for ranking are detailed in Table 9.

On the basis of Δ_i, it is possible to reach a ranking of the flouroquinolones library with a corrected ranking quality index (Ψ^*) of 0.313, representing a percentage of ranking quality ($R_{\%}$) of 68.7. This ranking compared with the perfect ranking is shown in Figure 3.

As can be noted, the quality of the ranking attained ($R_{\%} = 68.7$) is similar to the predictability values exhibited in the PMs as well as in the MOOP process ($Q_{MIC}^2 = 0.693$, $Q_{IC_{50}}^2 = 0.686$, $Q^2_{D_{MIC\text{-}IC50}} = 0.629$). This fact indicates that the quality of both process (desirability-based MOOP and ranking) are strongly dependent on the quality of the initial set of PMs. In addition, the similarity exhibited between these values suggests that the ranking algorithm reflects the quality of the PMs and the MOOP process on which it is based. The correspondence between the correlation results (low and similar residuals for each case) of the nonlinear curve-fitting process and the MLR modeling and the MOOP process support this choice. This can be verified in Table 10 (see also Tables 5, 6, and 9).

On the other hand, the main goal of ranking a library of compounds according to a pharmaceutically optimal candidate is to filter the fragment containing the most promising candidates (the closest and consequently more similar to the optimal candidate) to propose these for synthesis and biological assessment. Thus, if the best 10% (the best 9 candidates) of the library of flouroquinolones is proposed to be included on the drug development process, the probability of finding a promising candidate is increased. This fraction exhibits a percentage of quality ranking of 82.74 ($\Psi^* = 0.173$). The ranking of this fragment is shown in Figure 4.

Filtering the most promising candidates having the best compromise between pharmacological, toxicological, and pharmacokinetic properties confers to the process of discovery and development of new drugs an elevated degree of rationality which is not possible to reach via traditional QSAR which optimize sequentially each pharmaceutical property. The sequential optimization of the properties involved in the final pharmaceutical profile of a drug implies to overlook the rest of the properties equally determining on the success of the candidate as a drug or at least to leave to the serendipity to found a candidate with acceptable profiles of these properties simultaneously. That is, a potent candidate once identified via QSAR has a high probability of being discarded later as a drug because of unacceptable toxicological or pharmacokinetic profiles with the useless expenses of time and resources in synthesis and pharmacological assays.[69] Equally improvable is the choice of using a jury of models (pharmacological (QSAR), toxicological (QSTR) and pharmacokinetics (QPkR) prediction models) since that is not very probable to find a candidate with all the properties simultaneously optimized (in this way each property is optimized separately), and if this happens, the results is more by chance than the fruit of a rational drug development strategy.

As have been illustrated above, the MOOP-DESIRE methodology can be used as rational strategy of filtering new drug candidates from combinatorial libraries, always considering those candidates included on the applicability domain of the PMs on which are based the process of MOOP and ranking. In situations like this, where the main goal is the ranking and filtering, it is advisable to use descriptors leading to highly predictive structure−desirability relationships rather than interpretable descriptors to ensure the accuracy of the predictions and therefore, an accurate assessment of the molecule's overall desirability. This type of analysis is more appropriate for early stages of the drug development process. In contrast, the use of small and homogeneous data sets is more suitable for later stages of the drug development process, once a lead has been identified, rather than for early stages. Actually, specific structural modifications can be made over the lead according to the results of the optimization process. For this, the use of clearly defined structural or physicochemical descriptors can led to interpretable structure−desirability relationships which can be used to design new candidates with an improved pharmaceutical profile (see ref33). Figure 5 schematically summarizes the use of the MOOP-DESIRE methodology to aid the rational discovery and development of new drugs.

Acknowledgment. The authors acknowledge the Portuguese Fundação para a Ciência e a Tecnologia (FCT) (SFRH/BD/30698/2006 andSFRH/BPD/27167/2006) for financial support.

Supporting Information Available. The chemical structures and properties values of the library used in this work and details about the applicability domain and the parametrical assumptions of the MLR PMs, as well as a copy of the functions employed in the nonlinear curve-fitting process implemented in the "lsqcurvefit" function of MATLAB program, for the library of compounds. This material is available free of charge via the Internet at http://pubs.acs.org.

References and Notes

(1) Ekins, S.; Boulanger, B.; Swaan, P. W.; Hupcey, M. A. *J. Comput.-Aided Mol. Des.* **2002**, *16* (5−6), 381–401.
(2) Seifert, M. H. J.; Wolf, K.; Vitt, D. *Drug Discovery Today* **2003**, *1* (4), 143–149.
(3) Jorgensen, W. L. *Science* **2004**, *303* (5665), 1813–1818.
(4) Brown, N.; Lewis, R. A. *Curr. Opin. Drug Discovery Dev.* **2006**, *9* (4), 419–424.
(5) Hansch, C. *J. Med. Chem.* **1976**, *19* (1), 1–6.
(6) Walters, W. P.; Stahl, M. T.; Murcko, M. A. *Drug Discovery Today* **1998**, *3*, 160–178.
(7) Fox, S.; Farr-Jones, S.; Yund, M. A. *J. Biomol. Screening* **1999**, *4*, 183–186.
(8) Young, S.; Li, J. *Innovative Pharm. Technol.* **2000**, *28*, 24–26.
(9) Fukunaga, J. Y.; Hansch, C.; Steller, E. E. *J. Med. Chem.* **1976**, *19* (5), 605–11.
(10) Mayer, J. M.; van de Waterbeemd, H. *Environ. Health Perspect.* **1985**, *61*, 295–306.
(11) Moriguchi, I.; Hirano, H.; Hirono, S. *Environ. Health Perspect.* **1996**, *104* (Suppl 5), 1051–8.
(12) Estrada, E. *SAR QSAR Environ. Res.* **2000**, *11* (1), 55–73.
(13) Vilar, S.; Estrada, E.; Uriarte, E.; Santana, L.; Gutierrez, Y. *J. Chem. Inf. Model.* **2005**, *45* (2), 502–14.
(14) Marrero-Ponce, Y.; Marrero, R. M.; Torrens, F.; Martinez, Y.; Bernal, M. G.; Zaldivar, V. R.; Castro, E. A.; Abalo, R. G. *J. Mol. Model.* **2006**, *12* (3), 255–71.
(15) Helguera, A. M.; Cabrera Perez, M. A.; Gonzalez, M. P. *J Mol Model* **2006**, 1–12.

(16) Gonzalez-Diaz, H.; Cruz-Monteagudo, M.; Molina, R.; Tenorio, E.; Uriarte, E. *Bioorg. Med. Chem.* **2005**, *13* (4), 1119–29.
(17) Gonzalez-Diaz, H.; Cruz-Monteagudo, M.; Vina, D.; Santana, L.; Uriarte, E.; De Clercq, E. *Bioorg. Med. Chem. Lett.* **2005**, *15* (6), 1651–7.
(18) Cruz-Monteagudo, M.; Gonzalez-Diaz, H.; Borges, F.; Gonzalez-Diaz, Y. *Bull. Math. Biol.* **2006**, *68* (7), 1555–72.
(19) Cruz-Monteagudo, M.; Borges, F.; Perez Gonzalez, M.; Cordeiro, M. N. *Bioorg. Med. Chem.* **2007**, *15* (15), 5322–39.
(20) Cruz-Monteagudo, M.; Cordeiro, M. N.; Borges, F. *J. Comput. Chem.* **2008**, *29* (4), 533–49.
(21) Cruz-Monteagudo, M.; Gonzalez-Diaz, H.; Aguero-Chapin, G.; Santana, L.; Borges, F.; Dominguez, E. R.; Podda, G.; Uriarte, E. *J. Comput. Chem.* **2007**, *28* (11), 1909–23.
(22) Gonzalez-Diaz, H.; Aguero, G.; Cabrera, M. A.; Molina, R.; Santana, L.; Uriarte, E.; Delogu, G.; Castanedo, N. *Bioorg. Med. Chem. Lett.* **2005**, *15* (3), 551–7.
(23) Prado-Prado, F. J.; Gonzalez-Diaz, H.; Santana, L.; Uriarte, E. *Bioorg. Med. Chem.* **2007**, *15* (2), 897–902.
(24) Gonzalez-Diaz, H.; Prado-Prado, F. J.; Santana, L.; Uriarte, E. *Bioorg. Med. Chem.* **2006**, *14* (17), 5973–80, Unify QSAR approach to antimicrobials. Part 1: predicting antifungal activity against different species.
(25) Gonzalez-Diaz, H.; Prado-Prado, F. J. *J. Comput. Chem.* **2008**, *29* (4), 656–67.
(26) Nicolaou, A. C.; Brown, N.; Pattichis, C. S. *Curr. Opin. Drug Discovery Dev.* **2007**, *10* (3), 316–24.
(27) Yann, C.; Siarry, P. *Multiobjective Optimization: Principles and Case Studies*; Springer-Verlag: Berlin, Germany, 2004.
(28) Nicolotti, O.; Gillet, V. J.; Fleming, P. J.; Green, D. V. *J. Med. Chem.* **2002**, *45* (23), 5069–80.
(29) Stockfisch, T. P. *J. Chem. Inf. Comput. Sci.* **2003**, *43* (5), 1608–13.
(30) Rao, S. N.; Stockfisch, T. P. *J. Chem. Inf. Comput. Sci.* **2003**, *43* (5), 1614–22.
(31) Huang, J.; Ling, C. X., Rank Measures for Ordering. In *PKDD*; Jorge, A. e. a, Ed; Springer-Verlag Berlin: Heidelberg, Germany, 2005; pp 503−10.
(32) Derringer, G.; Suich, R. *J. Quality Technol.* **1980**, *12* (4), 214–9.
(33) Cruz-Monteagudo, M.; Borges, F.; Cordeiro, M. N. *J. Comput. Chem.* **2008**, *29* (14), 2445–59.
(34) Suto, M. J.; Domagala, J. M.; Roland, G. E.; Mailloux, G. B.; Cohen, M. A. *J. Med. Chem.* **1992**, *35*, 4745–50.
(35) Gracheck, S. J.; Mychalonka, M.; Gambino, L.; Cohen, M.; Roland, G. E.; Ciaravino, V.; Worth, D.; Theiss, J. C.; Heifetz, C. In *Correlations of Quinolone Inhibition Endpoints vs Bacterial DNA Gyrase, Topoisomerase I and II, Cell Clonogenic Survival and In Vitro Micronuclei Induction*; 31st Interscience Conference on Antimicrobial Agents and Chemotherapy, Chicago, Illinois; 1991; American Society for Microbiology: Chicago, IL.
(36) Holden, H. H.; Barret, J. F.; Huntington, C. M.; Muehlbauer, P. A.; Wahrenburg, M. G. *Environ. Mol. Mutagen.* **1989**, *13*, 238–25.
(37) Hoshino, K.; Sato, K.; Akahane, K.; Yoshida, A.; Hayakawa, I.; Sato, M.; Une, T.; Osada, Y. *Antimicrob. Agents Chemother.* **1991**, *35*, 309–12.
(38) Moreau, N. J.; Robaux, N.; Baron, L. *Antimicrob. Agents Chemother.* **1990**, *34*, 1955–60.
(39) Kohlbrenner, W. E.; Wideburg, N.; Weigl, D.; Saldivar, A.; Chu, D. T. W. *Antimicrob. Agents Chemother.* **1992**, *36*, 81–6.
(40) Ciaravino, V.; Suto, M. J.; Theiss, J. C. *Mutat. Res.* **1992**, *298*, 227–36.
(41) Cohen, M. A.; Griffen, T. J.; Bien, P. A.; Heifetz, C. L.; Domagala, J. M. *Antimicrob. Agents Chemother.* **1985**, *28*, 766–72.
(42) CambridgeSoft, ChemDraw Ultra, 9.0, 2004.
(43) Burkert, U.; Allinger, N. L., *Molecular Mechanics*; Americac Chemica Society: Washington, D.C., 1982.
(44) Clark, T., *Computational Chemistry*; Wiley: New York, 1985.
(45) Frank, J. *MOPAC*, version 6.0; Seiler Research Laboratory, U.S. Air Force Academy: Colorado Springs, CO, 1993.
(46) Todeschini, R.; Consonni, V.; Pavan, M. *DRAGON*, version 2.1; Milano Chemometrics: Milano, 2002.
(47) Todeschini, R.; Consonni, V. *Handbook of Moleccular Descriptors*; Wiley-VCH: Weinheim, Germany, 2000.
(48) Leardi, R.; Boggia, R.; Terrile, M. *J. Chemom.* **1992**, *6*, 267–81.
(49) Yasri, A.; Hartsough, D. *J. Chem. Inf. Comput. Sci.* **2001**, *41* (5), 1218–27.
(50) Hou, T. J.; Wang, J. M.; Liao, N.; Xu, X. J. *J. Chem. Inf. Comput. Sci.* **1999**, *39* (5), 775–81.
(51) Hasegawa, K.; Kimura, T.; Funatsu, K. *J. Chem. Inf. Comput. Sci.* **1999**, *39* (1), 112–20.
(52) Barbosa de Oliveira, D.; Gaudio, A. C. *BuildQSAR*, Physics Department-CCE, University of Espírito Santo: Vitória ES, Brasil, 2000.
(53) Barbosa de Oliveira, D.; Gaudio, A. C. *QSAR* **2000**, *19*, 599–601.
(54) Statsoft_Inc. *STATISTICA*, 6.0 for Windows, 2001.
(55) Eriksson, L.; Jaworska, J.; Worth, A. P.; Cronin, M. T.; McDowell, R. M.; Gramatica, P. *Environ. Health Perspect.* **2003**, *111* (10), 1361–75.
(56) Stewart, J.; Gill, L., *Econometrics*, 2nd ed.; Prentice Hall: London, 1998.
(57) Kutner, M. H.; Nachtsheim, C. J.; Neter, J.; Li, W. Multicollinearity and its effects. In *Applied Linear Statistical Models*, 5th ed.; McGraw Hill: New York, 2005; pp 27889.
(58) Atkinson, A. C., *Plots, Transformations and Regression*; Clarendon Press: Oxford, U.K., 1985.
(59) De Boor, C. *A Practical Guide to Splines*; Springer-Verlag: New York, 1978.
(60) Gerald, C. F.; Wheatley, P. O., *Applied Numerical Analysis*, 4th ed.; Addison Wesley: Reading, MA, 1989.
(61) Kutner, M. H.; Nachtsheim, C. J.; Neter, J.; Li, W., *Applied Linear Statistical Models*; 5th ed.; McGraw Hill: New York, 2005.
(62) Nelder, J. A.; Mead, R. *Comput. J.* **1965**, *7*, 308–13.
(63) Fletcher, R.; Reeves, C. M. *Comput. J.* **1964**, *7*, 149–54.
(64) Hooke, R.; Jeeves, T. A. *J. Assoc. Comput. Machine* **1961**, *8*, 212–29.
(65) Watson, I.; Marir, F. Case-Based Reasoning: A Review. *The Knowledge Engineering Review*; Cambridge University Press: Cambridge, U.K., 1994, Vol. 9.
(66) Coleman, T. F.; Li, Y. *SIAM J. Optim.* **1996**, *6*, 418–45.
(67) Coleman, T. F.; Li, Y. *Math Program* **1994**, *67* (2), 189–224.
(68) *MATLAB*, 7.2; The MathWorks, Inc.: Natick, MA, 2006.
(69) Drews, J. *Drug Discovery Today* **1998**, *3*, 491–4.

CC800115Y