Tese apresentada à Faculdade de Farmácia da Universidade do Porto para admissão as provas de Doutoramento em Ciências Farmacêuticas - Especialidade Toxicologia

MULTI-OBJECTIVE OPTIMIZATION BASED ON DESIRABILITY ESTIMATION OF SEVERAL INTERRELATED RESPONSES (MOOP-DESIRE): A COMPUTER-AIDED METHODOLOGY FOR MULTI-CRITERIA DRUG DISCOVERY

MAYKEL CRUZ MONTEAGUDO

Trabalho realizado sob a orientação da

Professora Doutora Maria Fernanda Martins Borges.
Departamento de Química e Bioquímica, Faculdade de Ciências, Universidade do Porto;

Professora Doutora Maria Natália Dias Soeiro Cordeiro.
Departamento de Química e Bioquímica, Faculdade de Ciências, Universidade do Porto;

Professor Doutor Fernando Manuel Gomes Remião.
Serviço de Toxicologia, Faculdade de Farmácia, Universidade do Porto.

U.PORTO
FACULDADE DE FARMÁCIA
UNIVERSIDADE DO PORTO

Julho 2010

Thesis in Toxicology
Submitted to the Faculty of Pharmacy, University of Porto
in fulfillment of the requirements for the degree
of Doctor in Pharmaceutical Sciences

MULTI-OBJECTIVE OPTIMIZATION BASED ON DESIRABILITY ESTIMATION OF SEVERAL INTERRELATED RESPONSES (MOOP-DESIRE): A COMPUTER-AIDED METHODOLOGY FOR MULTI-CRITERIA DRUG DISCOVERY

MAYKEL CRUZ MONTEAGUDO

Thesis Advisers

Professor Maria Fernanda Martins Borges, PhD
Department of Chemistry and Biochemistry, Faculty of Sciences,
University of Porto, Portugal.

Professor Maria Natália Dias Soeiro Cordeiro, PhD
Department of Chemistry and Biochemistry, Faculty of Sciences,
University of Porto, Portugal.

Professor Fernando Manuel Gomes Remião, PhD
Department of Toxicology. Faculty of Pharmacy,
University of Porto, Portugal

U.PORTO

FACULDADE DE FARMÁCIA
UNIVERSIDADE DO PORTO

July 2010

Os estudos efectuados no âmbito desta dissertação tiveram lugar, em sua grande parte, no Serviço de Toxicologia da Faculdade de Farmácia da Universidade do Porto. Outros estudos foram realizados no Grupo de Química Teórica e Bioquímica Computacional, Departamento de Química e Bioquímica da Faculdade de Ciências.
Alguns trabalhos foram ainda realizados em estreita colaboração com o Grupo de Simulação Molecular e Desenho de Fármacos no Centro de Bioactivos Químicos, o Centro de Estudos de Química Aplicada da Faculdade de Química e Farmácia e o Centro de Estudos Informáticos da Universidade Central de Las Villas, Cuba, e com o Departamento de Química Orgânica da Universidad de Vigo, Espanha.
Esta dissertação teve o apoio financeiro da Fundação para a Ciência e Tecnologia (SFRH/BD/30698/2006).

...Thus, QSAR lives on, not only as a stand-alone technique, but even more so in disguised forms within the more popular drug design approaches of the modern era. Correlative thinking has pervaded humankind's existence for eons, evolving from the recognition of danger engendered by the hairy fellow with a rock in his hand to the present day molecular nuance of a well-placed methyl group and its predicted effect on activity. Rebirth gives rise to novel applications of the technique. To paraphrase, "QSAR is dead, QSAR is dead, long live QSAR!"

(Arthur M. Doweyko. J. Comput. Aided Mol. Des. (2008) 22:81–89)

To Adriana... and Silvana, of course.

ACKNOWLEDGEMENTS

Marie desJardins, in her guide to graduate students and advisors (*desJardins, M. How to be a good graduate student and advisor, 1994; marie@erg.sri.com*.), wrote:

"…A good advisor will serve as a mentor as well as a source of technical assistance. A mentor should provide, or help you find, the resources you need (financial, equipment, and psychological support); introduce you and promote your work to important people in the field; encourage your own interests, rather than promoting their own; be available to give you advice on the direction of your thesis and your career; and help you to find a job when you finish…"

I was lucky to get all of the above (except for a job that was not necessary); however, this work was directed not by a single advisor, but by three advisors: Prof. Maria Fernanda Martins Borges, Prof. Maria Natália D.S. Cordeiro and Prof. Fernando Manuel Gomes Remião.

I would like to thank Fernanda and Natalia for their wise guidance, constant support, sincere friendship, and most important, for their patient and understanding.

Although Chemoinformatics was not precisely their field of expertise, Fernanda gave me the freedom to pursue my interests by applying it to medicinal chemistry. Fernanda is the perfect example of an altruistic person; he will go out of his way to help others, and he has helped me many times in different ways. Fernanda have been also my mentor for Portuguese language and culture; she tried to teach me some good manners, but with only partial success.

Natalia provided me with many of the resources I needed, and beyond; and I don't know how, but she always had an encouraging phrase in the precise moment I needed. Thank you very much for that!

I want also thank specially Prof. Fernando Manuel Gomes Remião for been there in the right moment.

Furthermore, I would like to thank my colleagues from the Molecular Simulation and Drug Design Group, Aliuska, Liane, Gisselle, jabao, Guille, Migue, chiqui, Daimel, Hai, for their valuable technical contribution to this project and for their encouragement. I am cordially thankful to my friends Dr. Marta Teijeira Bautista and Luis "el gallego" for scientific input and willingness to assist whenever they could.

Finally, I want to thank all members of the Faculty of Sciences (Portugal), Chemical Bioactive Center (Cuba) and the Faculty of Chemistry and Pharmacy (Cuba), specifically to friends I have encountered in the Theoretical Chemistry and Biochemistry Computational Group and Chemical Bioactive Synthesis Group (Portugal) who have enlightened me with their knowledge and long-lasting friendship.

Finally I gratefully acknowledge support for this work from the Portuguese Fundação para a Ciência e Tecnologia (FCT) (SFRH/BD/30698/2006).

About acknowledgments, someone said (I do not remember where I read it) something like “I often wondered why individuals exaggerate by thanking almost everyone on the planet“. I have thanked a small fraction of the planet and ask forgiveness from those I have omitted unintentionally. Thank you all!

LIST OF ORIGINAL PAPERS AND CONGRESS PRESENTATIONS

Papers

The results achieved in this thesis are based on the following articles, which are collected in the Annexes Section and are referred in the text by Roman numerals:

I. Cruz-Monteagudo M, Borges F, Cordeiro MNDS. Desirability-Based Multi-Objective Optimization for Global QSAR Studies. Application to the Design of Novel NSAIDs with Improved Analgesic, Anti-Inflammatory and Ulcerogenic Profiles. ***Journal of Computational Chemistry*** 2008, 29, 2445–2459.

II. Cruz-Monteagudo M, Borges F, Cordeiro MNDS, Fajin JLC, Morell C, Molina RR, Cañizares-Carmenate Y, Domínguez ER. Desirability-Based Methods of Multi-Objective Optimization and Ranking for Global QSAR Studies. Filtering Safe and Potent Drug Candidates from Combinatorial Libraries. ***Journal of Combinatorial Chemistry***. 2008, 10, 897–913.

III. Cruz-Monteagudo M, The HP, Cordeiro MNDS, Borges F. Prioritizing Hits With Appropriate Trade-offs Between HIV-1 Reverse Transcriptase Inhibitory Efficacy and MT4 Blood Cells Toxicity Through Desirability-Based Multi-Objective Optimization and Ranking. ***Molecular Informatics*** 2010, 29, 303–321.

IV. Cruz-Monteagudo M, Cordeiro MNDS, Teijeira M, González MP, Borges F. Multidimensional Drug Design: Simultaneous Analysis of Binding and Relative Efficacy Profiles of N6-substituted-4´-thioadenosines A3 Adenosine Receptor Agonists. ***Chemical Biology & Drug Design*** 2010, 75, 607–618.

Congress Presentations

1. Cruz-Monteagudo M, Borges F, Cordeiro MNDS. MOOP-DESIRE-based simultaneous optimization of the analgesic, anti-inflammatory and ulcerogenic profiles of 3-(3-methylphenyl)-2-substituted amino-3H-quinazolin-4-ones. **Proceedings of the 12th International Electronic Conference on Synthetic Organic Chemistry, ECSOC-12**.CD-ROM edition ISBN 3-906980-20-0.**Universidad de Santiago de Compostela, Santiago de Compostela, Spain**. (November 2008).
   http://www.usc.es/congresos/ecsoc/12/hall_gCC/g016/g016.pdf
2. Cruz-Monteagudo, M.; CagideFajin JL, Molina Ruiz R, Cordeiro MNDS, Borges F. Filtering Safe and Potent Drug Candidates from Combinatorial Libraries throughout Desirability-Based Methods of Multi-Objective Optimization and Ranking. **1º Encontro Nacional de Química Terapêutica. Universidade do Porto, Porto, Portugal**.(November 13-15, 2008).
3. Cruz-Monteagudo M, The HP, Cordeiro MNDS, Borges F. A Multi-Objective Strategy for Ligand Based Virtual Screening. **VII European Workshop in Drug Design, Certosa di Pontignano – Siena, Italy**. (May 24-30, 2009).
4. Cruz-Monteagudo M, Cordeiro MNDS, Teijeira M, González MP, Borges F. Desirability-based simultaneous analysis of binding and relative efficacy profiles of A3 adenosine receptor agonists. **Third Joint Italian-German Purine Club Meeting Purinergic Receptors: New Frontiers for Novel Therapies. Camerino, Italy**. (July 17-20, 2009).
5. Cruz-Monteagudo M, Cordeiro MNDS, Teijeira M, González MP, Borges F. Desirability-based simultaneous analysis of binding and relative efficacy profiles of A3 adenosine receptor agonists. *Abstract published at:* ***Purinergic Signaling*** 2010, 6, 72–73.
6. Cruz-Monteagudo M, Cañizares-Carmenate Y, Borges F, Cordeiro MNDS, Fajín JLC, Morell C, Molina RR, Domínguez ER, Moreno E. Desirability-based methods of multiobjective optimization and filtering for the discovery of potent drug candidates. **7th Seminars of Advanced Studies on Molecular Design and Bioinformatics, VII SEADIMB. Faculty of Chemistry, University of Havana, Havana, Cuba.** (August 23-28, 2009).
7. Cruz-Monteagudo M, The HP, Cordeiro MNDS, Borges F. Prioritizing Hits With Appropriate Trade-offs Between HIV-1 Reverse Transcriptase Inhibitory Efficacy and MT4 Blood Cells Toxicity Through Desirability-Based Multi-Objective

Optimization and Ranking. **IV Simposio Internacional de Química. Universidad Central de Las Villas, Santa Clara, Villa Clara, Cuba**. (June 1-4, 2010).

8. Cruz-Monteagudo M, Cordeiro MNDS, Teijeira M, González MP, Borges F. Multidimensional Drug Design: Simultaneous Analysis of Binding and Relative Efficacy Profiles of N6-substituted-4´-thioadenosines A3 Adenosine Receptor Agonists. **IV Simposio Internacional de Química. Universidad Central de Las Villas, Santa Clara, Villa Clara, Cuba**. (June 1-4, 2010).

ABSTRACT

The ability to improve the pharmaceutical profile of drugs on the sole basis of their activity has been often overestimated. The adjustment of multiple criteria in hit-to-lead identification and lead optimization is considered to be a major advance in the rational drug discovery process. Thus, the development of approaches able to handle additional criteria for the early simultaneous treatment of the most important properties, potency, safety, and bioavailability, determining the pharmaceutical profile of a drug candidate, is an emergent issue in drug discovery and development.In this Thesis, it is introduced a multi-objective optimization (MOOP) method based on Derringer's desirability functions that allows conducting global QSAR studies considering simultaneously the potency, bioavailability and/or safety of a set of drug candidates. The results of the desirability-based MOOP (the levels of the predictor variables producing concurrently the best possible compromise between the properties determining an optimal drug candidate) are used for the implementation of a ranking method, 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 optimal candidate previously determined. The whole process is condensed in a methodology that we decided to name as **MOOP-DESIRE**, acronym of **M**ulti-**O**bjective **OP**timization based on the **D**esirability **E**stimation of **S**everal **I**nterrelated **RE**sponses. Their suitability for key tasks involving the use of chemoinformatics methods in drug discovery– drug design, library ranking, and virtual screening – is evaluated besides the use of Desirability Theory as a tool for the interpretation of multi-criteria prediction models. Each task was challenged through four different data sets enabling to evaluate the performance of the methodology in the corresponding task, each representing a current drug discovery problem. The overall results herein obtained suggest that the identification of hits with appropriate trade-offs between potency and safety, rather than fully optimized hits solely based on potency, can facilitate the hit to lead transition and increase the likelihood of the candidate to evolve into a successful drug. So, it is possible to assert that the desirability-based MOOP method proposed seems to be a valuable tool for rational drug discovery and development.

Keywords: Computer-Aided Drug Design - Desirability Functions - Drug Discovery - Multi-Objective Optimization - Virtual Screening

RESUMO

A capacidade de melhorar o perfil farmacêutico de um fármaco baseado exclusivamente na sua eficácia terapcêutica ha sido freqüentemente superestimada. O ajuste de critérios múltiplos na identificação de candidatos potenciais (*hit-to-lead identification*) e na otimização dos líderes (*lead optimization*) é considerado um progresso fundamental no processo de descobrimento racional de fármacos. Assim, o desenvolvimento de aproximações capazes de manejar critérios adicionais para o tratamento prematuro e simultâneo das propriedades mais importantes que determinam o perfil farmacêutico de um candidato de fármaco como a sua potência, segurança, e biodisponibilidade, é uma questão emergente no processo de descobrimento e desenvolvimento de fármacos.Nesta Tese, é introduzido um método de otimização multi-objetivos (OMO) baseado nas funções de conveniência de Derringer, que permite conduzir estudos QSAR globais considerando simultaneamente a potência, segurança e/ou a biodisponibilidade de um conjunto de candidatos de fármaco. Os resultados do processo de OMO (os níveis das variáveis explicativas que simultaneamente produzem o melhor equilibrio possível entre as propriedades que determinam um ótimo candidato de fármaco) é usado para a implementação de um método de ordenação, também baseado na aplicação de funções de conveniência. Este método permite ordenar grandes bibliotecas de compostos (reais ou virtuais) com propriedades farmacêuticas desconhecidas de acordo com o grau de semelhança com o candidato ótimo previamente determinado.O processo inteiro é condensado em uma metodologia que nós decidimos nomear como **MOOP-DESIRE**, acrônimo em idioma inglês para **M**ulti-**O**bjective **OP**timization based on the **D**esirability **E**stimation of **S**everal **I**nterrelated **RE**sponses. A sua conveniência para as principais tarefas que envolvem o uso de métodos quimioinformáticos no descobrimento de fármacos - desenho de fármacos, ordenação de bibliotecas, e screening virtual - é avaliado além do uso da Teoria da Conveniência como uma ferramenta para a interpretação de modelos de predição multi-critérios. Cada tarefa foi avaliada mediante quatro conjuntos de dados diferentes permitindo a verificação do desempenho da metodologia na tarefa correspondente, representando cada uma de estas um problema atual na área de descobrimento de fármacos. Os resultados globais obtidos sugerem que a identificação de *hits* com um equilibrio apropriado entre potência e segurança, em lugar de *hits* completamente otimizados baseados unicamente na potência, pode facilitar a transição "*hit-to-lead*" e aumentar a probabilidade do candidato para evoluir num fármaco próspero. Assim, é possível afirmar que a metodologia de OMO

proposta pode ser considerada uma valiosa ferramenta para o processo de descobrimento e desenvolvimento racional de fármacos.

Palavras Chave: Descobrimento de fármacos -Desenho de fármacos assistido por computador - Funções de conveniencia - Otimização multi-objetivos - Screening virtual

TABLE OF CONTENTS

INDEX OF FIGURES

INDEX OF TABLES

LIST OF ABBREVIATIONS

The acronyms and symbols used in this thesis to define research fields, methods, molecular descriptors, etc. are listed below, in alphabetical order.

1D: unidimensional.
2D: bi-dimensional.
3D: tri-dimensional.
A_3AR: A_3 adenosine receptor.
Aa: anti-inflammatory activity.
ACF: atom centred fragments molecular descriptor
ALOGP2: square of the Ghose-Crippen octanol water coefficient.
An: analgesic activity.
ARR: fraction of aromatic atoms in the hydrogen suppressed molecule graph.
AUAC: area under the accumulation curve.
a(Q^2): Y-scrambling statistic based on the determination coefficient of the leave one out cross validation.
a(R^2): Y-scrambling statistic based on the determination coefficient.
B_D: joint belief based on desirability values.
C-001: atom centred fragment descriptor accounting for the number of methyl groups.
C-037: atom centred fragment descriptor accounting for the number of heteroatoms attached to a sp_2 carbon atom linked to the aromatic side ring.
CBR: case-based reasoning.
CoMFA: comparative molecular field analysis.
CoMSIA: comparative molecular similarity index analysis.
d: individual desirability.
D: overall desirability.
DF: desirability function.
DST: Dempster-Shafer theory, also known as belief theory.
DUD: directory of useful decoys.
EF: enrichment factor.
F: Fisher's statistics.
FP: false positive case.
GA: Genetic Algorithm.
H-046: atom centred fragment descriptor accounting for the number of hydrogen atoms attached to a sp_3 carbon no heteroatom attached to another carbon.
HIV-1: human immunodeficiency virus type-1.
HTS: high-throughput screening.
IC_{50}: concentration of compound yielding 50% cell survival compared to untreated control cells.
Ki_{A3}: binding affinities for the A_3 Adenosine Receptor.
LBVS: ligand-based virtual screening.
MD: molecular descriptor.
MIC: minimal inhibitor concentration.
MLR: multiple linear regression.
MOEA: multi-objective evolutionary algorithm.
MOOP: multi-objective optimization.
MOOP-DESIRE: multi-objective optimization technique based on the desirability estimation of several interrelated responses.
NCE: new chemical entity.
nCIR: number of circuits in the molecule graph.
nCs: number of total secondary sp3 carbon atoms.

NNRTI: non nucleoside reverse transcriptase inhibitor.
NSAID: non steroid analgesic/anti-inflammatory drug.
p: level of statistical significance.
PM: prediction model.
$\mathbf{Q^2}$: determination coefficient of the leave-one-out cross validation.
$\mathbf{Q^2_{Boots}}$: determination coefficient of the bootstrapping cross validation.
$\mathbf{Q^2_D}$: overall desirability's leave one out cross validation determination coefficient.
$\mathbf{Q^2_{LOO}}$: determination coefficient of the leave-one-out cross validation.
QSAR: quantitative structure-activity relationship.
QSBR: quantitative structure-biotransformation relationship.
QSPR: quantitative structure-property relationship.
QSTR: quantitative structure-toxicity relationship.
$\mathbf{R_{\%}}$: percentage of ranking quality.
R: correlation coefficient.
$\mathbf{R^2}$: determination coefficient.
$\mathbf{R^2_D}$: overall desirability's determination coefficient.
$\mathbf{RE_{A3}}$: relative maximal efficacy in the activation of the A_3AR.
ROC: receiver operating characteristic.
RT: reverse transcriptase enzyme.
s: fitting standard error.
$\mathbf{s_{Boots}}$: bootstrapping cross validation standard error.
$\mathbf{s_{LOO}}$: leave-one-out cross validation standard error.
$\mathbf{TP/FP_{ROC\text{-}OP}}$: operating point of the receiver operating characteristic curve.
TP: true positive case.
U: ulcerogenic index.
VS: virtual screening.
WSOF: weighted-sum-of-objective-functions.
$\mathbf{Y_a}$: yield of actives at certain filtered fraction.
$\mathbf{\Delta_i}$: parameter used to describe the similarity between a case i and the optimal case as a function of the subset of descriptive variables used for the multi-objective optimization process.
$^{\mathbf{D}}\mathbf{\Delta_i}$: desirability-normalized Δ_i.
$\boldsymbol{\rho}$: ratio between the number of compounds and the number of adjustable parametersin the model.
$\boldsymbol{\Psi}$: ranking quality index.
$\boldsymbol{\Psi^*}$: corrected ranking quality index.

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). Although “Costs of Goods” has been claimed as one of the major reasons for the end of a research & development (R&D) project (2) one cannot disregard the idea that toxicity and/or pharmacokinetics profiles of the clinical candidates are still decisive causes of failure in drug development process (3-6). Roughly 75% of the total costs during the development of a drug is attributed to poor pharmacokinetics or to toxicity (7).

In the 1980's, the development of high throughput technologies was expected to solve the drug discovery problem by a massive parallelization of the process. In practice, it turned out that, if they were not carefully deployed, these new technologies could lead to such a tremendous increase of candidate molecules that the drug discovery process became like finding a needle in a haystack. As a result, the large-scale approach has been progressively abandoned over the recent years, for the profit of more rationalized process. In this regard, Professor Hugo Kubinyi nicely pointed out: “*If you search a needle in a haystack, the best strategy might not be to increase the haystack*” (8-10).

The importance and possibility of jointly considering the multiple aspects of drug action was recognized and suggested since 1985 by Mayer and Van de Waterbeemd (11). As a possible way to achieve this goal, they suggest a stepwise multiple QSAR (MUQSAR) technique. In MUQSAR technique each step in drug action should be analyzed by using a quantitative method [i.e.: quantitative structure-activity/property/biotransformation/toxicity relationships (QSAR/QSPR/QSBR/QSTR)], thus permitting to fully conceive an “*overall QSAR*”: $OverallQSAR = f(QSAR_i, QSPR_i, QSBR_i, QSTR_i)$ (11).

Not without advising that some practical problems surely would have to be tackled, more than twenty years ago Mayer and Van de Waterbeemd were already confident about the feasibility of this approach and that the information finally obtained would worth the effort (11).

Improvement of the profile of a drug candidate 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 multi-objective 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 (12).

In fact, the ability to improve the pharmaceutical profile of candidates in lead optimization process on the sole basis of their activity has been often overestimated (3, 6). The adjustment of the multiple criteria in hit-to-lead identification and lead optimization is considered to be a major advance in the rational drug discovery process. The aim of this paradigm shift is the prompt identification and elimination of candidate molecules that are unlikely to survive later stages of discovery and development. In turn, this new approach will reduce clinical attrition, and as a consequence, the overall cost of the process (3, 13).

All these arguments put forward the need for approaches able to early integrate drug- or lead-likeness, toxicity and bioavailability criteria in the drug discovery phase as an emergent issue (3, 6). That is, methods that can handle additional criteria for the early simultaneous treatment of the most important properties, potency, safety, and bioavailability, determining the pharmaceutical profile of a drug candidate (14-22).

At the same time, the virtual screening (VS) (23, 24) 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 compounds (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 (25). VS technologies have thus emerged as a response to the pressure from the combinatorial/high-throughput screening (HTS) community.

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 (12). In particular, the QSAR paradigm has long been of interest in the drug design process (26), redirecting our thinking about structuring medicinal chemistry (27).

Yet standard chemoinformatics approaches usually ignore multiple objectives and optimize each biological property sequentially (11, 28-38). Nevertheless, some efforts have been made recently toward unified approaches capable of modeling multiple pharmacological, pharmacokinetic, or toxicological properties onto a single QSAR equation (39-43).

Multi-objective 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 (44).

Several applications of MOOP methods in the field of drug development have appeared lately, ranging from substructure mining to docking, including inverse QSPR and QSAR (44). Most of these MOOP applications have been based on the following approaches: weighted-sum-of-objective-functions (WSOF) (45) and pareto-based methods(44). An excellent review on the subject has been recently published by Nicolaou et al (44).

Concerning substructure mining, MOOP applications have focused on molecular alignment and pharmacophore identification. Examples of MOOPs tackling the substructure mining from a multi-objective perspective include the Genetic Algorithm Similarity Program method (GASP; a WSOF-based method) (46) and some pareto-based methods, such as the Genetic Algorithm for Multiple Molecular Alignment method (GAMMA; probably the first application of a pareto-based approach in chemoinformatics) (47) and the Genetic Algorithm with Linear Assignment for the Hypermolecular Alignment of Datasets (GALAHAD) (48).

As regards docking, several research groups are particularly active using pareto-based MOOP methods. For instance, Janson *et al.* (49) described a docking optimization application termed ClustMPSO, based on the particle swarm optimization (PSO) algorithm that minimizes simultaneously the intermolecular energy between the protein and the ligand and the intramolecular energy of the ligand. A multi-objective evolutionary algorithm (MOEA) has also been used by Zoete *et al.* (50) in their docking program EADock.

Recently, the application of the concept of multiple objectives have been introduced to the optimization of new chemical entities (NCEs) via *de novo* molecular design and inverse QSPR, standing out applications such as the CoG approach introduced by Brown *et al.* (51) to solve the inverse QSPR problem and the Molecule Evaluator proposed by Lameijer *et al.* (52) where the user assume the role of the fitness function by selecting candidate molecules for further evolution after each iteration.

Despite the availability of numerous optimization objectives, MOOP techniques have only recently been applied to the building of QSAR models. Nicolotti *et al.* (17) employed a variant of an evolutionary algorithm called multi-objective genetic

programming that used pareto ranking to optimize the QSAR models. A number of conflicting objectives including model accuracy, number of terms, internal complexity and interpretability of the descriptors used in the model were considered. On the other hand, Stockfisch (53) proposed a non-evolutionary multi-objective technique called the partially unified multiple property recursive partitioning (PUMP-RP) method for building QSAR models. This method was successfully used to construct models to analyze selectivity relationships between cyclooxygenase (COX) 1 and 2 inhibitors (54). More recently, a multi-objective optimization algorithm was proposed by Nicolotti*et.al.* for the automated integration of structure- and ligand-based molecular design (15). Actually, very few reports exist of the application of MOOP methods to QSAR, and even scarcer are the reports concerning the simultaneous optimization of competing objectives directly related with the definitive pharmaceutical profile of drugs, such as therapeutic efficacy, bioavailability, and/or toxicity.

Classic QSAR approaches usually ignore the multi-objective nature of the problem focusing on the evaluation of each single property as they became available during the drug discovery process (44). So, an approach offering a simultaneous study of several biological properties determinants for a specific therapeutic activity is considered a very attractive option in computational medicinal chemistry.

In this sense, desirability functions (DF) are well-known multi-criteria decision-making methods (55, 56). This approach has been extensively employed in several fields (57-68). However, despite of perfectly fit with the drug development problem, reports of computational medicinal chemistry applications are at present very limited (16, 18).

In the present work, we are proposing a MOOP methodology based on Derringer's desirability functions (56) that allows global QSAR studies to be run jointly, considering multiple properties of interest to the drug design process (16, 18). 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 (69). Specifically, in rational drug development, the availability of accurate ranking methods is highly desirable for virtual screening and filtering of promising new drug candidates from combinatorial libraries (7).

So, the results of the desirability-based MOOP will be used for the implementation of a ranking algorithm also based on the application of desirability functions. This desirability-based ranking algorithm it is proposed as multi-criteria virtual screening tool.

Summarizing, the knowledge involved in the development of new drugs is necessarily multidisciplinary. Like drugs, optimal QSAR models are a trade-off between several

objectives. At the same time, the process of computational drug discovery is conducted in many different ways and through diverse approaches, each with their own advantages and limitations. All these facts expose the multidimensional nature of the drug discovery and development process as well as an urgent need of methods able to integrate the plethora of approaches (mostly used as separate and independent pieces) and knowledge accumulated up to date, for the final and common goal: to develop efficient and safe drugs in a rational and cost-effective way. MOOP methods offer the potential to do this and the efforts involved in the present work attempted to approach to one of the countless routes to the complex goal of finding "good needles" on the vastness of that haystack that is the chemical space.

So, the specific objectives of this thesis can be summarized as follows:

i) To establish a multi-objective optimization & ranking methodology based on Derringer's desirability functions (MOOP-DESIRE Methodology),enabling global QSAR studies to be run jointly, considering multiple properties of interest to the drug discovery and development process.

ii) To evaluate the applicability of the MOOP-DESIRE methodology to the task of multi-criteria drug design by applying it to the design of novel NSAIDs quinazolinones with simultaneously improved analgesic, antiinflammatory, and ulcerogenic profiles.

iii) To evaluate the usefulness of the MOOP-DESIRE methodology as multi-criteria library ranking tool by applying it to the filtering of safe and potent antibacterial candidates from a heterogeneous library of antibacterial fluoroquinolones.

iv) To assess the potential of the MOOP-DESIRE methodology as multi-criteria virtual screening tool through the application of a MOOP-DESIRE-based prioritization of hits with appropriate trade-offs between human immunodeficiency virus type-1 (HIV-1) reverse transcriptase (RT) inhibitor efficacy and MT4 blood cells toxicity.

v) To evaluate the suitability of Desirability Theory as an interpretation tool for multi-criteria prediction models (PMs) by using it for the extraction of useful information on the desired trade-offs between binding and relative efficacy of N^6-substituted-4´-thioadenosines A_3 adenosine receptor (A_3AR) agonists.

2 RESULTS AND DISCUSSION

The results presented in this Thesis are reported by means of the author's original articles. First, the MOOP-DESIRE methodology is introduced and depicted in section 2.1. Next, the potential of the methodology proposed in the field of drug discovery are described by means of four practical applications in sections 2.2 to 2.5. Section 2.2 describes the potential of the methodology as a multi-criteria drug design tool. In section 2.3 is described their use as a multi-criteria library ranking algorithm. A multi-criteria virtual screening strategy is depicted in section 2.4 and finally, in section 2.5 is illustrated the use of Desirability Theory for the interpretation of a multi-criteria prediction model. Although the methodology itself involves several steps, only details pertaining to the respective applications are presented in these sections. The reader is referred to the author's original articles for more information.

The author's original articles (14, 16, 18, 22) have been attached under the heading "ANNEXES" of the present report. Pages containing explanatory sections follow the Thesis appropriate Arabic numeration system, whereas the pages belonging to the published works keep the actual journal numbering.

2.1 MOOP-DESIRE METHODOLOGY: MULTI-OBJECTIVE 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. 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 multi-objective 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, toxicological, and/or pharmacokinetic profiles of a set of drug candidates. The MOOP-DESIRE methodology is intended to find the most desirable solution that optimizes a multi-objective problem by using the Derringer's desirability function (70, 71), specifically addressed to confer rationality to the drug development process.

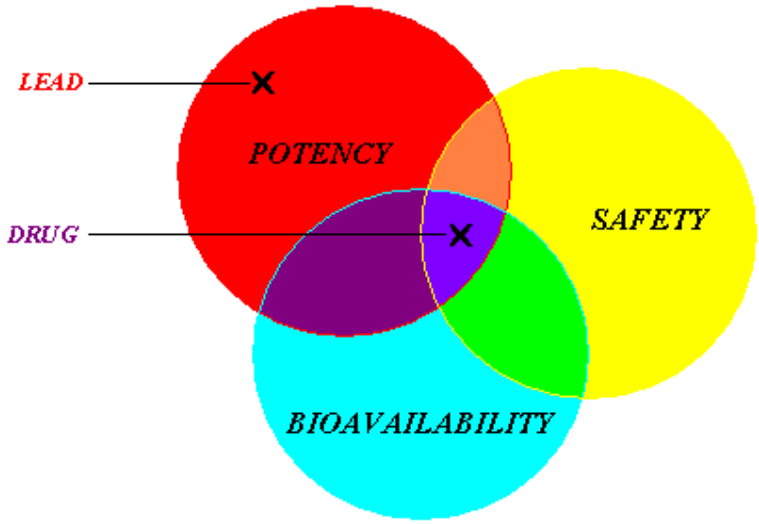


Figure 1. Graphic representation of the compromise between therapeutic efficacy (potency), bioavailability (ADME properties) and toxicity (safety) required to reach a successful drug.

Phase I: Desirability-based Multi-Objective Optimization

The process of simultaneous optimization of multiple properties of a drug candidate can be described as follows. 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 (MDs) used to model each property, respectively.

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 ($\hat{Y}_i$) 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 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 several 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. Desirability Function Selection and Evaluation

For each predicted response $\hat{Y}_i$, a desirability function d_i assigns values between 0 and 1 to the possible values of $\hat{Y}_i$. This transformed response d_i, can have many different shapes. Regardless of the shape, d_i= 0 represents a completely undesirable value of $\hat{Y}_i$, 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:*

$$D = (d_1 \times d_2 \times ... \times d_k)^{\frac{1}{k}} \qquad (\mathbf{1})$$

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, 1] 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 (56).

Let L_i, U_i, and T_i be the lower, upper, and target values, respectively, that are desired for the response $\hat{Y}_i$, 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:

$$d_i = \begin{cases} \left[\dfrac{\hat{Y}_i - L_i}{T_i - L_i}\right]^s & \text{if} \quad L_i \leq \hat{Y}_i \leq T_i \\ \left[\dfrac{\hat{Y}_i - U_i}{T_i - U_i}\right]^t & \text{if} \quad T_i < \hat{Y}_i \leq U_i \\ 0 & \text{if} \quad \hat{Y}_i < L_i \quad \text{or} \quad \hat{Y}_i > U_i \end{cases} \qquad (\mathbf{2})$$

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

$$d_i = \begin{cases} 0 & \text{if} \quad \hat{Y}_i \le L_i \\ \left[\dfrac{\hat{Y}_i - L_i}{T_i - L_i}\right]^s & \text{if} \quad L_i < \hat{Y}_i < T_i \\ 1 & \text{if} \quad \hat{Y}_i \ge T_i = U_i \end{cases} \qquad (\mathbf{3})$$

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:

$$d_i = \begin{cases} 1 & \text{if} \quad \hat{Y}_i \le T_i = L_i \\ \left[\dfrac{\hat{Y}_i - U_i}{T_i - U_i}\right]^s & \text{if} \quad U_i < \hat{Y}_i < T_i \\ 0 & \text{if} \quad \hat{Y}_i \ge U_i \end{cases} \qquad (\mathbf{4})$$

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 $\hat{Y}_i$ 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 $\hat{Y}_i$ above L_i and below U_i are acceptable or if having values of $\hat{Y}_i$ considerably above L_i are not of critical importance(56).

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 $\hat{Y}_i$, 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^2_D), 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 $\hat{Y}_i$ values (eqs2-4), and the overall desirability *D* is a continuous function of the *di* values s (eq. 1), then *D* is also a continuous function of the X_n. Therefore, R^2_D can be computed in analogy with the so-called determination coefficient R^2.Specifically, R^2_D is computed by using the observed D_{Yi} (calculated from Y_i) and the predicted $D_{\hat{Y}i}$ (calculated from $\hat{Y}_i$) overall desirability values instead of using directly the measured (Y_i) and predicted ($\hat{Y}_i$) response values.

$$R_D^2 = 1 - \frac{SSE}{SSTO} = 1 - \frac{\sum (D_{Y_i} - D_{\hat{Y}_i})^2}{\sum (D_{Y_i} - \overline{D}_{Y_i})^2} \quad (\mathbf{5})$$

where D_{Yi} and $D_{\hat{Y}i}$ have been defined previously. $\overline{D}_{Y_i}$ 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* ($Adj.R^2_D$) can be computed as shown below.

$$Adj.\ R_D^2 = 1 - \frac{SSE}{SSTO} = 1 - \frac{\dfrac{\sum (D_{Y_i} - D_{\hat{Y}_i})^2}{N-2}}{\dfrac{\sum (D_{Y_i} - \overline{D}_{Y_i})^2}{N-1}} \quad (\mathbf{6})$$

Like this, both R^2_D and $Adj.R^2_D$have the same properties of R^2 and $Adj.R^2$. Thus, both will fall in the range [0, 1],and the larger R^2_D/$Adj.R^2_D$ is, the lower is the uncertainty in predicting D by using a specific set of independent variables X_n (72).

Since $R^2_D$and $Adj.R^2_D$ 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 (LOO-CV) determination coefficient (Q^2_{LOO}) to establish the reliability of the method in predicting D. For this, the *overall desirability's LOO-CV determination coefficient* (Q^2_D) can be defined in a manner analogous to that of R^2_D.

$$Q_D^2 = 1 - \frac{SSE_{LOO-CV}}{SSTO} = 1 - \frac{\sum (D_{Y_i} - D_{\hat{Y}_i}(LOO-CV))^2}{\sum (D_{Y_i} - \overline{D}_{Y_i})^2} \quad (\mathbf{7})$$

where $SSE_{LOO\text{-}CV}$ and $D_{\hat{Y}i}$ (*LOO-CV*) are the leave one outcross validation square sum of residuals and the predicted overall desirability by LOO-CV, respectively.

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.

3. Multi-Objective 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 (73). The overall desirability D is optimized with the "*Use general function optimization*" option, which is, the *simplex* method of function optimization (74-76),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).

Phase II: 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) (77). 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 multi-objective optimization process previously described.

1. Similarity Assessment

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 multi-objective optimization process, which is defined as:

$$\Delta_i = \sum_{X=1}^{m} \delta_{i,X} \cdot w_X \qquad (8)$$

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:

$$E = \left[\sum (X_i - X_j)^2\right]^{1/2} \quad (\mathbf{9})$$

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:

$$\delta_{i,X} = \left|X_i - X_{OPT}\right| \quad (\mathbf{10})$$

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

2. Desirability Scaling of Similarity Metrics and Minimization of Differences Between Case Description (Δ_i) and Case Solution (D_i)

The Δ_i values are normalized by means of the application of the Derringer desirability functions(56) 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.4 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 (78, 79), implemented in the "*lsqcurvefit*" function of MATLAB program, version 7.2 (80).

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 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 ordered 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.

3. 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.

We will use some simple notations to represent ordering throughout this work. 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 OT = 1, 2, 3, ...,n. For any ordered list generated by a ranking algorithm, it is a permutation of OT. We use OR to denote the ordered list generated by the ranking algorithm R. OR can be written as a_1, a_2, ..., a_i, where a_i is the actual ordering position of the case that is ranked *ith* in OR (see Table 1).

Table 1. An 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

The ranking validation includes the following steps:

I. 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 (OT).

II. Invert OT. This new ordering corresponds to the worst order list (OW).

III. 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 ordering corresponds to the order generated by the ranking algorithm *R* (*OR*).

IV. Normalize (through eq.4) 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* (${}^{OR}d_i$).

V. Use the respective normalized order values to determine the difference between *OR* and *OT* (${}^{OT\text{-}OR}\delta_i$)

$$ {}^{OT-OR}\delta_i = \left| {}^{OT}d_i - {}^{OR}d_i \right| \qquad (\mathbf{11}) $$

and between *OW* and *OT* (${}^{OT\text{-}OW}\delta_i$)

$$ {}^{OT-OW}\delta_i = \left| {}^{OT}d_i - {}^{OW}d_i \right| \qquad (\mathbf{12}) $$

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

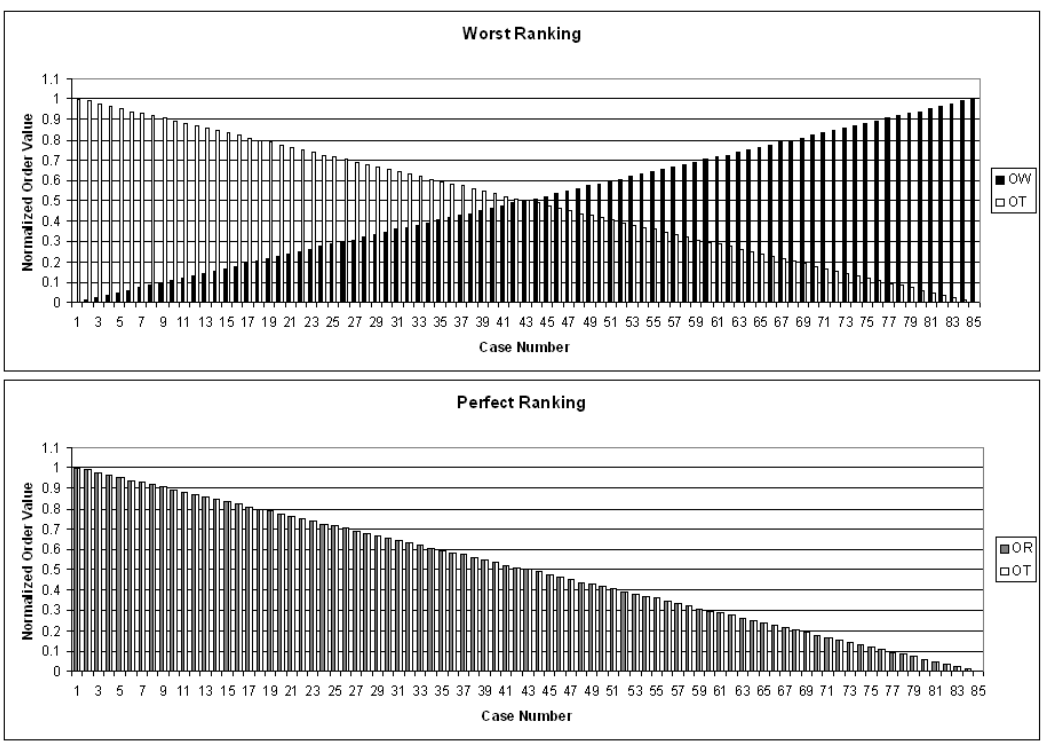


Figure 2. Worst (top) and perfect (bottom) ranking.

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

$$\Psi = \left| \frac{\sum_{i=1}^{n} {}^{OT-OR}\delta_i}{n} \right| \qquad (\mathbf{13})$$

Ψ is in the range [0, 0.5], being *Ψ*= 0 if a ranking is perfect and *Ψ* = 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*:

$$F = \frac{2}{\Psi^{OW}} \qquad (\mathbf{14})$$

where Ψ^{OW} is the quality index for the worst ranking. *F* is used here to obtain a more representative indicator 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^*):

$$\Psi^* = \left| \frac{\sum_{i=1}^{n} {}^{OT-OR}\delta_i}{n} \right| \cdot F = \left| \frac{\sum_{i=1}^{n} {}^{OT-OR}\delta_i}{n} \right| \cdot \frac{2}{\Psi^{OW}} \qquad (\mathbf{15})$$

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

$$R_{\%} = (1 - \Psi^*) \cdot 100 \qquad (\mathbf{16})$$

Finally, the Figure 3 summarizes schematically the above detailed MOOP-DESIRE methodology as a computer-aided tool for multi-criteria drug discovery.

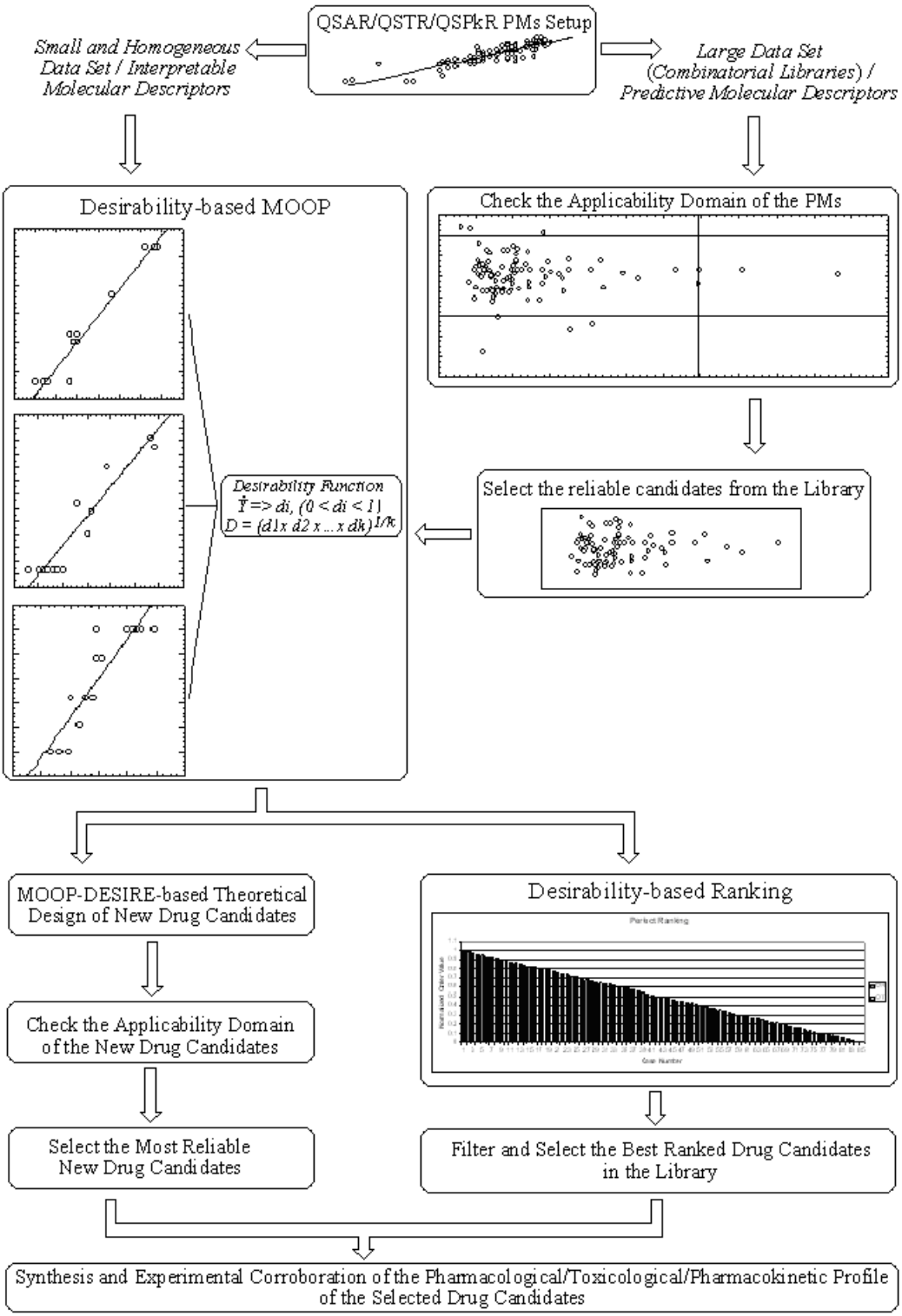


Figure 3.MOOP-DESIRE-based rational drug discovery and development.

2.1.1 Data Sets and QSAR Modeling Details

The respective data sets used in this work as well as the general aspects of QSAR modeling are depicted below. Details can be accessed from the respective author´s original articles (14, 16, 18, 22) which have been attached under the heading "ANNEXES" of the present report.

Data Sets

Design of Novel NSAIDs quinazolinones with Simultaneously Improved Analgesic, Antiinflammatory, and Ulcerogenic Profiles. A library of fifteen 3-(3-methylphenyl)-2-substituted amino-3*H*-quinazolin-4-one compounds published by Alagarsamy *et al.* (81) was used as starting point for the design of novel NSAIDS quinazolinones with simultaneously improved analgesic, antiinflammatory, and ulcerogenic profiles. See Annex I (16)for details.

Filtering Safe and Potent Antibacterial Candidates from a heterogeneous library of Antibacterial Fluoroquinolones. The multi-objective strategy for the filtering of safe and potent antibacterial candidates was based on a library of 117 fluoroquinolones published by Suto *et al*. reporting the cytotoxicity on Chinese hamster V79 cells expressed as the $IC_{50}$and the geometric mean of the minimal inhibitor concentration (MIC) for five Gram-negative bacteria (82). See Annex II (18)for details.

Prioritizing Hits with Appropriate Trade-Offs Between HIV-1 Reverse Transcriptase Inhibitory Efficacy and MT4 Blood Cells Toxicity. The prediction models for inhibitory efficacy over the HIV-1 RT and toxicity over MT4 blood cells, as well as the desirability-based MOOP and ranking process were performed using a library of non nucleoside reverse transcriptase inhibitors (NNRTIs) collected from previous literature reports(83-86).See Annex III (22) for details.

Extracting Useful Information on the Desired Trade-Offs Between Binding and Relative Efficacy of N_6-Substituted-4´-Thioadenosines A_3 Adenosine Receptor Agonists. The multiple linear regression (MLR) PMs developed were based on the binding affinities (Ki_{A3}) and relative maximal efficacy (RE_{A3}) in the activation of the A_3AR reported by Jeong *et al.* (87) for a library of thirty two N^6-substituted-4´-thioadenosines A_3AR agonists. See Annex IV (14) for details.

Molecular Structure Representation and Geometry Optimization

The structures of all compounds were first drawn with the aid of ChemDraw Ultra 9.0 software package (88), and reasonable starting geometries obtained by resorting to the MM2 molecular mechanics force field (89, 90). Molecular structures were then

fully optimized with the PM3 semi-empirical Hamiltonian (88), implemented in the MOPAC 6.0 program (91). Here, it should be remarked that the final molecular structures selected as prototype of the ¨bioactive¨ conformation pertain only to the compounds' global minimum energy conformations. We perfectly understand the limits of our selection criteria, but we can consider this a reasonable compromise to standardize the conformational selection.

Molecular Descriptors Calculation and Data Dimension Reduction

The 1664 MDs included in 20 different families implemented on software DRAGON 5.4 (92) were computed for each molecular structure previously optimized. The graphical user interface of DRAGON software is represented in Figure 4 allowing the inspection of the 20 families of MDs implemented. As a general rule, MDs having constant or near constant values as well as highly pair-correlated ($|R| > 0.95$) were automatically excluded in order to reduce the data dimension as well as noisy information that could lead to chance correlations.

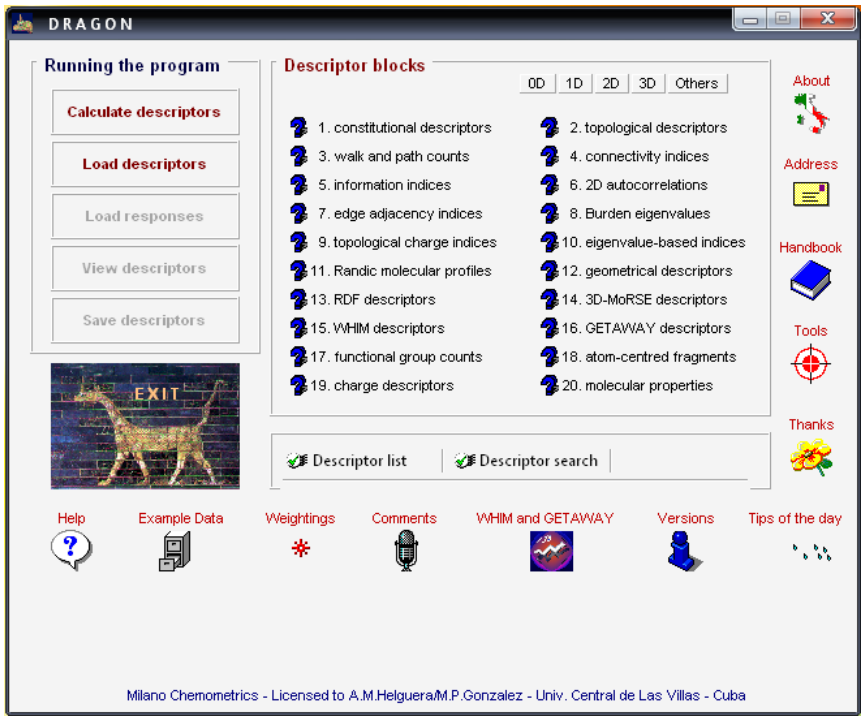


Figure 4. Graphical user interface of DRAGON software.

Selection of Relevant Molecular Descriptors

The task of selecting the descriptors that will be more suitable to model the activity of interest is complicated, as there are no absolute criteria for ruling such selection. Herein, an optimization technique – the *G*enetic *A*lgorithm (GA) – was applied for

variable selection (93-96). 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). The GA search was conducted in this work by using the software BuildQSAR (97, 98) as well as the MobyDigs 1.1 software package (99). The GA selection parameters to setup were: population size, maximum allowed variables in the model, and reproduction/mutation trade-off. The correlation coefficient (R), and the determination coefficient of the leave-one-out cross validation (Q^2_{LOO}) are the respective fitness functions employed in BuildQSAR (97, 98) and MobyDigs (99) GA variable selection.

Mapping the Molecular Descriptors to Activity

As to 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 analysis by means of the STATISTICA software package (73). 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.

Validation

The stability and predictive ability of the models was approached by means of both internal cross-validation and external validation methods. The leave-one-out (LOO) (71) and bootstrapping (100) techniques were the internal cross-validation methods employed. Basically, LOO consists of forming N subsets from the entire dataset, each missing one point, which in turn is used to validate a new model that is trained with the corresponding subset. The bootstrap validation procedure implemented on the software MobyDigs (99) was determined by 8000 re-substitutions. Additionally, a Y-scrambling procedure (101) (based on 500 random permutations of the Y-response vector) implemented on MobyDigs (99) was also applied to check whether the correlations established by the respective PMs were due to chance correlations or not. For details on the specific validation procedures applied to each particular work see the respective author´s papers (ANNEXES I-IV) (14, 16, 18, 22).

Parametrical Assumptions and Applicability Domain

We have also checked the validity of the pre-adopted parametric assumptions, another important aspect in the application of linear multivariate statistical-based

approaches (102). These include the linearity of the modeled property, homoscedasticity (or homogeneity of variance) as well as the normal distribution of the residuals and non-multicollinearity between the descriptors (103).

The applicability domain of the final PMs was identified by a leverage plot, that is to say, a plot of the standardized residuals *.vs.* leverages for each training compound (70, 71). The leverage (h_i) of a compound in the original variable space measures its influence on the model, and is calculated as follows:

$$h_i = \mathbf{t}_i (\mathbf{T}^T \mathbf{T})^{-1} \mathbf{t}_i^{\,T} \qquad (i = 1,\ldots,N) \tag{17}$$

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

$$h^* = 3 \times p'/N \tag{18}$$

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 three 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 were considered reliable.

Desirability Function Specifications

The optimization of the overall desirability was carried on by the *Use general function optimization* option (56) of the general regression module of STATISTICA (73). Furthermore, the spline method (104, 105) was used for fitting the desirability function and surface/contours maps, and the current level of each independent variable was set equal to its optimum value. As to the *s* and *t* parameters, these were fixed at 1.00 by assuming that the desirability functions increase linearly towards T_i on the three responses. For details on the desirability function specifications for each particular work see the respective author´s papers (ANNEXES I-IV).

2.2 DESIRABILITY-BASED MULTI-CRITERIA DRUG DESIGN

The evaluation of the capabilities of MOOP-DESIRE methodology to theoretically design new drug candidates with several biological properties simultaneously optimized was the main goal of this section. That is, not only to be able to translate the chemical structure into numbers to find out which are significantly related with a specific property, but in addition, to go back from these numbers to structure, or at least to some clues suggesting the structural modifications required to improve that property, or even better, more than one property at once. In doing so, we used as starting point a library of fifteen 3-(3-methylphenyl)-2-substituted amino-3*H*-quinazolin-4-one compounds reporting their respective analgesic (*An*) and anti-inflammatory (*Aa*) activities among the ulcerogenic index (*U*) (81).

2.2.1 Design of Novel NSAIDs quinazolinones with Simultaneously Improved Analgesic, Antiinflammatory, and Ulcerogenic Profiles

Following the strategy outlined previously, we began by seeking the best linear models relating each property to the atom centred fragments (ACF) molecular descriptors (106). One MLR-based PM containing two ACF variables previously selected by genetic algorithms was developed for each property (see Table 2).

Table 2. Regression coefficients and statistical parameters for the MLR models.

Analgesic Activity (*An*) Model								
$An = 51.762(\pm 2.155) + 8.333(\pm 0.957) \cdot C-001 - 6.929(\pm 1.534) \cdot C-037$								
N	***R***	***R^2***	***R^2 Adj.***	***Q^2***	***SPRESS***	***ρ***	***F***	***P***
15	0.967	0.935	0.923	0.905	3.143	5.000	85.15699	0.000000
Anti-Inflammatory Activity (*Aa*) Model								
$Aa = 36.708(\pm 1.789) + 5.527(\pm 1.232) \cdot C-001 + 1.475(\pm 0.430) \cdot H-046$								
N	***R***	***R^2***	***R^2 Adj.***	***Q^2***	***SPRESS***	***ρ***	***F***	***P***
15	0.942	0.887	0.869	0.827	3.526	5.000	47.46719	0.000002
Ulcerogenic Index (*U*) Model								
$U = 0.718(\pm 0.044) - 0.056(\pm 0.020) \cdot C-001 + 0.137(\pm 0.032) \cdot C-037$								
N	***R***	***R^2***	***R^2 Adj.***	***Q^2***	***s***	***ρ***	***F***	***P***
15	0.896	0.803	0.771	0.713	0.065	5.000	24.56766	0.000057

As can be noticed, the models are good in both statistical significance and predictive ability. Good overall quality of the models is revealed by the large *F* and small *p* values, satisfactory *ρ* values (*ρ* = 5), along with R^2 and *Adj.*R^2 (goodness of fit) values ranging from 0.803 to 0.935 and 0.771 to 0.923, respectively; as well asQ^2 (predictivity) values between 0.713 and 0.905. In addition, the overall desirability function exhibits good statistical quality as indicated by the R^2_D and *Adj.* R^2_D values (~1). Moreover, the high Q^2_D value (0.905) provides an adequate level of reliability on the method in predicting the overall desirability *D*.

By using these models as evaluation functions we may now thus proceed with an adequate level of confidence to the simultaneous optimization of the analgesic, anti-inflammatory and ulcerogenic properties for the set of compounds.

We intend to find a candidate with analgesic and anti-inflammatory activities as high as possible while keeping their ulcerogenic ability as low as possible. So, previous to the *simplex* optimization of the overall desirability *D*, the desirability function specifications were applied to each property accordingly (see Table 3). Here it is important to remark that, since *D* is maximized directly over the independent variables domain, and at the same time, the predicted *D* values depend on the initial set of PMs, one should consider the applicability domain of each PM to determine the optimum level of each independent variable as well as for the selection of the optimal solution(s).

Table 3. Desirability functions specifications.

Response	***OPT***	***DES***	L_i	U_i	T_i
An(%)	Max.	eq. 3	25	100	100
Aa(%)	Max.	eq. 3	25	100	100
U	Min.	eq. 4	0	1.73[a]	0

OPT: Type of optimization task; DES: Desirability function applied; L_i: Lower bound; U_i: Upper bound; T_i: Target; [a] Ulcerogenic Index of aspirin used as ulcerogenic reference drug.

Finally, the optimization of the overall desirability was carried out to obtain the levels of the ACF descriptors that simultaneously produce the most desirable combination of all properties. Figure 5 shows the multiple response overall desirability, as well as the individual desirability functions determined by the respective pairs of predictor variables included on the three MLR models.

By inspecting the form of each individual desirability function, it is possible to know the influence of a certain variable over each individual objective. In so doing, one can conclude that C-001 has a significant influence over the three properties, while H-046 has only a remarkable influence on the *Aa* activity. Here, one should note that the form of the *An* individual desirability function is similar to that obtained for the *Aa* activity (for these non competing objectives, both curves show a positive slope). However, opposite individual desirability function forms were obtained for competing objectives like *Aa* and *U* (*i.e.* the curve related to the ulcerogenic index has a negative slope).

Moreover, the data reveal that a 3-(3-methylphenyl)-2-substituted amino-3*H*-quinazolin-4-one optimized candidate must have analgesic and anti-inflammatory activities of 93.43% and 82.04%, respectively, plus an ulcerogenic index of 0.44. This represents an overall desirability of 0.8; that can be attained if the candidate has C-001, C-037 and H-046 values equal to 5, 0 and 12, respectively, being C-001 the

most influencing variable. The significant slope of the C-001 curve suggests that more attractive candidates could be designed if its values are greater than 5.

However, due to the high influence of C-001 over the overall desirability, the optimal range for this variable should be close to 5. But one must also consider the applicability domain of the original PMs. In fact, the training set show C-001 values up to 3 and thus, if the new candidate has a C-001 value extremely far from 3, it might be out of the applicability domain of the original PMs. On the other hand, as the shape of the H-046 desirability function reveals no significant influence (slope near zero), the overall desirability could be increased by large departures from its optimum value (= 12). But again the applicability domain of the original PMs should be taken into account.

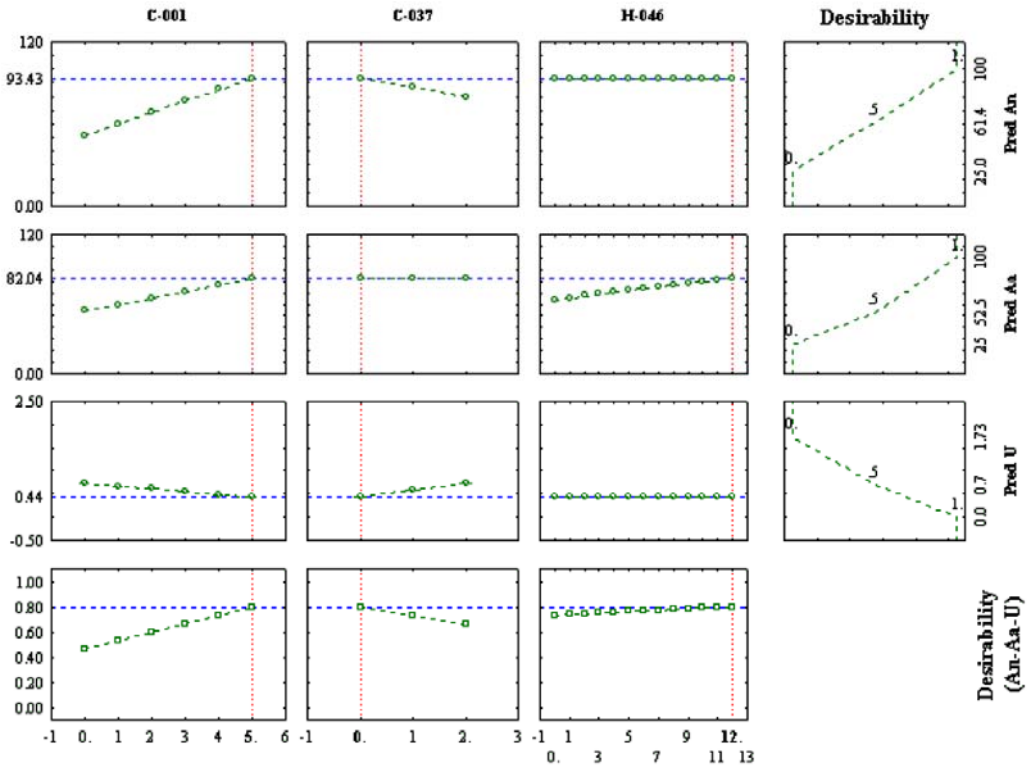


Figure 5. Multiple response desirability function due to the analgesic activity, anti-inflammatory activity and ulcerogenic index –*D*(*An-Aa-U*) (last row), along with the individual desirability functions coming from the pairs of predictor variables included on the three MLR models(first three rows).

According to the previous results, the most important variable was found to be descriptor C-001 and the second one descriptor C-037. These two ACF descriptors represent, respectively, the number of methyl groups and heteroatoms attached to a sp_2 carbon atom linked to the aromatic side ring in the drug candidates. On the other hand, the less influencing ACF descriptor, H-046, represents the number of hydrogen

atoms attached to a sp_3 carbon no heteroatom attached to another carbon. For a better understanding, this set of ACF molecular descriptors is depicted in Figure 6 for one on the training compounds (AS14).

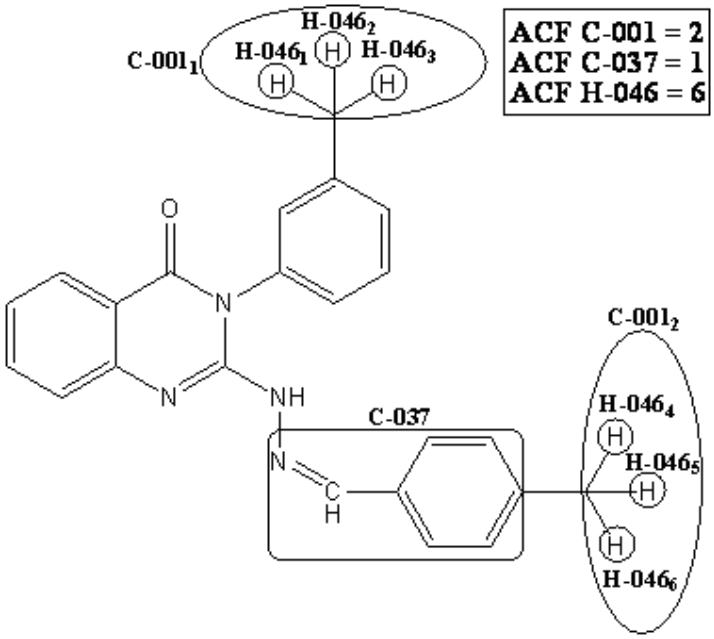


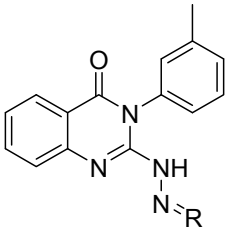
Figure 6. Atom-Centered Fragments (ACF) descriptors for compound **AS14**.

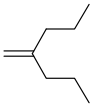
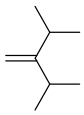
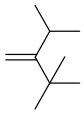
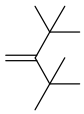
This information allows one guessing the main chemical modifications needed to improve the overall desirability of the present compounds. Considering the positive/negative influence of C-001/C-037 a different number *.vs.* type of alkyl groups on the C-2 position of the quinazoline ring should be introduced. In fact, the introduction of branched alkyl substituents might lead to a positive role due to the bulkiness of the substituents.

So, a new set of nine compounds was designed in which several different alkyl substituents were linked to the C-2 position of the quinazoline ring. The chemical modifications and the predicted values of the expected pharmaceutical properties are shown in Table 4. The leverage values obtained for each new designed candidate were also considered to check whether or not each new candidate falls within the applicability domain of the original PMs.

Table 4. Computed ACF descriptors (C-001, C-037 and H-046), predicted and leverage values for the analgesic (*An*) and anti-inflammatory (*Aa*) activities, plus the ulcerogenic index (*U*) of the nine new designed compounds.

3-(3-methylphenyl)-2-substituted amino-3H-quinazoline-4-one



Compound	*R*	*C-001*	*C-037*	*H-046*	An_{pred} (%)	Aa_{pred} (%)	U_{pred} (%)	*h*(*An*)	*h*(*Aa*)	*h*(*U*)
ASNEW1		3	0	11	77	70	0.55	0.216	0.361	0.216
ASNEW2		3	0	13	77	72	0.55	0.216	0.496	0.216
ASNEW3		4	0	12	85	77	0.49	0.403	0.453	0.403
ASNEW4*		5	0	15	93	86	0.44	0.573	**0.614**	0.573
ASNEW5*		6	0	18	102	96	0.38	**0.695**	**0.724**	**0.695**
ASNEW6*		7	0	21	110	106	0.33	**0.777**	**0.796**	**0.777**
ASNEW7		4	0	9	85	72	0.49	0.403	0.401	0.403
ASNEW8		5	0	12	93	82	0.44	0.573	0.562	0.573
ASNEW9*		5	0	15	93	86	0.44	0.573	**0.614**	0.573

* Compounds out of the predictions model's applicability domain; leverage values greater than *h** are marked in bold.

After an overall data analysis, compound **ASNEW8** can be claimed to be the most desirable and reliable candidate designed in this study, displaying predicted percentages of analgesic and anti-inflammatory activities of 93 and 82, respectively, plus a predicted ulcerogenic index of 0.44. Further, an excellent predicted overall desirability (0.8) is obtained.

A noticeable profile improvement can be observed between the predicted properties displayed by compound **ASNEW8** and the most promising compound reported by Alagarsamy*et al.* (**AS3**) (81). Explicitly, **ASNEW8** displays analgesic and anti-

inflammatory activities 15% and 13% higher, respectively. At the same time, **ASNEW8** shows only the 78.6% of the ulcerogenic ability of **AS3**. On the other hand, if we compare the performance of **ASNEW8** with diclofenac (a known NSAIDs used as reference compound (81)), one can easily notice its enhanced predicted pharmaceutical properties. In effect, **ASNEW8** displays analgesic and anti-inflammatory activities 31% and 22% higher than diclofenac, respectively. In addition, the ulcerogenic index is extensively reduced (**ASNEW8** has almost a quarter (3.75 times lower) of the ulcerogenic ability of diclofenac).

In summary, a remarkable simultaneously improvement on the analgesic and anti-inflammatory activities plus ulcerogenic profile of the new designed candidates was obtained throughout MOOP-DESIRE-based methods. The data suggest a positive role of the bulkiness of the alkyl substituents on the C-2 position of the quinazoline ring on the ulcerogenic properties. Anyhow, in the future, an experimental study of the analgesic, anti-inflammatory and ulcerogenic properties of the designed candidates should be carried out to validate the process.

Though the limited size and homogeneity of our data set, this work offers the possibility of a deeper and case by case analysis of the results obtained by using the MOOP-DESIRE methodology. The use of small and homogeneous data set is more suitable for later stages of the drug development process once identified a lead 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.

2.3 DESIRABILITY-BASED MULTI-CRITERIA LIBRARY RANKING

The MOOP-DESIRE methodology can also be applied to handle larger and/or more diverse data sets, such as those frequently obtained in *High-Throughput Screening* processes, being there more appropriate for early stages of the drug development process. That is, molecules coming from large and heterogeneous data sets can be ranked and filtered according to a certain criterion rather than applying the results of the optimization process to design new candidates. To accomplish that, one can resort to the overall desirability of each molecule as a ranking criterion or to several distance measures between the optimal values of the descriptors determined by MOOP-DESIRE and the computed values of the descriptors. In this case, it is advisable to use descriptors leading to highly predictive structure-desirability

relationships rather than interpretable descriptors in order to ensure the accuracy of the predictions and therefore, an accurate assessment of the molecule's overall desirability. So, in order to test the utility of the MOOP-DESIRE methodology as a multi-criteria library ranking algorithm it was applied to a library of 95 fluoroquinolones reported by Suto *et al*. (82). It was done with the aim of optimize simultaneously their antibacterial activity over gram-negative microorganisms (MIC) and their cytotoxic effects over mammalian cells (IC_{50}) and use these results as a pattern for a multi-criteria ranking algorithm.

2.3.1 Filtering Safe and Potent Antibacterial Candidates from a Heterogeneous Library of Antibacterial Fluoroquinolones

The desirability-based multi-objective optimization process was conducted in a similar manner to previous work. The best linear models relating each property to the DRAGON molecular descriptors are shown in Table 5 together with the statistical regression parameters. As can be noticed, the models are good in both statistical significance and predictive ability. In addition, the overall desirability function exhibits good statistical quality as indicated by the R^2_D and *Adj.* R^2_D values (~0.7). Moreover, a Q^2_D value of 0.63 provides an adequate level of reliability on the method in predicting *D*. So these models can be considered suitable as evaluation functions of the further simplex optimization process of the overall desirability *D*.

Table 5. Regression coefficients and statistical parameters for the MLR models.

Antibacterial Activity MLR Model (*MIC* = 1/1+*MIC*)

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

N	*R*	R^2	R^2 *Adj.*	*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})

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

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

Once the models has been set up and previous the optimization process of *D*, the desirability functions for each property (d_i's) might be specified. In order 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. **3**) and 1/1+IC_{50} minimized (eq.

4). 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 referred in Table 6.

Table 6. Desirability functions specifications.

Response	*Transformed Response*	*OPT*	*DES*	*(Response / Transformed Response)*		
				L_i	U_i	T_i
MIC (µg/mL)	**1/(1+MIC)**	Max.	eq. 3	25 µg/mL / 0.038	0.01 µg/mL / 0.99	0.01 µg/mL / 0.99
IC_{50} (µg/mL)	**1/(1+IC_{50})**	Min.	eq. 4	380 µg/mL / 0.002	8 µg/mL / 0.1	380 µg/mL / 0.002

OPT: Type of optimization task; DES: Desirability function applied; L_i: Lower bound; U_i: Upper bound; T_i: Target.

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 ($D_{MIC\text{-}IC50}$ = 1) fluoroquinolone-like candidate with the best possible compromise between antibacterial and cytotoxicity properties.

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

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. Through a nonlinear curve-fitting process implemented in MATLAB is 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.

Table 8. Optimal set of weighs.

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

Next, $\Delta_i$is used as a ranking criterion in order 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. 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.

Table 9. Δ_i, $^D\Delta_i$ and D_i values of the library of compounds used for ranking.

Compound ID	Δ_i	$^D\Delta_i$	Pred. $D_{(MIC-IC50)}$	Compound ID	Δ_i	$^D\Delta_i$	Pred. $D_{(MIC-IC50)}$
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				

Based on Δ_i 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 7.

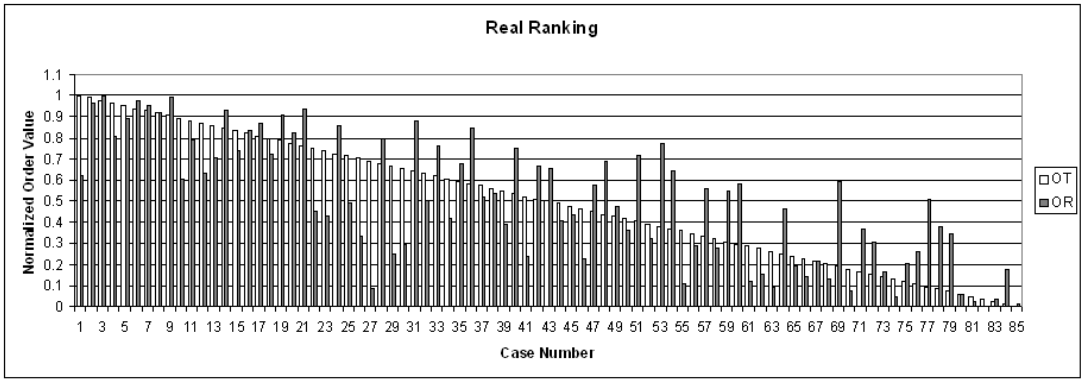


Figure 7.Δ_i-based ranking of the fluoroquinolone library.

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^2(MIC) = 0.693, $Q^2(IC_{50})$ = 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 of 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 in which it is based.

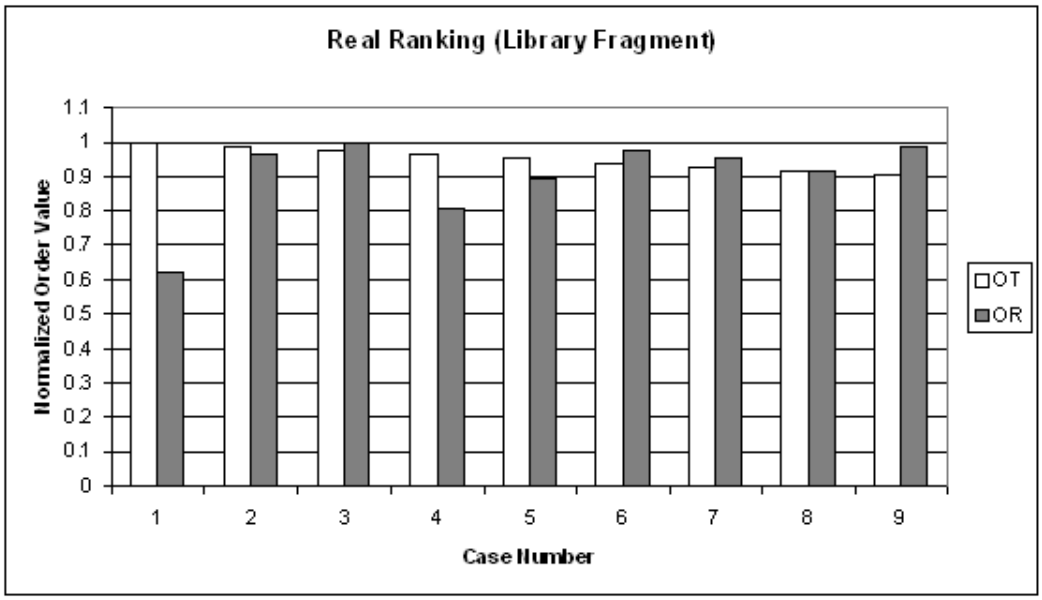


Figure 8. Ranking attained for the 10% of the library of compounds.

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 ones 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 exhibit a percentage of quality ranking of 82.74 ($\Psi^* = 0.173$). The ranking of this fragment is shown in Figure 8.

2.4 DESIRABILITY-BASED MULTI-CRITERIA VIRTUAL SCREENING

Filtering the most promising candidates having the best compromise between several properties comprising the final pharmaceutical profile confers to the process of discovery and development of new drugs an elevated degree of rationality which is difficult to reach via traditional QSARs which optimize sequentially each property. The sequential optimization of the properties comprising the final pharmaceutical profile of a drug candidate implies to overlook at some stage properties equally decisive to reach a successful drug or, at least, to find only by chance a candidate with acceptable profiles of all properties simultaneously. That is, a potent candidate once identified via QSAR has a high probability of being discarded later as a drug because of an unacceptable toxicological profile with the useless expenses of time and resources in synthesis and pharmacological assays (107). Equally difficult is the choice of using a panel of models (*i.e.*: a parallel screening based on QSAR models to respectively map the therapeutic efficacy and toxicity) since it is not very probable to find a candidate with all the properties simultaneously optimized and if this happens the results are more by chance than fruit of a rational drug development strategy.

In this regard, we describe in this section the application of the MOOP-DESSIRE methodology for simultaneously probe the inhibitory efficacy towards HIV-1 RT, and the toxic effects towards MT4 blood cells, of a diverse set of HIV-1 NNRTIs reported in the literature (83-86). This methodology is proposed as a rational strategy of multi-criteria virtual screening to prioritize HIV-1 NNRTIs hits with acceptable trade-offs between the above mentioned properties. Finally, a retrospective analysis of the training set, based on well-known enrichment measures (108-110), will be done allowing to compare the performance of several approaches (sequential, parallel and multi-objective) as VS strategies. The performance of this multi-criteria VS strategy to retrieve pharmaceutically

acceptable NNRTI candidates from a pool of NNRTI decoys is also tested. Since the capabilities of the methodology for multi-objective optimization and ranking has been well documented in the two previous sections, this section will be focused on the evaluation of the VS potential. The details regarding prediction models setup, multi-objective optimization and ranking can be assessed in the original publication(22) (ANNEX III).

2.4.1 Prioritizing Hits with Appropriate Trade-Offs Between HIV-1 Reverse Transcriptase Inhibitory Efficacy and MT4 Blood Cells Toxicity

The main goal in a VS effort is to select a subset from a large pool of compounds (typically a compound database or a virtual library) and try to maximize the number of known actives in this subset. That is, to select the most "enriched" subset as possible. Several enrichment metrics have been proposed in the literature to measure the enrichment ability of a VS protocol (108, 109). In this work, we use some of the most extended.

Based on the analysis of the receiver operating characteristic (ROC) curve (109) it is possible to derive the area under the ROC curve (*ROC Metric*) (108), as well as the ratio of true positive (TP) cases and false positive (FP) cases found at the operating point of the ROC curve ($TP/FP_{ROC\text{-}OP}$) (111).

From the accumulation curve we can deduce enrichment from the area under the curve (*AUAC*) (108), from the yield of actives (*Ya*) at certain filtered fractions (*i.e.*10%), as well as from the fraction of the database that has to be screened in order to retrieve a certain percentage (100%) of the TP cases (screening percentage, $\chi_{100\%}$).

On the other hand, the enrichment factor (*EF*) takes into account the improvement of the hit rate by a VS protocol compared to a random selection.

$$EF = \frac{TP/n}{N_+/N} \qquad (19)$$

where *TP* is the number of true positive cases retrieved, *n* the number of selected cases, *N* and N_+ are the total number of cases, and the number of positive cases in the library, respectively (108).

In a first experiment we are searching for the VS approach able to maximize the number of NNRTI candidates with a pharmaceutical profile equal or superior to 50% ($D_{IC50\text{-}CC50} \geq 0.5$) in a predefined fraction (χ) of the library (χ = 0.1 = top 10%; first 12 compounds). That is, to include in the top 10% fraction of the ordered library as much

candidates as possible exhibiting a favorable compromise between HIV-1 RT inhibitory efficacy and MT4 blood cells toxicity. The experiment is applied to the full set of 122 NNRTIs (83/21/18 from training/validation/test set) containing 41 compounds with a pharmaceutical profile equal or superior to 50%.

The sequential VS is conducted in this work by ranking independently the library of compounds according to the two objectives considered, HIV-1 RT inhibitory efficacy (IC_{50}) and MT4 blood cells toxicity (CC_{50}). The predicted values of IC_{50} and CC_{50} derived from the initial QSAR PMs were the ranking criteria employed. After ranking, a fraction of the library is first filtered according to a predefined threshold value of inhibitory efficacy (inhibitory efficacy profile ≥ 50%; $d_{IC50} \geq 0.5$; $-logIC_{50} \geq 0.196$; $IC_{50} \leq 0.64$ μM). Next, those candidates not fulfilling a predefined threshold value of safety (safety profile ≥ 50%; $d_{CC50} \geq 0.5$; $-logCC_{50} \leq -1.794$; $CC_{50} \geq 62.23$ μM) are eliminated in order to keep those with adequate inhibitory efficacy and safety profiles. In this approach; as well as in the multi-objective one, the true positive fraction (χ_+) can be equal or smaller than the filtered fraction χ (i.e., $0 \leq \chi_+ \leq \chi$).

The parallel VS, as the name implies, is based on running in parallel the independent analysis of the two objectives involved on the pharmaceutical profile of the candidate (IC_{50} and CC_{50}). The conditions in this case are identical to those defined for the sequential approach, but applied in a parallel fashion. In this case, those candidates included in each top 10% filtered fraction, and fulfilling the predefined threshold value for both criteria, are selected. In this case, if the retrieved compounds in both filtered fractions are the same, the retrieved fraction = χ = 0.1 = 12 compounds, otherwise the retrieved fraction ≤ 2χ. Consequently, $0 \leq \chi_+ \leq 2\chi$, depending of the efficacy and safety profiles of the candidates filtered in each top 10% filtered fraction.

The multi-objective VS approach proposed in this work considers the pharmaceutical profile of the candidate, rather than separately consider each property related with it. As detailed previously, the overall desirability of the candidate is the criterion employed here to measure their pharmaceutical profile. The library of NNRTIs is ranked according to a structural similarity criterion (Δ_i), top ranking those candidates structurally closer to the previously determined optimal candidate. Like in the sequential and parallel VS approaches, the top 10% of the ordered library is filtered, searching for those candidates with $D_{IC50\text{-}CC50}$ values ≥ 0.5.

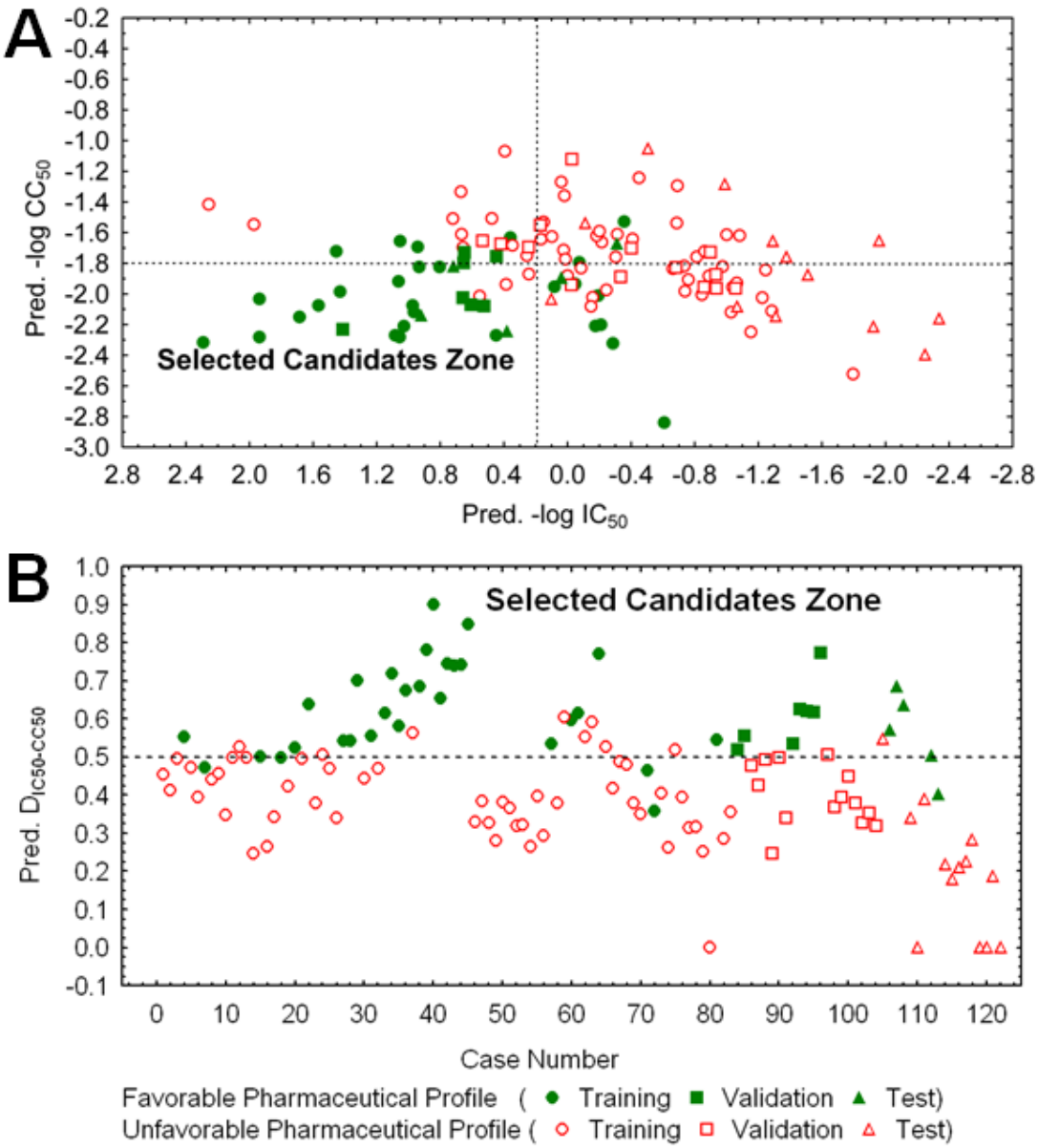


Figure 9. Graphical representation of the results for (A) a sequential screening [based on the inhibitory efficacy (*Pred.–logIC*$_{50}$) and safety (*Pred.–logCC*$_{50}$) profiles], and (B) a multi-objective screening [based on the pharmaceutical profile (*Pred.D*$_{IC50-CC50}$)], of the full set of 122 NNRTI compounds.

The suitability of a multi-objective VS approach can be checked if we compare the enrichment achieved in the screening of NNRTI candidates with a favorable pharmaceutical profile from the full set of 122 NNRTI compounds, sequentially considering the inhibitory efficacy (the predicted values of *–logIC*$_{50}$) and safety (the predicted values of *–logCC*$_{50}$) profiles in opposition to use the pharmaceutical profile information (*Pred.D*$_{IC50-CC50}$).

So, if the screening is conducted in a sequential manner, starting with the selection of candidates fulfilling a previously established threshold for the inhibitory efficacy ($Pred.{-}logIC_{50} \geq 0.196$; $Pred.IC_{50} \leq 0.64$ μM; $Pred.d_{IC50} \geq 0.5$) and further eliminating those candidates with an unfavorable safety profile ($Pred.{-}logCC_{50} \leq -1.794$; $Pred.CC_{50} \geq 62.23$ μM; $Pred.d_{CC50} \geq 0.5$), the area of selected candidates is reduced. As a consequence, 41% of the candidates (17 out of 41) with favorable pharmaceutical profiles ($D_{IC50\text{-}CC50} \geq 0.5$) are mistakenly discarded (see Figure 9A). However, by considering the compromise between inhibitory efficacy and safety of the candidates through a multi-objective virtual screening ($Pred.D_{IC50\text{-}CC50} \geq 0.5$) is possible to retrieve up to 88% of the candidates with acceptable pharmaceutical profiles included on the library (see Figure 9B).

This reveals the importance of considering multiple properties simultaneously since the sequential application of property filters could have led to the elimination of the candidate, despite it having a good balance between most of the properties (112). The importance of achieving a balance across a range of criteria is also recognized by other groups (113).

However, that can be settle on in a more detailed way by simulating a VS attempt over the same data set through three different VS approaches, and conducting a retrospective analysis of the performance of each approach by comparing the respective degree of enrichment achieved at the top 10% of the data set. As referred to above, the multi-objective VS approach proposed in this work is compared with two of the approaches – QSAR-based sequential and parallel VS – currently employed on drug discovery.

The sequential selection guides retrieving 75% of the pharmaceutically acceptable compounds included on the top 10% fraction of the data set, which represents an $EF_{10\%} = 2.232$. Similar but inferior results were achieved through a parallel screening ($Ya_{10\%} = 0.6$; $EF_{10\%} = 1.785$). These results although very good are outperformed when the selection of compounds was made based on a multi-objective criterion (the structural similarity to an optimal candidate,Δ_i). In the latter case, it was possible to retrieve 100% included on the same fraction of the data, reaching the maximum possible *EF* value for this fraction ($EF_{10\%} = 2.976$). More significant is the fact that compounds, initially selected, were rejected by the sequential or the parallel VS approach, even when they actually exhibited a pharmaceutically acceptable profile (false negative compounds, *FN*). Specifically, one out of twelve, and three out of twenty compounds were mistakenly discarded through the sequential and the parallel

approach, respectively. All these results are detailed in the original publication (22) (See Tables 9-11 in ANNEX III).

Finally, we decided to test the ability of the multi-objective VS strategy proposed to prioritize NNRTI candidates with favorable pharmaceutical profiles ($D_{IC50\text{-}CC50} \geq 0.5$) disperse in a data set of NNRTI decoys. NNRTI decoys are physically similar but chemically distinct from NNRTIs, so that they are unlikely to be binders of the HIV reverse transcriptase (RT). Specifically, we used as positive cases the 12 HIV RT known ligands with favorable pharmaceutical profiles included on the validation and test sets, and 36 decoys (negative cases) for each known ligand (432 decoys) were randomly selected from the database of HIV RT decoys included on the directory of useful decoys (DUD) (114).

We only considered those decoys included on the applicability domain of our prediction models at a ratio of 36 decoys per ligand, as recommended by Huang *et al.* (114). The final set of 444 compounds is ranked according to their structural similarity (Δ_i) with the previously determined optimal candidate, and the enrichment ability of this strategy is finally tested according to the enrichment metrics previously detailed and now depicted in Table 10.

Table 10. Enrichment metrics for Δ_i-based ranking of the data set collected form DUD.

ENRICHMENT METRICS	**MOOP Rank**
ROC Curve Information	
ROC Metric	0.798
$TP/FP_{ROC\text{-}OP}$	0.833/0.215
Accumulation Curve Information	
AUAC	0.828
$X_{100\%}$	0.320
$Ya_{10\%}$	0.333
Enrichment Curve Information	
$EF_{10\%}$	3.364
EF_{Max}	3.592

The respective values of *AUAC* and *ROC Metric* obtained suggest that the method is able to rank a NNRTI candidate with a favorable pharmaceutical profile earlier than a NNRTI decoy with a probability around 0.8. At the same time, $TP/FP_{ROC\text{-}OP}$ informs that, to obtain the best performance is necessary to filter 23.2 % of the library, in turn leading to find 83.3% of the TP cases at a cost of only 21.5 % of FP cases, which represents a EF_{MAX} = 3.592. Furthermore, all the positive cases can be found at the first 32% of the library. On the other hand, a third of the compounds retrieved, after filtering the top 10% of the library, were NNRTI candidates with a favorable pharmaceutical profile ($Ya_{10\%}$ = 0.33), which represents an $EF_{10\%}$ = 3.364, being 10.09 the maximum possible value of *EF* for this data fraction. The respective ROC, accumulation, and enrichment curves can be checked in Figure 10.

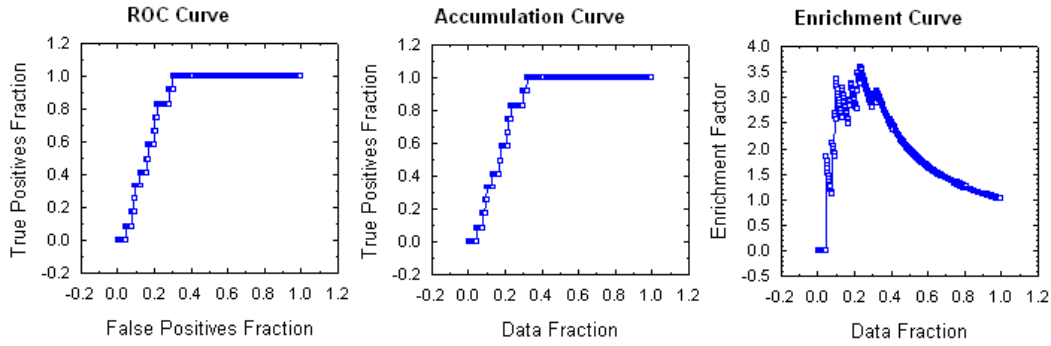


Figure 10. ROC, accumulation, and enrichment curves for the Δ_i-based ranking of the data set collected form DUD.

So, considering the previous results, one may well expect that larger (real or virtual) libraries of molecules (always inside the applicability domain of the PMs), like combinatorial libraries, could be correctly ranked; prioritizing in this way those candidates (top ranked) with more favorable compromise between inhibitory efficacy and safety.

2.5 DESIRABILITY-BASED INTERPRETATION OF MULTI-CRITERIA PREDICTION MODELS

Until now, have been exposed the multi-objective nature of the drug discovery process in which the modeling of decision preferences and constraints and the visualization and assessment of the trade-offs among objectives is yet a great challenge. As the desirability theory is a well-known multi-criteria decision-making approach it has decided to apply it, instead for multi-objective optimization, as a tool for the interpretation of multi-criteria prediction models. That is, instead of running a simultaneous optimization task over multiple properties of interest for drug discovery, such properties are directly combined into an overall desirability value (representing the compromise between the properties determining their pharmaceutical profile), predicted as a linear function of multiple molecular descriptors, and such a relationship is profiled in order to extract useful information on the desired trade-offs between such properties.

Specifically, we propose in this section the use of the desirability theory as a tool to extract useful information on the desired trade-offs between binding and relative efficacy of N^6-substituted-4´-thioadenosines A_3AR agonists. In doing so, we used the binding affinities (Ki_{A3}) and relative maximal efficacy (RE_{A3}) in the activation of the A_3AR reported by Jeong *et al.* (87) for a library of thirty two N^6-substituted-4´-thioadenosines A_3AR agonists.

2.5.1 Extracting Useful Information on the Desired Trade-Offs Between Binding and Relative Efficacy of N_6-Substituted-4´-Thioadenosines A_3 Adenosine Receptor Agonists

Once desirability scaled both Ki_{A3} and RE_{A3} responses for each compound, the corresponding overall desirability ($D_{KiA3\text{-}REA3}$) values were derived. In order to identify the factors governing the trade-offs between binding affinity and efficacy of this family of A_3AR agonists, the combined response $D_{KiA3\text{-}REA3}$ was mapped as a function of four simple 1D MDs with a direct structural and/or physiochemical explanation. The resulting best-fit model together with the statistical regression parameters is given in Table 11.

Table 11. Regression coefficients and statistical parameters for the overall desirability MLR model ($D_{KiA3\text{-}REA3}$).

$$D_{KiA3-REA3} = 1.557(\pm 0.292) - 0.107(\pm 0.013) \times \boldsymbol{ALOGP2} + 0.203(\pm 0.033) \times \boldsymbol{nCIR} - 2.783(\pm 0.595) \times \boldsymbol{ARR} - 0.092(\pm 0.027) \times \boldsymbol{nCs}$$

N	R^2	F	p	s	Q^2_{LOO}	s_{LOO}	Q^2_{Boost}	s_{Boost}	$a(R^2)$	$a(Q^2)$
32	0.781	24.13	< 0.01	0.127	0.566	0.138	0.539	0.179	0.0063	-0.0039

Based on the satisfactory accuracy, statistical significance, predictive ability and fulfilment of the pre-adopted MLR parametrical assumptions of the overall desirability PM ($D_{KiA3\text{-}REA3}$ model) we can proceed, with an adequate level of confidence to the simultaneous analysis of the factors governing the balance between the binding affinity and relative efficacy profiles of A_3AR agonists.

Although the main variation of the subset of compounds employed is over the N^6 position of the adenine ring, the MDs employed in mapping $D_{KiA3\text{-}REA3}$ are global and not fragment based. So, any inference made have to be only based on the influence of N^6 substituents over the global molecular system.

First, the information encoded in the MDs included on the model was analyzed. According to the model regression parameters, the most influencing MD is the aromatic ratio (*ARR*), followed by the Ghose-Crippen octanol water partition coefficient (*ALOGP2*), the number of circuits (*nCIR*) and the number of total secondary sp3 carbon atoms (*nCs*). All MDs were inversely related with the overall desirability $D_{KiA3\text{-}REA3}$ of N^6-substituted-4´-thioadenosine A3AR agonists, except *nCIR*.

Specifically, *ARR* is the fraction of aromatic atoms in the hydrogen suppressed molecule graph and encodes the degree of aromaticity of the molecule. According to the model parameters, N^6 substitutions increasing the aromaticity of the molecule do not favor $D_{KiA3\text{-}REA3}$.*ALOGP2* is simply the square of the Ghose-Crippen octanol water coefficient (*ALOGP*). Since these MD encodes the hydrophobic/hydrophilic character of the molecule, $D_{KiA3\text{-}REA3}$ could be favored by the presence of N^6 substituents

contributing to reduce the hydrofobicity of the molecule. The *nCIR* is a complexity descriptor, which is related to the molecular flexibility. Since *nCIR* serve as a measure of rigidity with higher numbers of circuits corresponding to reduced flexibility; cyclic and rigid or conformationally restricted N^6 substituents could increase the overall desirability of the molecular system. Finally, the presence of secondary sp^3 carbon atoms in the molecule appears to be detrimental for $D_{KiA3\text{-}REA3}$.

According to the model, a molecule with a low aromaticity degree, without secondary sp^3 carbon atoms, and containing cyclic and rigid N^6 substituents which contributes to reduce the hydrofobicity of the system could favor the balance of the binding affinity and relative efficacy profiles of N^6-substituted-4´-thioadenosine A_3AR agonists.

To note that these conclusions, although derived from a simple 1D model, are very similar to that obtained by 3D-CoMFA/CoMSIA approaches (115). Kim and Jacobson have concluded that a bulky group, conformationally restricted, at the N^6 position of the adenine ring will increases the A_3AR binding affinity, and that a small bulky group, at this position, might be crucial for A_3AR activation. Note the accordance of data obtained in the previous and present work: a “conformationally restricted bulky group” is suggested by Kim and Jacobson and herein a “cyclic and rigid substituents” on the N^6 position.

To note that although *nCIR* is not the MD more significantly related with $D_{KiA3\text{-}REA3}$, it is very informative for the property. From *nCIR* we can infer that the bulkiness of the N_6 substituent suggested in (115) can be characterized by a cyclic rather than an alkyl substituent.

Although useful, this information is found to be incomplete since it is well-known that steric factors are determinant for the design of A_3AR agonists, especially for binding affinity (115). Consequently, it is found to be important to determine the optimal size of the conformationally restricted cyclic N_6 substituent. Unfortunately, the simple inspection of the regression parameters of the PM do not offers this information. In consequence, a property/desirability profiling was carried out to identify the levels of the MDs included in the PM that simultaneously generate the most desirable combination of binding affinity and relative efficacy.

As the main goal of this analysis is to extract information on the factors governing $D_{KiA3\text{-}REA3}$ rather than optimize it, the behaviour of $D_{KiA3\text{-}REA3}$ was profiled at the mean values of the four MDs rather than looking for their optimal values (see first row in Figure 11). Accordingly, it was possible to find the levels of the MDs simultaneously producing the best possible $D_{KiA3\text{-}REA3}$ in the training set employed.

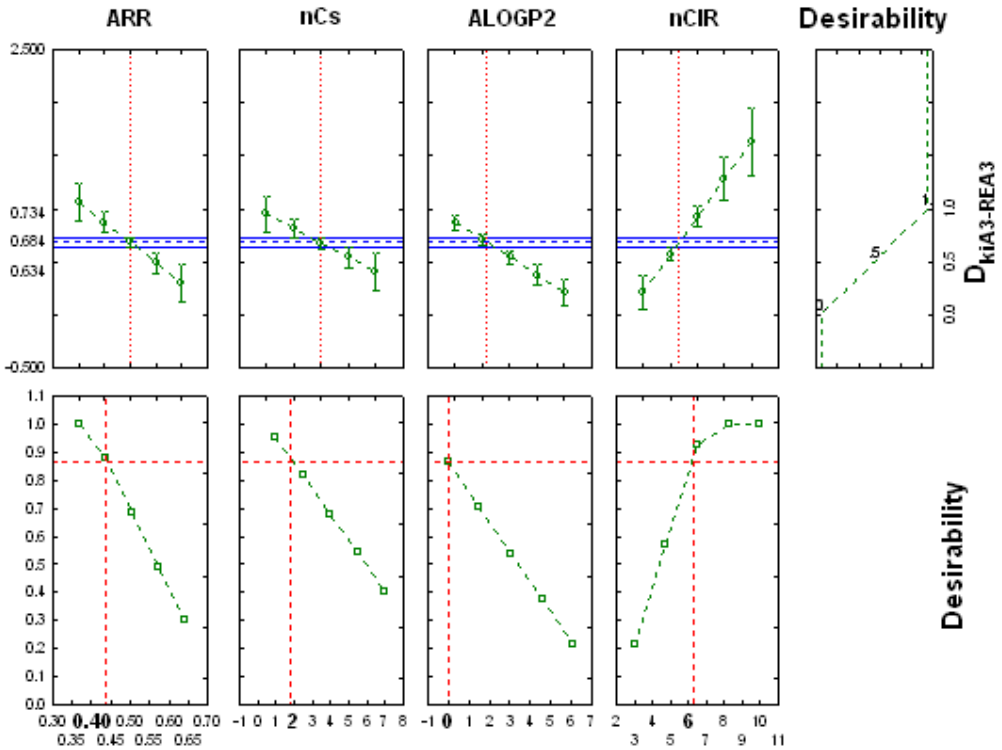


Figure 11. Property/desirability profiling of the levels of the MDs that simultaneously produce the most desirable combination of binding affinity and relative efficacy of N^6-substituted-4´-thioadenosine A_3AR agonists.

The analysis reveal that for the most favorable balance of binding affinity and agonist efficacy: the *ARR* should be not just low but near to 0.4; *ALOGP2* should be as low as possible; the number of secondary sp3 carbon atoms should be kept around two; and *nCIR* should be not just high but close to six.

Since the thioadenosine nucleus already contain three secondary sp3 carbon atoms, at least on the applicability domain of the present model, the minimum number of such atoms should be kept at three. So, this type of carbons must be excluded in the substituents located at N^6 position.

At the same time, considering that the *nCIR* value of the thioadenosine nucleus is four, one can deduce that the ideal *nCIR* value of the N^6 substituent should be two. This information can be structurally translated into bicyclic N^6 type of substituents.

The inclusion in the PM of *nCIR*, instead of the number of rings in the chemical graph (*nCIC*) is also significant. Although the structural information of this pair of MDs is very similar (the number of cyclic structures in a chemical graph) their graph-theoretical information is quite different. While *nCIC* encodes the number of rings, *nCIR* includes both rings and circuits (a circuit is a larger loop around two or more rings). So, additional information can be inferred: the bicyclic N^6 substituent should not be fused. This assumption could be related to the binding interaction of this type

of fragments with the A_3AR. In fact, the presence of a certain degree of rotational freedom between the two rings of the fragment could favor its docking into the receptor cavity.

This result matches with previous experimental findings on the structure-activity relationship (SAR) of this family of thioadenosine derivatives (87). The SAR obtained for this family suggests that compounds with bulky N^6 substituents lost their binding to the A_3AR. Paradoxically, among compounds showing high binding affinity at the human A_3AR, two compounds substituted with a N^6-(*trans*-2-phenylcyclopropyl) amino group were found to be full agonists at the human A_3AR. In addition, it was found that compounds with α-naphthylmethyl N^6 substituents lost their binding to the A_3AR (87), which reinforce the present proposal.

From the study it was also concluded that bulky N^6 substituents only affects the binding affinity, however bulky (bicyclic) substituents such as a *trans*-2-phenylcyclopropyl group, could be beneficial for agonist efficacy without lost their binding affinity. Although that experimental study does not deal with the simultaneous analysis of both properties, their experimental findings properly match with our theoretical results.

The previous information can be employed for the theoretical design of new N^6-substituted-4´-thioadenosine analogues with adequate balances between binding affinity and agonist efficacy. Since *ARR* and *ALOGP2* cannot be easily manipulated by structural modifications, the design efforts will be mainly focused on *nCs* and *nCIR*. Thus, a combinatorial library focused on the generation of N^6-substituted-4´-thioadenosine candidates was assembled with *nCs* ≈ 3 and *nCIR* ≈ 6. This approach was performed with the aid of the SmiLib software (116), for the rapid assembly of combinatorial Libraries in SMILES notation. The library was directed to produce candidates with conformationally restricted bicyclic N^6 substituents while keeping at minimum the presence of secondary sp^3 carbon atoms using the 4´-thioadenosine nucleus as scaffold and a set of 25 cyclic or heterocyclic structures as linkers and building blocks. The working combinatorial scheme is shown in Table 12.

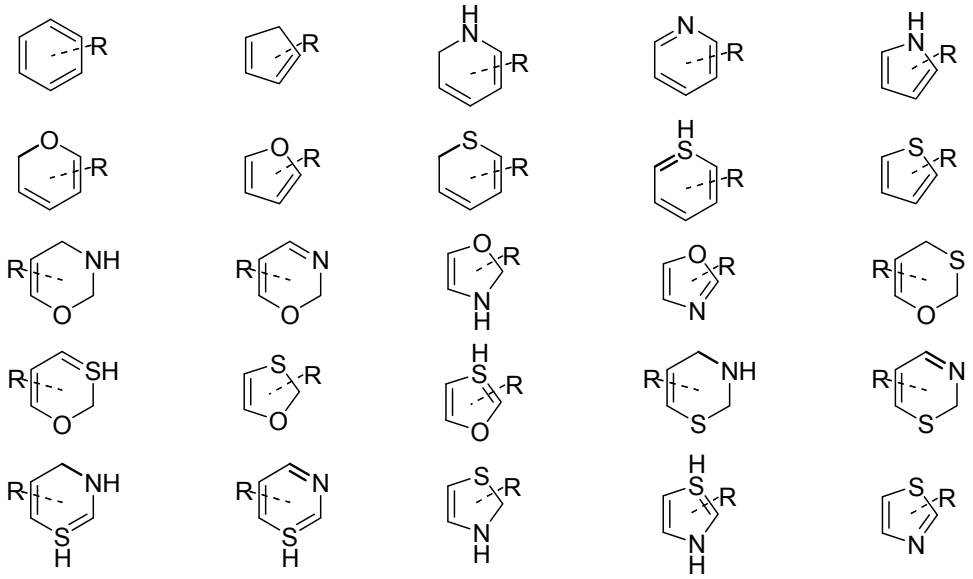
This combinatorial strategy produced a focused combinatorial library of more than 9 000 candidates which according to previous results, can be employed in a subsequent virtual screening campaign using as ranking criterion the predicted value of $D_{KiA3\text{-}REA3}$ of each candidate. As mentioned before, only candidates included on the applicability domain of the overall desirability PM (3395 candidate molecules) should be submitted to the ranking process. As a result, it is possible to propose for

biological screening a reduced set of candidates with a promissory balance between A_3AR binding affinity and agonist efficacy.

Table 12. Scaffolds, linkers and building blocks employed to assemble the combinatorial library

Scaffolds

Linkers/Building Blocks



2.5.2 Multi-Criteria Virtual Screening based on the Combined Use of Desirability and Belief Theories

Although the idea of desirability-transforming and combining a number of related properties is in accordance with the concept of pharmaceutical profile (16, 18), the usefulness of a parallel approach allowing obtaining a feedback on the reliability of the properties predicted as a unique overall desirability D_i value, is also desirable.

If two or more property values Y_i (previously scaled to the respective d_i values with proper desirability functions) of a compound are combined into a unique D_i value, in order to map it as a MLR function of n molecular descriptors X_i (denoted as approach A_1), it is rational to expect that the resultant predicted D_i value should be similar to the inverse approach. The inverse approach consist in the independent mapping of the k properties Y_i as a MLR function of n molecular descriptors X_i, the subsequent

desirability-scaling of each predicted Y_i value and the final combination of the corresponding d_i values into a unique predicted D_i value (denoted as approach A_2).

$$Y_i \rightarrow d_i \rightarrow D_i = f(X_i) \rightarrow \mathrm{Pr}ed.D_i = A_1 \approx A_2 = \mathrm{Pr}ed.D_i \leftarrow \mathrm{Pr}ed.d_i \leftarrow \mathrm{Pr}ed.Y_i \leftarrow Y_i = f(X_i) \quad (\mathbf{20})$$

Assuming true the previous analysis, one must anticipate that the higher is the degree of similarity between the predicted D_i values of both approaches, the higher should be their reliability, and vice versa. Clearly, the results will depend of the goodness of fit and prediction of the set of PMs involved. In addition, the degree of uncertainty of PMs with different sets of MDs will be diverse.

So, it is required a framework allowing the fusion of results from different approaches in order to access the reliability of predictions from several approaches with different degrees of uncertainty. In the present work we select Dempster-Shafer Theory (DST) (117-119) (also known as belief theory) to achieve that goal (120). DST is based on two ideas: the idea of obtaining degrees of belief for one question from subjective probabilities for a related question, and Dempster's rule for combining such degrees of belief when they are based on independent items of evidence (120).

These two rules are quite simple. The *rule for successive testimony* says that if a report has been relayed to us through a chain of n reporters, each having a degree of credibility p, then the credibility of the report is p^n. The *rule for concurrent testimony* says that if a report is concurrently attested to by n reporters, each with credibility p, then the credibility of the report is $1-(1-p)^n$; where $0 \leq p \leq 1$. Thus, the credibility of a report is weakened by transmission through a chain of reporters but strengthened by the concurrence of reporters (118, 119).

If we make a simple analogy of this situation with the situation previously exposed regarding two parallel overall desirability PMs, each approached inversely, is possible to note that DST theory, specifically, the Hospers's rule for combining concurrent evidence (118, 119), is fully applicable to our problem. There it is only needed to replace "report" with "prediction" and "reporter" with "prediction model", and the previous paragraph will almost literally describe our problem.

Developing a *probability assignment* is the basic function in DST, and is an expression of the level of confidence that can be ascribed to a particular measurement. However, in this work we are interested on the desirability of a compound. Consequently, rather than a probability assignment for each compound, we will use the desirability values coming from both overall desirability PMs approaches (D_1 and D_2) to derive the final joint belief values (B_D):

$$B_D = 1 - (1 - D_1)(1 - D_2) \quad (\mathbf{21})$$

While desirability is not itself a probability, like probabilities their values also range from 0 to 1. Therefore it can be used to derive the values of B_D for each compound. So, in this way it is possible to encode the reliability of the predicted desirability of a compound along with to two inverse but complementary prediction approaches. Given this information, B_D can be used as ranking criterion in a virtual screening scheme, resulting particularly useful for ligand-based virtual screening (LBVS).

A LBVS strategy based on B_D can be described in the sequence of steps detailed bellow:

1- *Prediction Models setup*.

Here, the predicted D_i values for each compound are derived from A_1 and A_2 as expressed in eq. **20**.

2- *Desirability assignment*.

Due to limitations inherent to the MLR approach, the predicted desirability values not always will be included in the interval [0,1] and consequently is not possible to use it as is to derivate B_D. So, in the case of the desirability values derived from the approach A_1, it is necessary to re-scale using eq.**3** considering that D have to be maximized.

In the case of the approach A_2, the derivation of the respective D_i values is affected by the above mentioned limitations of MLR, but the process is complicated by the wider range of the mapped Y_i properties. Consequently, d_i is scaled by using a two-tale (eq.**2**) using the same target T_i values employed in A_1 for each Y_i.

3- *Derivation of Joint Belief B_D by the application of Hospers's Rule for Combining Concurrent Evidence*.

4- *B_D-Based Ranking*.

The resultant ranking should render an ordered list, top-ranking the most reliable compounds with the highest desirability values. The compounds with a higher chance to exhibit a desirable combination of the k properties modeled.

Two QSAR PMs (for Ki_{A3} and RE_{A3}) focused on their predictive ability (prediction approach A_2) were derived in order to use both in combination with the previously described overall desirability PM ($D_{KiA3\text{-}REA3}$ model, identified as prediction approach A_1) in a LBVS strategy based on the combination of their concurrent predictions through belief theory.

The resulting best-fit models together with the statistical regression parameters are depicted in Table 13. According to their statistics, the models are good in terms of their statistical significance and predictive ability.

Table 13. Regression coefficients and statistical parameters for the MLR models involved on the prediction approach A_2 (Ki_{A3and} RE_{A3}).

Ki_{A3} MLR Model

$$Ki_{A3} = -8857.67(\pm 331.482) + 10.36(\pm 1.019)\cdot \boldsymbol{D/Dr03} + 502.99(\pm 99.263)\cdot \boldsymbol{GATS3m}$$
$$+ 5217.43(\pm 188.103)\cdot \boldsymbol{BELe3} - 453.64(\pm 45.869)\cdot \boldsymbol{Mor13u} + 1110.88(\pm 57.144)\cdot \boldsymbol{Mor09v}$$
$$- 1258.23(\pm 101.691)\cdot \boldsymbol{Mor23v} + 26703.72(\pm 3542.089)\cdot \boldsymbol{R7u} +$$

N	R^2	F	p	s	Q^2_{LOO}	s_{LOO}	Q^2_{Boots}	s_{Boots}	a(R^2)	a(Q^2)
32	0.985	230.82	< 0.01	48.80	0.977	56.35	0.957	61.25	0.0017	-0.0052

RE_{A3} MLR Model

$$RE_{A3} = 2559(\pm 413.56) - 3307(\pm 373.0.4)\cdot \boldsymbol{PW2} - 0.44(\pm 0.038)\cdot \boldsymbol{D/Dr06}$$
$$- 143.68(\pm 28.85)\cdot \boldsymbol{ATS5v} + 344.25(\pm 25.72)\cdot \boldsymbol{EEig10d} + 114.72(\pm 10.54)\cdot \boldsymbol{VEA1}$$
$$+ 89.91(\pm 20.18)\cdot \boldsymbol{H8p} - 15.68(\pm 2.32)\cdot \boldsymbol{ALOGP}$$

N	R^2	F	p	s	Q^2_{LOO}	s_{LOO}	Q^2_{Boots}	s_{Boots}	a(R^2)	a(Q^2)
32	0.966	96.79	< 0.01	5.52	0. 942	6.37	0.921	7.18	0.0017	-0.0055

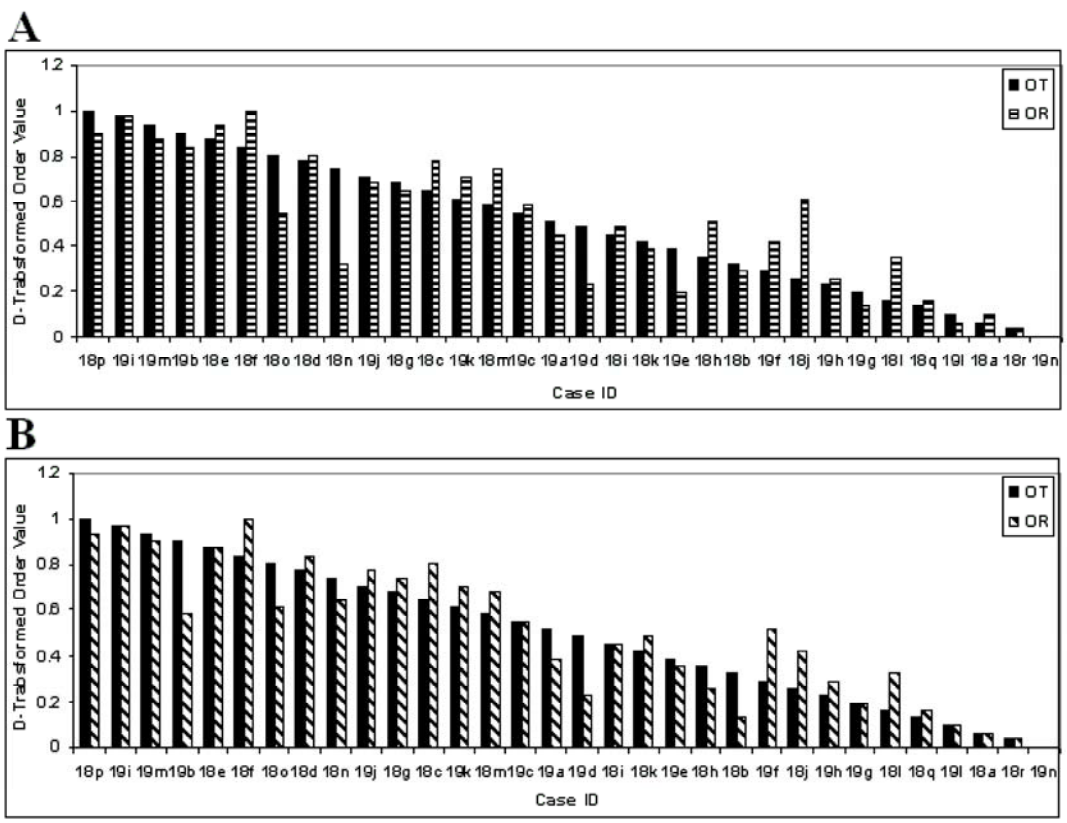


Figure 12. Ranking of the training set compounds based on $D_{KiA3\text{-}REA3}$ (A) and B_D (B), respectively.

Considering the structural similarity between both (the combinatorial library assembled and our training set), is possible to use the latter to infer the reliability of

the ranking attained for the combinatorial library. The predicted values of $D_{KiA3\text{-}REA3}$ (according to approach A_1) was also tested as ranking criterion in order to compare a VS strategy based on predictions coming from a single approach with a VS strategy based on the combination of concurrent predictions. The quality of the respective ranking obtained was compared according to the corrected ranking quality index Ψ*, previously defined in section 2.1.

Based on the analysis of our training set, the quality of the ranking attained using the predicted values of $D_{KiA3\text{-}REA3}$ is around 80% which suggest an acceptable degree of confidence if the scheme is applied to our combinatorial library (R% = 80.08%; Ψ* = 0.1992). As can be noted in Figure 12, the use of B_D as ranking criterion (R% = 82.81%; Ψ* = 0.1719) slightly overcomes the performance of the predicted values of $D_{KiA3\text{-}REA3}$. Considering that B_D encodes in addition to the desirability of the compound, the reliability of such a prediction, it is clear their suitability at the moment to screen higher and/or structurally diverse libraries with a wider range of the mapped properties.

3 CONCLUSIONS

i) A new multi-objective optimization & ranking methodology based on Derringer's desirability functions (MOOP-DESIRE Methodology), enabling global QSAR studies to be run jointly, considering multiple properties of interest to the drug discovery and development process, was introduced in this Thesis. The necessary steps for applying the methodology were detailed in addition to statistical parameters accounting for the suitability of the QSAR prediction models developed as evaluation functions for the desirability-based multi-objective optimization process: the *overall desirability determination coefficient* (R^2_D) and the *overall desirability's LOO–CV determination coefficient* (Q^2_D). A ranking procedure is also proposed to order libraries of compounds according to their structural similarity with an optimal theoretical candidate, as well as a measure of the quality of the ranking obtained: the *ranking quality index* (Ψ).

ii) The application it of the MOOP-DESIRE methodology lead to the design of a set of novel NSAIDs quinazolinones with simultaneously improved analgesic, antiinflammatory, and ulcerogenic profiles. The best compromise between the mentioned properties was established and new drug candidates with the highest overall desirability then designed. In particular, one of the designed candidates (compound **ASNEW8**) reached 93% of analgesic activity, 82% of inflammatory inhibition and an ulcerogenic index of 0.44, which represents an excellent overall desirability (= 0.8), being this accomplished by modifying the compounds' structure in such a way that pushed the values of the C-001, C-037 and H-046 predictor variables to 5, 0 and 12, respectively. Furthermore, it was observed that the presence of bulky alkyl substituents at the C-2 position of the quinazoline ring displayed a positive role on the ulcerogenic ability without a negative influence in the other properties. These results support the applicability of the MOOP-DESIRE methodology to the task of multi-criteria drug design.

iii) The usefulness of the MOOP-DESIRE methodology as multi-criteria library ranking tool was challenged by using it as a rational strategy for filtering safe and potent antibacterial candidates from a heterogeneous library of antibacterial fluoroquinolones. Each compound in the library was ranked according to a criterion of structural similarity with a pharmaceutically

optimal candidate (with the best possible compromise between antibacterial efficacy and cytotoxicity) previously obtained. Based on this criterion (Δ_i) 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. On the other hand, if the top 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 since this fraction exhibit a percentage of quality ranking of 82.74 ($\Psi^* = 0.173$).

iv) The MOOP-DESIRE methodology was applied to the prioritization of hits with appropriate trade-offs between HIV-1 RT inhibitor efficacy and MT4 blood cells toxicity. In this work was determined the theoretical levels of a set of molecular descriptors leading to a pharmaceutically desirable HIV-1 NNRTI candidate, using it as a pattern to rank libraries of new compounds according to the degree of structural similarity. The developed multi-objective optimization strategy was efficiently employed as a virtual screening tool by the prioritization of 12 NNRTI candidates with favourable pharmaceutical profiles disperse in a library of 432 NNRTI decoys extracted from DUD. In such a difficult task was possible to retrieve in the top 10% of the ordered library up to a third of the NNRTI candidates with favourable pharmaceutical profiles. The comparative study between the sequential, parallel and multi-objective virtual screening approaches of the selected library of compounds revealed that the multi-objective approach can be superior to the other approaches. Moreover, it can rule out the exclusion of pharmaceutically acceptable candidates. The data obtained so far evidences the potential of the MOOP-DESIRE methodology as multi-criteria virtual screening tool.

v) The development of a linear 1D prediction model of the A_3ARagonists overall desirability based on four simple molecular descriptors with a direct physicochemical or structural explanation, as well as the desirability analysis of this model was described in this work. The results obtained provided significant clues on desired trade-offs between binding and relative efficacy of N^6-substituted-4´-thioadenosines A_3AR agonists. The desirability-based prediction model interpretation strategy proposed here suggest a favorable effect over binding affinity and agonist efficacy of conformationally restricted, but not fused bicyclic N^6 substituents. The

overall data provide guides to the rational design of new A_3AR agonist candidates by assembling a combinatorial library useful for the prioritization of candidates with a promissory balance between A_3AR binding affinity and agonist efficacy through a virtual screening campaign. These results evidence the suitability of the Desirability Theory as interpretation tool for multi-criteria prediction models.

REFERENCES

1. Ekins S, Boulanger B, Swaan PW, Hupcey MA. Towards a new age of virtual ADME/TOX and multidimensional drug discovery. J Comput Aided Mol Des. 2002;16(5-6):381-401.
2. Federsel HJ. In search of sustainability: process R&D in light of current pharmaceutical industry challenges. Drug Discov Today. 2006 Nov;11(21-22):966-74.
3. Bleicher KH, Bohm HJ, Muller K, Alanine AI. Hit and lead generation: beyond high-throughput screening. Nat Rev Drug Discov. 2003 May;2(5):369-78.
4. Xu J, Hagler A. Chemoinformatics and Drug Discovery. Molecules. 2002;7:566-700.
5. Butina D, Segall MD, Frankcombe K. Predicting ADME properties in silico: methods and models. Drug Discov Today. 2002;7(11 Suppl):S83-S8.
6. Manly CJ, Louise-May S, Hammer JD. The impact of informatics and computational chemistry on synthesis and screening. Drug Discov Today. 2001 Nov 1;6(21):1101-10.
7. Seifert MHJ, Wolf K, Vitt D. Virtual high-throughput in silico screening. Drug Discov Today: Biosilico. 2003;1(4):143-9.
8. Kubinyi H. Virtual Screening (Lectures of the Drug Design Course)2006: Available from: http://kubinyi.de/dd-19.pdf.
9. Kubinyi H. Virtual Screening - The Road to Success (Lecture at the XIX International Symposium on Medicinal Chemistry, Istanbul)2006: Available from: http://kubinyi.de/istanbul-09-06.pdf.
10. Lahana R. How many leads from HTS? Drug Discov Today. 1999;4(10):447-8.
11. Mayer JM, van de Waterbeemd H. Development of quantitative structure-pharmacokinetic relationships. Environ Health Perspect. 1985 Sep;61:295-306.
12. Jorgensen WL. The many roles of computation in drug discovery. Science. 2004;303(5665):1813-8.
13. DiMasi JA, Hansen RW, Grabowski HG. The price of innovation: new estimates of drug development costs. J Health Econ. 2003 Mar;22(2):151-85.
14. Cruz-Monteagudo M, Cordeiro MNDS, Teijeira M, González MP, Borges F. Multidimensional Drug Design: Simultaneous Analysis of Binding and Relative Efficacy Profiles of N6-substituted-4´-thioadenosines A3 Adenosine Receptor Agonists. Chem Biol Drug Des. 2010;75(607–618).
15. Nicolotti O, Giangreco I, Miscioscia TF, Carotti A. Improving quantitative structure-activity relationships through multiobjective optimization. J Chem Inf Model. 2009 Oct;49(10):2290-302.
16. Cruz-Monteagudo M, Borges F, Cordeiro MN. Desirability-based multiobjective optimization for global QSAR studies: Application to the design of novel NSAIDs with improved analgesic, antiinflammatory, and ulcerogenic profiles. J Comput Chem. 2008 May 1;29(14):2445-59.
17. Nicolotti O, Gillet VJ, Fleming PJ, Green DV. Multiobjective optimization in quantitative structure-activity relationships: deriving accurate and interpretable QSARs. J Med Chem. 2002 Nov 7;45(23):5069-80.
18. Cruz-Monteagudo M, Borges F, Cordeiro MNDS, Cagide Fajin JL, Morell C, Molina Ruiz R, et al. Desirability-Based Methods of Multiobjective Optimization and Ranking for Global QSAR Studies. Filtering Safe and Potent Drug Candidates from Combinatorial Libraries. J Comb Chem. 2008;10(6):897-913.
19. Nicolaou CA, Apostolakis J, Pattichis CS. De Novo Drug Design Using Multiobjective Evolutionary Graphs. J Chem Inf Model. 2009 Jan 26.

20. Yamashita F, Hara H, Ito T, Hashida M. Novel hierarchical classification and visualization method for multiobjective optimization of drug properties: application to structure-activity relationship analysis of cytochrome P450 metabolism. J Chem Inf Model. 2008 Feb;48(2):364-9.
21. Machado A, Tejera E, Cruz-Monteagudo M, Rebelo I. Application of desirability-based multi(bi)-objective optimization in the design of selective arylpiperazine derivates for the 5-HT1A serotonin receptor. Eur J Med Chem. 2009 Dec;44(12):5045-54.
22. Cruz-Monteagudo M, The HP, Cordeiro MNDS, Borges F. Prioritizing Hits With Appropriate Trade-offs Between HIV-1 Reverse Transcriptase Inhibitory Efficacy and MT4 Blood Cells Toxicity Through Desirability-Based Multi-Objective Optimization and Ranking. Mol Inf. 2010;29:303–21.
23. Walters WP, Stahl MT, Murcko MA. Virtual screening - an overview. Drug Discov Today. 1998;3:160-78.
24. Fox S, Farr-Jones S, Yund MA. High throughput screening for drug discovery: continually transitioning into new technology. J Biomol Screening. 1999;4:183-6.
25. Young S, Li J. Virtual screening of focused combinatorial libraries. Innov Pharm Technol. 2000;28:24-6.
26. Brown N, Lewis RA. Exploiting QSAR methods in lead optimization. Curr Opin Drug Discov Devel. 2006;9(4):419-24.
27. Hansch C. On the structure of medicinal chemistry. J Med Chem. 1976 Jan;19(1):1-6.
28. Fukunaga JY, Hansch C, Steller EE. Inhibition of dihydrofolate reductase. Structure-activity correlations of quinazolines. J Med Chem. 1976 May;19(5):605-11.
29. Moriguchi I, Hirano H, Hirono S. Prediction of the rodent carcinogenicity of organic compounds from their chemical structures using the FALS method. Environ Health Perspect. 1996 Oct;104 Suppl 5:1051-8.
30. Estrada E. On the topological sub-structural molecular design (TOSS-MODE) in QSPR/QSAR and drug design research. SAR QSAR Environ Res. 2000;11(1):55-73.
31. Vilar S, Estrada E, Uriarte E, Santana L, Gutierrez Y. In silico studies toward the discovery of new anti-HIV nucleoside compounds through the use of TOPS-MODE and 2D/3D connectivity indices. 2. Purine derivatives. J Chem Inf Model. 2005 Mar-Apr;45(2):502-14.
32. Marrero-Ponce Y, Marrero RM, Torrens F, Martinez Y, Bernal MG, Zaldivar VR, et al. Non-stochastic and stochastic linear indices of the molecular pseudograph's atom-adjacency matrix: a novel approach for computational in silico screening and "rational" selection of new lead antibacterial agents. J Mol Model (Online). 2006 Feb;12(3):255-71.
33. Helguera AM, Cabrera Perez MA, Gonzalez MP. A radial-distribution-function approach for predicting rodent carcinogenicity. J Mol Model (Online). 2006 Jan 19:1-12.
34. Gonzalez-Diaz H, Cruz-Monteagudo M, Molina R, Tenorio E, Uriarte E. Predicting multiple drugs side effects with a general drug-target interaction thermodynamic Markov model. Bioorg Med Chem. 2005 Feb 15;13(4):1119-29.
35. Gonzalez-Diaz H, Cruz-Monteagudo M, Vina D, Santana L, Uriarte E, De Clercq E. QSAR for anti-RNA-virus activity, synthesis, and assay of anti-RSV carbonucleosides given a unified representation of spectral moments, quadratic, and topologic indices. Bioorg Med Chem Lett. 2005 Mar 15;15(6):1651-7.

36. Cruz-Monteagudo M, Gonzalez-Diaz H, Borges F, Gonzalez-Diaz Y. Simple stochastic fingerprints towards mathematical modeling in biology and medicine. 3. Ocular irritability classification model. Bull Math Biol. 2006 Oct;68(7):1555-72.
37. Cruz-Monteagudo M, Borges F, Perez Gonzalez M, Cordeiro MN. Computational modeling tools for the design of potent antimalarial bisbenzamidines: Overcoming the antimalarial potential of pentamidine. Bioorg Med Chem. 2007;15(15):5322-39.
38. Cruz-Monteagudo M, Cordeiro MN, Borges F. Computational chemistry approach for the early detection of drug-induced idiosyncratic liver toxicity. J Comput Chem. 2008;29(4):533-49.
39. Cruz-Monteagudo M, Gonzalez-Diaz H, Aguero-Chapin G, Santana L, Borges F, Dominguez ER, et al. Computational chemistry development of a unified free energy Markov model for the distribution of 1300 chemicals to 38 different environmental or biological systems. J Comput Chem. 2007 Aug;28(11):1909-23.
40. Gonzalez-Diaz H, Aguero G, Cabrera MA, Molina R, Santana L, Uriarte E, et al. Unified Markov thermodynamics based on stochastic forms to classify drugs considering molecular structure, partition system, and biological species: distribution of the antimicrobial G1 on rat tissues. Bioorg Med Chem Lett. 2005 Feb 1;15(3):551-7.
41. Prado-Prado FJ, Gonzalez-Diaz H, Santana L, Uriarte E. Unified QSAR approach to antimicrobials. Part 2: Predicting activity against more than 90 different species in order to halt antibacterial resistance. Bioorg Med Chem. 2007 Jan 15;15(2):897-902.
42. Gonzalez-Diaz H, Prado-Prado FJ, Santana L, Uriarte E. Unify QSAR approach to antimicrobials. Part 1: predicting antifungal activity against different species. Bioorg Med Chem. 2006 Sep 1;14(17):5973-80.
43. Gonzalez-Diaz H, Prado-Prado FJ. Uniffied QSAR and network-based computational chemistry approach to antimicrobials, Part 1: Multispecies activity models for antifungals. J Comput Chem. 2008;29(4):656-67.
44. Nicolaou AC, Brown N, Pattichis CS. Molecular optimization using computational multi-objective methods. Curr Opin Drug Discov Devel. 2007;10(3):316-24.
45. Yann C, Siarry P, editors. Multiobjective Optimization: Principles and Case Studies. Berlin, Germany: Springer-Verlag; 2004.
46. Jones G, Willett P, Glen RC. A genetic algorithm for flexible molecular overlay and pharmacophore elucidation. J Comput Aided Mol Des. 1995;9(6):532-49.
47. Handschuh S, Wagener M, Gasteier J. Superposition of three-dimensional chemical structures allowing for conformational flexibility by a hybrid method. J Chem Inf Comput Sci. 1998;38(2):220-32.
48. Shepphird JK, Clarck RD. A marriage made in torsional space: Using GALAHAD models to drive pharmacophore multiple searches. J Comput Aided Mol Des. 2006;20(12):735-49.
49. Janson S, Merkle D. A new multi-objective particle swarm optimization algorithm using clustering applied to automated docking. In: Blesa MJ, Blum C, Roli A, Sampels M, editors. Hybrid Metaheuristics Second International Workshop, HM 2005; August 29-30; Barcelona, Spain: Springer-Verlag; 2005. p. 128-41.
50. Zoete V, Grosdidier A, Michielin O, editors. EADock: A new approach to the docking of small molecules to protein active sites. High Performance Computing for the Life Sciences Symposium; 2005; Lausanne, Switzerland.

51. Brown N, Mckay B, Gasteiger J. A novel workflow for the inverse QSPR problem using multiobjective optimization. J Comput Aided Mol Des. 2006;20(5):333-41.
52. Lameijer EW, Kok JN, Back T, Ijerman AP. The molecule avoluator. An iteractive evolutionary algorithm for the design of drug-like molecules. J Chem Inf Model. 2006;46(2):545-52.
53. Stockfisch TP. Partially unified multiple property recursive partitioning (PUMP-RP): A new method for predicting and understanding drug selectivity. J Chem Inf Comput Sci. 2003;43(5):1608-13.
54. Rao SN, Stockfisch TP. Partially unified multiple property recursive partitioning (PUMP-RP) analyses of cyclooxygenase (COX) inhibitors. J Chem Inf Comput Sci. 2003;43(5):1614-22.
55. Harrington EC. The Desirability Function. Ind Quality Control. 1965;21(10):494-8.
56. Derringer G, Suich R. Simultaneous Optimization of Several Response Variables. J Quality Technol. 1980;12(4):214-9.
57. Outinen K, Haario H, Vuorela P, Nyman M, Ukkonen E, Vuorela H. Optimization of selectivity in high-performance liquid chromatography using desirability functions and mixture designs according to PRISMA. Eur J Pharm Sci. 1998 Jul;6(3):197-205.
58. Garcia-Gonzalez DL, Aparicio R. Detection of vinegary defect in virgin olive oils by metal oxide sensors. J Agric Food Chem. 2002 Mar 27;50(7):1809-14.
59. Shih M, Gennings C, Chinchilli VM, Carter WH, Jr. Titrating and evaluating multi-drug regimens within subjects. Stat Med. 2003 Jul 30;22(14):2257-79.
60. Kording KP, Fukunaga I, Howard IS, Ingram JN, Wolpert DM. A neuroeconomics approach to inferring utility functions in sensorimotor control. PLoS Biol. 2004 Oct;2(10):e330.
61. Safa F, Hadjmohammadi MR. Simultaneous optimization of the resolution and analysis time in micellar liquid chromatography of phenyl thiohydantoin amino acids using Derringer's desirability function. J Chromatogr A. 2005 Jun 17;1078(1-2):42-50.
62. Pavan M, Todeschini R, Orlandi M. Data mining by total ranking methods: a case study on optimisation of the "pulp and bleaching" process in the paper industry. Ann Chim. 2006 Jan-Feb;96(1-2):13-27.
63. Coffey T, Gennings C, Moser VC. The simultaneous analysis of discrete and continuous outcomes in a dose-response study: using desirability functions. Regul Toxicol Pharmacol. 2007 Jun;48(1):51-8.
64. Rozet E, Wascotte V, Lecouturier N, Preat V, Dewe W, Boulanger B, et al. Improvement of the decision efficiency of the accuracy profile by means of a desirability function for analytical methods validation. Application to a diacetyl-monoxime colorimetric assay used for the determination of urea in transdermal iontophoretic extracts. Anal Chim Acta. 2007 May 22;591(2):239-47.
65. Wong WK, Furst DE, Clements PJ, Streisand JB. Assessing disease progression using a composite endpoint. Stat Methods Med Res. 2007 Feb;16(1):31-49.
66. Cojocaru C, Khayet M, Zakrzewska-Trznadel G, Jaworska A. Modeling and multi-response optimization of pervaporation of organic aqueous solutions using desirability function approach. J Hazard Mater. 2008 Dec 25.
67. Fajar NM, Carro AM, Lorenzo RA, Fernandez F, Cela R. Optimization of microwave-assisted extraction with saponification (MAES) for the determination of polybrominated flame retardants in aquaculture samples. Food Addit Contam Part A Chem Anal Control Expo Risk Assess. 2008 Aug;25(8):1015-23.

68. Jancic-Stojanovic B, Malenovic A, Ivanovic D, Rakic T, Medenica M. Chemometrical evaluation of ropinirole and its impurity's chromatographic behavior. J Chromatogr A. 2009 Feb 20;1216(8):1263-9.
69. Huang J, Ling CX. Rank Measures for Ordering. In: Jorge Aea, editor. PKDD 2005; LNAI 3721: Springer-Verlag Berlin Heidelberg; 2005. p. 503–10.
70. Atkinson AC. Plots, Transformations and Regression: Oxford:Clarendon Press; 1985.
71. Eriksson L, Jaworska J, Worth AP, Cronin MT, McDowell RM, Gramatica P. Methods for reliability and uncertainty assessment and for applicability evaluations of classification- and regression-based QSARs. Environ Health Perspect. 2003 Aug;111(10):1361-75.
72. Kutner MH, Nachtsheim CJ, Neter J, Li W. Applied Linear Statistical Models. Fifth ed. New York: McGraw Hill; 2005.
73. Statsoft_Inc. STATISTICA. 6.0 for Windows ed2001.
74. Nelder JA, Mead R. A Simplex method for function minimization. Computer Journal. 1965;7:308-13.
75. Fletcher R, Reeves CM. Function minimization by conjugate gradients. Computer Journal. 1964;7:149-54.
76. Hooke R, Jeeves TA. Direct search solution of numerical and statistical problems. J Assoc Comp Machin. 1961;8:212-29.
77. Watson I, Marir F. Case-based reasoning: a review. The knowledge engineering review. 1994;9(4):Cambridge, UK: Cambridge University Press.
78. Coleman TF, Li Y. An Interior, Trust Region Approach for Nonlinear Minimization Subject to Bounds. SIAM J Optim. 1996;6:418-45.
79. Coleman TF, Li Y. On the Convergence of Reflective Newton Methods for Large-Scale Nonlinear Minimization Subject to Bounds. Math Program. 1994;67(2):189-224.
80. MATLAB. 7.2 ed: The MathWorks, Inc.; 2006.
81. Alagarsamy V, Dhanabal K, Parthiban P, Anjana G, Deepa G, Murugesan B, et al. Synthesis and pharmacological investigation of novel 3-(3-methylphenyl)-2-substituted amino-3H-quinazolin-4-ones as analgesic and anti-inflammatory agents. J Pharm Pharmacol. 2007;59:669–77.
82. Suto MJ, Domagala JM, Roland GE, Mailloux GB, Cohen MA. Fluoroquinolones: Relationships between Structural Variations, Mammalian Cell Cytotoxicity, and Antimicrobial Activity. J Med Chem. 1992;35:4745-50.
83. Ranise A, Spallarossa A, Schenone S, Bruno O, Bondavalli F, Vargiu L, et al. Design, synthesis, SAR, and molecular modeling studies of acylthiocarbamates: a novel series of potent non-nucleoside HIV-1 reverse transcriptase inhibitors structurally related to phenethylthiazolylthiourea derivatives. J Med Chem. 2003 Feb 27;46(5):768-81.
84. Sun GF, Chen XX, Chen FE, Wang YP, De Clercq E, Balzarini J, et al. Nonnucleoside HIV-1 reverse-transcriptase inhibitors, part 5. Synthesis and anti-HIV-1 activity of novel 6-naphthylthio HEPT analogues. Chem Pharm Bull (Tokyo). 2005 Aug;53(8):886-92.
85. Ji L, Chen FE, De Clercq E, Balzarini J, Pannecouque C. Synthesis and anti-HIV-1 activity evaluation of 5-alkyl-2-alkylthio-6-(arylcarbonyl or alpha-cyanoarylmethyl)-3,4-dihydropyrimidin-4(3H)-ones as novel non-nucleoside HIV-1 reverse transcriptase inhibitors. J Med Chem. 2007 Apr 19;50(8):1778-86.
86. Xiong YZ, Chen FE, Balzarini J, De Clercq E, Pannecouque C. Non-nucleoside HIV-1 reverse transcriptase inhibitors. Part 11: structural modulations of diaryltriazines with potent anti-HIV activity. Eur J Med Chem. 2008 Jun;43(6):1230-6.

87. Jeong LS, Lee HW, Kim HO, Jung JY, Gao ZG, Duong HT, et al. Design, synthesis, and biological activity of N6-substituted-4'-thioadenosines at the human A3 adenosine receptor. Bioorg Med Chem. 2006 Jul 15;14(14):4718-30.
88. CambridgeSoft. ChemDraw Ultra. 9.0 ed2004.
89. Burkert U, Allinger NL. Molecular Mechanics. Washington, D.C., USA: ACS; 1982.
90. Clark T. Computational Chemistry. N.Y., USA: Wiley; 1985.
91. Frank J. MOPAC. 6.0 ed: Seiler Research Laboratory, US Air Force Academy, Colorado Springs, CO.; 1993.
92. Todeschini R, Consonni V, Pavan M. DRAGON Software. 5.4 ed. Milano: Talete srl; 2006.
93. Leardi R, Boggia R, Terrile M. Genetic algorithms as a strategy for feature selection. J Chemom. 1992;6:267-81.
94. Yasri A, Hartsough D. Toward an optimal procedure for variable selection and QSAR model building. J Chem Inf Comput Sci. 2001 Sep-Oct;41(5):1218-27.
95. Hou TJ, Wang JM, Liao N, Xu XJ. Applications of genetic algorithms on the structure-activity relationship analysis of some cinnamamides. J Chem Inf Comput Sci. 1999 Sep-Oct;39(5):775-81.
96. Hasegawa K, Kimura T, Funatsu K. GA strategy for variable selection in QSAR studies: application of GA-based region selection to a 3D-QSAR study of acetylcholinesterase inhibitors. J Chem Inf Comput Sci. 1999 Jan-Feb;39(1):112-20.
97. Barbosa de Oliveira D, Gaudio AC. BuildQSAR. Vitória ES, Brasil: Physics Department-CCE, University of Espírito Santo; 2000.
98. Barbosa de Oliveira D, Gaudio AC. BuildQSAR: A new computer program for QSAR analysis. Quant Struct-Act Relat. 2000;19:599-601.
99. Todeschini R, Consonni V, Pavan M. MOBY DIGS. 1.2 for Windows ed. Milan, Italy: Talete srl 2002. p. Software for Multilinear Regression Analysis and Variable Subset Selection by Genetic Algorithm.
100. Efron B. Better Bootstrap Confidence Intervals. Journal of American Statistical Association. 1987;82:171-200.
101. Lindgren F, Hansen B, Karcher W, Sjöström M, Eriksson L. Model validation by permutation tests: Applications to variable selection. J Chemom. 1996;10(5):521-32.
102. Stewart J, Gill L. Econometrics. 2nd edition ed. Allan P, editor. London: Prentice Hall; 1998.
103. Kutner MH, Nachtsheim CJ, Neter J, Li W. Multicollinearity and its effects. Applied Linear Statistical Models. Fifth ed. New York: McGraw Hill; 2005. p. 278-89.
104. De Boor C. A practical guide to splines. New York: Springer-Verlag; 1978.
105. Gerald CF, Wheatley PO. Applied numerical analysis. 4th ed. Reading, MA: Addison Wesley; 1989.
106. Adamson GW, Lynch MF, Town WG. Analysis of structural characteristics of chemical compounds in a large computer-based tile. Part 2. Atom centred fragments. J Chem Sot. 1970;C:3702-6.
107. Drews J. Innovation deficit revisited: Reflections on the productivity of pharmaceutical R&D. Drug Discov Today. 1998;3:491-4.
108. Truchon JF, Bayly CI. Evaluating virtual screening methods: good and bad metrics for the "early recognition" problem. J Chem Inf Model. 2007 Mar-Apr;47(2):488-508.
109. Kirchmair J, Markt P, Distinto S, Wolber G, Langer T. Evaluation of the performance of 3D virtual screening protocols: RMSD comparisons,

enrichment assessments, and decoy selection--what can we learn from earlier mistakes? J Comput Aided Mol Des. 2008 Mar-Apr;22(3-4):213-28.
110. Fechner U, Schneider G. Evaluation of distance metrics for ligand-based similarity searching. Chembiochem. 2004 Apr 2;5(4):538-40.
111. Witten IH, Frank E. Data Mining: Practical Machine Learning Tools and Techniques. 2nd ed. Gray J, editor. San Francisco: Morgan Kaufmann; 2005.
112. Gillet VJ. New directions in library design and analysis. Curr Opin Chem Biol. 2008 Jun;12(3):372-8.
113. Segall MD, Beresford AP, Gola JM, Hawksley D, Tarbit MH. Focus on success: using a probabilistic approach to achieve an optimal balance of compound properties in drug discovery. Expert Opin Drug Metab Toxicol. 2006 Apr;2(2):325-37.
114. Huang N, Shoichet BK, Irwin JJ. Benchmarking sets for molecular docking. J Med Chem. 2006 Nov 16;49(23):6789-801.
115. Kim SK, Jacobson KA. Three-dimensional quantitative structure-activity relationship of nucleosides acting at the A3 adenosine receptor: analysis of binding and relative efficacy. J Chem Inf Model. 2007 May-Jun;47(3):1225-33.
116. Schüller A, Schneider G, Byvatov E. SMILIB: Rapid Assembly of Combinatorial Libraries in SMILES Notation. QSAR & Comb Science. **2003**;22:719-21.
117. Dempster AP. Upper and Lower Probabilities Induced by a Multivalued Mapping. Ann Stat. 1967;28:325-39.
118. Hooper G. A calculation of the credibility of human testimony. Philosophical Transaction of the Royal Society. 1699;21:359-65.
119. Shafer G. The combination of evidence. Int J Intell Syst. 1986;1(3):155-79.
120. Muchmore SW, Debe DA, Metz JT, Brown SP, Martin YC, Hajduk PJ. Application of belief theory to similarity data fusion for use in analog searching and lead hopping. J Chem Inf Model. 2008 May;48(5):941-8.

ANNEXES

No.	Ref.	
I	(16)	Cruz-Monteagudo M, Borges F, Cordeiro MNDS. Desirability-Based Multi-Objective Optimization for Global QSAR Studies. Application to the Design of Novel NSAIDs with Improved Analgesic, Anti-Inflammatory and Ulcerogenic Profiles. ***Journal of Computational Chemistry*** 2008, 29, 2445–2459.
II	(18)	Cruz-Monteagudo M, Borges F, Cordeiro MNDS, Fajín JLC, Morell C, Molina RR, Cañizares-Carmenate Y, Domínguez ER. Desirability-Based Methods of Multi-Objective Optimization and Ranking for Global QSAR Studies. Filtering Safe and Potent Drug Candidates from Combinatorial Libraries. ***Journal of Combinatorial Chemistry***. 2008, 10, 897–913.
III	(22)	Cruz-Monteagudo M, The HP, Cordeiro MNDS, Borges F. Prioritizing Hits With Appropriate Trade-offs Between HIV-1 Reverse Transcriptase Inhibitory Efficacy and MT4 Blood Cells Toxicity Through Desirability-Based Multi-Objective Optimization and Ranking. ***Molecular Informatics*** 2010, 29, 303–321.
IV	(14)	Cruz-Monteagudo M, Cordeiro MNDS, Teijeira M, González MP, Borges F. Multidimensional Drug Design: Simultaneous Analysis of Binding and Relative Efficacy Profiles of N6-substituted-4´-thioadenosines A3 Adenosine Receptor Agonists. ***Chemical Biology & Drug Design*** 2010, 75, 607–618.

ANNEX I

Desirability-Based Multiobjective Optimization for Global QSAR Studies: Application to the Design of Novel NSAIDs with Improved Analgesic, Antiinflammatory, and Ulcerogenic Profiles

MAYKEL CRUZ-MONTEAGUDO,[1,2,3] FERNANDA BORGES,[1] M. NATÁLIA D. S. CORDEIRO[4]
[1]*Physico-Chemical Molecular Research Unit, Department of Organic Chemistry, Faculty of Pharmacy, University of Porto, 4150-047 Porto, Portugal*
[2]*Applied Chemistry Research Center, Faculty of Chemistry and Pharmacy, Central University of "Las Villas", Santa Clara 54830, Cuba*
[3]*Chemical Bioactive Center, Central University of "Las Villas", Santa Clara 54830, Cuba*
[4]*REQUIMTE, Department of Chemistry, Faculty of Sciences, University of Porto, 4169-007 Porto, Portugal*

Received 24 October 2007; Accepted 6 March 2008
DOI 10.1002/jcc.20994
Published online 1 May 2008 in Wiley InterScience (www.interscience.wiley.com).

Abstract: Up to now, very few reports have been published concerning the application of multiobjective optimization (MOOP) techniques to quantitative structure–activity relationship (QSAR) studies. However, none reports the optimization of objectives related directly to the desired pharmaceutical profile of the drug. In this work, for the first time, it is proposed a MOOP method based on Derringer's desirability function that allows conducting global QSAR studies considering simultaneously the pharmacological, pharmacokinetic and toxicological profile of a set of molecule candidates. The usefulness of the method is demonstrated by applying it to the simultaneous optimization of the analgesic, antiinflammatory, and ulcerogenic properties of a library of fifteen 3-(3-methylphenyl)-2-substituted amino-3*H*-quinazolin-4-one compounds. The levels of the predictor variables producing concurrently the best possible compromise between these properties is found and used to design a set of new optimized drug candidates. Our results also suggest the relevant role of the bulkiness of alkyl substituents on the C-2 position of the quinazoline ring over the ulcerogenic properties for this family of compounds. Finally, and most importantly, the desirability-based MOOP method proposed is a valuable tool and shall aid in the future rational design of novel successful drugs.

 J Comput Chem 29: 2445–2459, 2008

Key words: chemoinformatics; drug discovery; global QSAR; multiobjective optimization; NSAIDs; overall desirability function; ulcerogenic index

Introduction

Developing a successful drug is a complex and lengthy process and failure at the development stage is due to multiple factors, such as lack of efficacy, poor bioavailability, and toxicity.[1] Improving 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 highlights 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.[2] Roughly 75% of the total costs during the development of a drug are attributed to poor pharmacokinetics and/or toxicity.[3]

In the last years, the drug discovery/development process has been gaining in efficiency and rationality because of the continuous progress and application of chemoinformatics methods.[2] In

Additional Supporting Information may be found in the online version of this article.

Correspondence to: M. Cruz-Monteagudo; e-mail: gmailkelcm@yahoo.es

Contract/grant sponsor: Fundação para a Ciência e a Tecnologia (FCT); contract/grant numbers: SFRH/BD/30698/2006

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] Yet, standard chemoinformatics approaches usually ignore multiple objectives and optimize each biological property sequentially.[6–17] Nevertheless, some efforts have been made recently toward unified approaches able of modeling multiple pharmacological, pharmacokinetic, or toxicological properties onto a single QSAR equation.[18–21]

Multiobjective optimization (MOOP) methods introduce a new philosophy for reaching optimality based on compromises among the various objectives. These methods aim at discovering the global optimal solution by optimizing 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, leading thus to a more efficient overall process.[22]

Several applications of MOOP methods have appeared lately ranging from substructure mining to docking, including inverse quantitative structure property relationship (QSPR) and QSAR.[22] Most of these MOOP applications have been based on the following approaches: weighted-sum-of-objective-functions (WSOF)[23] and pareto-based methods.[22] An excellent review on the subject has been most recently published by Nicolaou et al.[22]

Concerning substructure mining, MOOP applications have focused on molecular alignment and pharmacophore identification. Examples of MOOPs tackling the substructure mining from a multiobjective perspective include the genetic algorithm similarity program method (a WSOF-based method)[24] and some pareto-based methods, such as the genetic algorithm for multiple molecular alignment method (probably the first application of a pareto-based approach in chemoinformatics),[25] and the genetic algorithm with linear assignment for the hypermolecular alignment of datasets.[26]

As regards docking, several research groups are particularly active using pareto-based MOOP methods. For instance, Janson et al.[27] described a docking optimization application termed ClustMPSO, based on the particle swarm optimization algorithm that minimizes simultaneously the intermolecular energy between the protein and the ligand and the intramolecular energy of the ligand. A multiobjective evolutionary algorithm has also been used by Zoete et al.[28] in their docking program EADock.

Recently, MOOP methods have been applied to the optimization of new chemical entities via *de novo* molecular design and inverse QSPR. In this area, there are notable applications such as the CoG approach introduced by Brown et al.[29] to solve the inverse QSPR problem as well as the Molecule Evoluator proposed by Lameijer et al.,[30] where the user assumes the role of the fitness function by selecting candidate molecules for further evolution after each iteration.

Finally, 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. Nicolotti et al.[31] employed a variant of an evolutionary algorithm called multiobjective genetic programming that used pareto ranking to optimize the QSAR models. A number of conflicting objectives including model accuracy, number of terms, internal complexity, and interpretability of the descriptors used in the model were considered. On the other hand, Stockfisch[32] proposed a nonevolutionary multiobjective technique called the partially unified multiple property recursive partitioning method for building QSAR models. This method was successfully used to construct models to analyze selectivity relationships between cyclooxygenase 1 and 2 inhibitors.[33] Up to now, no QSAR study has nevertheless reported the simultaneous optimization of competing objectives directly related with the definitive pharmaceutical profile of drugs, such as therapeutic efficacy, bioavailability, and/or toxicity.

In the present work, we are proposing for the first time a MOOP method based on Derringer's desirability function[34] that allows running global QSAR studies jointly considering multiple properties of interest to the drug-design process. The method proposed is applied to a small set of 2-substituted amino-3*H*-quinazolin-4-one compounds with the aim of simultaneously optimizing their analgesic, antiinflammatory and ulcerogenic properties, as well as suggesting new improved drug candidates of this kind.

Materials and Methods

Data Set

Our prediction models (PMs) were developed using a library of fifteen 3-(3-methylphenyl)-2-substituted amino-3*H*-quinazolin-4-one compounds published by Alagarsamy et al.[35] The analgesic activity (*An*) reported for these compounds (in %) was measured using the tail-flick method in Wistar albino mice,[36] whereas the antiinflammatory activity (*Aa*) reported (in %) was evaluated using the carrageenan-induced paw oedema test in rats.[36] The ulcerogenic index (*U*) was determined by the method of Ganguly and Bhatnagar,[37] and the ulcers were induced in rats using the method described by Goyal et al.[38] All these assays[35] were performed by administering a maximum dose of 20 mg kg^{-1}.

Computational Methods

The structures of all compounds were first drawn with the aid of ChemDraw software package,[39] and reasonable starting geometries obtained by resorting to the MM2 molcular mechanics force field.[40,41] Molecular structures were then fully optimized with the PM3 semiempirical Hamiltonian,[39] implemented in the MOPAC 6.0 program.[42] Here, it should be remarked that the final molecular structures pertain only to the compounds' global minimum energy conformations, and indeed, further molecular simulations and/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[43] for computing a total of 120 atom-centered fragment (ACF) molecular descriptors.[44] ACF

Table 1. Symbols and Description for the 12 ACF Descriptors Remaining After Variable Reduction.

Symbol	Description	Symbol	Description
C-001	CH3R/CH4	C-038	Al−C(=X)−Al
C-002	CH2R2	C-039	Ar−C(=X)−R
C-024	R...CH...R	H-046	H attached to C0(sp^3) no X attached to next C
C-025	R...CR...R	H-052	H attached to C0(sp^3) with 1X attached to next C
C-026	R...CX...R	O-061	O...
C-037	Ar−CH=X	Cl-089	Cl attached to C1(sp^2)

descriptors were chosen because their simple nature offers easy structural interpretation. To reduce noisy information that could lead to chance correlations, descriptors having constant or near constant values as well as highly pair-correlated ($|R| > 0.95$) were excluded. Thus, from an initial set of 120 ACF molecular descriptors only 12 remained for further variable selection. Table 1 summarizes and describes the ACF molecular descriptors used in this work.

The task of selecting the descriptors that will be more suitable to model the activity of interest is complicated, as there are no absolute criteria for ruling such selection. Approaches implementing genetic algorithms (GA) for solving optimization problems in ANN[45–47] and SVM[48] based QSAR have been recently reported. 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). Herein, the GA optimization technique was applied for variable selection[49–52] by using the BuildQSAR software package.[53,54] The particular GA simulation conditions applied here were 10,000 generations, 300 model populations and 35% of mutation probabilities. Figure 1 depicts the ACF molecular descriptors selected by the GA method, which were finally applied to model the analgesic, antiinflammatory, and ulcerogenic properties of the present compounds.

As to 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.[55] 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 predictive ability of the models was evaluated by means of internal cross-validation (CV), specifically by the leave-one-out (LOO) technique.[56] Basically, LOO consists of forming N subsets from the entire dataset, each missing one point, which in turn is used to validate a new model that is trained with the corresponding subset. Quality of the new models (CV R^2: Q^2_{LOO}) gives then 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.[57] These include the linearity of the modeled property, homoscedasticity (or homogeneity of variance) as well as the normal distribution of the residuals and nonmulticollinearity between the descriptors.[58]

Finally, the applicability domain of the final PMs was identified by a leverage plot, that is to say, a plot of the standardized residuals.vs. leverages for each training compound.[56,59] The leverage (h_i) of a compound in the original variable space measures its influence on the model, and is calculated as follows:

$$h_i = \mathbf{t}_i(\mathbf{T}^T\mathbf{T})^{-1}\mathbf{t}_i^T \quad (i = 1, \ldots, N) \tag{1}$$

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

$$h^* = 3 \times p'/N \tag{2}$$

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 were considered reliable.

MOOP Based on the Desirability Estimation of Several Interrelated Responses

Improving the profile of a molecule for the drug discovery and development process requires the simultaneous optimization of

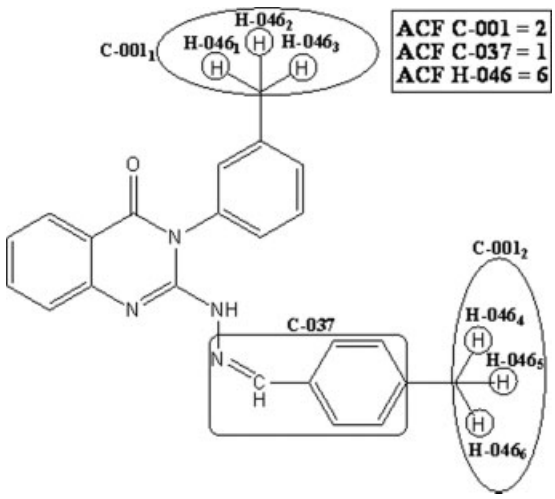


Figure 1. Atom-centered fragments (ACF) descriptors for compound **AS14**.

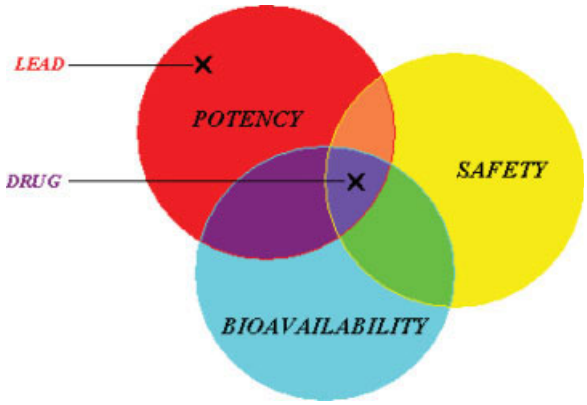


Figure 2. Graphic representation of the compromise between therapeutic efficacy (potency), bioavailability (ADME properties), and toxicity (safety) required to reach a successful drug.

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, to discover such a drug is almost a chimera and, if possible, an extremely difficult, expensive and time-consuming task. However, finding the best compromise between such objectives is an accessible and more realistic target (see Figure 2).

In this work, we are proposing a MOOP technique based on the desirability estimation of several interrelated responses (MOOP-DESIRE) as a tool for performing global QSAR studies, taking into account both the pharmacological, pharmacokinetic and toxicological profiles of a set of candidates. MOOP-DESIRE methodology is intended to find the most desirable solution that optimizes a multiobjective problem by using the Derringer's desirability function,[56,59] specifically addressed to confer rationality to the drug development process.

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

1. Prediction Models Set-Up

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 ($\hat{Y}_i$) is then estimated by a least-squares regression technique.

In some cases, the developed PM for some response may share the same independent variables of the other responses' PMs, but with different coefficients. In this atypical case, attaining the best compromise among the responses turns out to be simpler. Actually, due to 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, i.e. to set-up a global PM where the predicted values of each response are fitted to a linear function using all the independent variables employed in modeling the k original responses. Here, the independent variables used in computing the predicted values for the original responses will be used. Independent variables not used in computing the predicted values for the original responses will be zero.

2. Desirability Functions Selection and Evaluation

For each predicted response $\hat{Y}_i$, a desirability function d_i assigns values between 0 and 1 to the possible values of $\hat{Y}_i$. This transformed response, d_i, can have many different shapes. Regardless of the shape, $d_i = 0$ represents a completely undesirable value of $\hat{Y}_i$, 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:

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

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, 1] 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.[34]

Let L_i, U_i and T_i be the lower, upper, and target values, respectively, that are desired for the response $\hat{Y}_i$, 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

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

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

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

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:

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

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 towards T_i. Large values for s and t should be selected if it is very desirable that the value of $\hat{Y}_i$ 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 $\hat{Y}_i$ above L_i, and below U_i are acceptable or if having values of $\hat{Y}_i$ considerably above L_i are not of critical importance.[34]

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 $\hat{Y}_i$, there is no way to know how reliable the predicted D value of each candidate is.

To overcome this shortcoming, we propose here a statistical parameter, the overall desirability's determination coefficient (R^2_{D}), which measures the effect of the set of independent variables X_n in reducing 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 $\hat{Y}_i$'s [eqs. (2–4)], and the overall desirability D is a continuous function of the d_i's [eq. (1)], then D is also a continuous function of the X_n. Therefore, R^2_{D} can be computed in analogy with the so-called determination coefficient R^2. Specifically, R^2_{D} is computed by using the observed D_{Y_i} (calculated from Y_i) and the predicted $D_{\hat{Y}_i}$ (calculated from $\hat{Y}_i$) overall desirability values instead of using directly the measured (Y_i) and predicted ($\hat{Y}_i$) response values.

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

where D_{Y_i} and $D_{\hat{Y}_i}$ have been defined previously. $\overline{D}_{Y_i}$ 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^2_{D}) can be computed as shown below.

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

Like this, both R^2_{D} and *Adj.* R^2_{D} have the same properties of R^2 and Adj. R^2. Thus, both will fall in the range [0, 1] and the larger R^2_{D}/Adj. R^2_{D} is, the lower is the uncertainty in predicting D by using a specific set of independent variables X_n.[60]

Since R^2_{D} and Adj. R^2_{D} measure the goodness of fit rather than the predictive ability of a certain PM, it is advisable to use an analogous of the leave one out CV determination coefficient (Q^2_{LOO}) to establish the reliability of the method in predicting D. For this, the overall desirability's LOO–CV determination coefficient (Q^2_{D}) can be defined in an analogous way as R^2_{D}:

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

where $\mathrm{SSE}_{\mathrm{LOO-CV}}$ and $D_{\hat{Y}_i}$(LOO–CV) are the leave one out CV square sum of residuals and the predicted overall desirability by LOO–CV, respectively.

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.

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.[55] The overall desirability D is optimized with the "Use general function optimization" option, that is, the simplex method of function optimization,[61–63] 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 latter 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.

An added benefit of the method is the ability to plot D as a function of one or more independent variables. This allows the user to find a tendency in the relationship between responses and independent variables by considering the shape of the desirability function related to each independent variable, which then permits to establish an optimal range for each independent variable over the optimum values determined in the optimization process.

The final goal is to find the optimum levels (or an optimum 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).

Desirability Functions Specifications

Response/desirability profiling allows one to trace the response surface produced by fitting the observed response(s) using equation(s) based on the levels of the independent variables.[34] That is to say, one can inspect the predicted values for the response(s) at different combinations of levels of the independent variables,

Table 2. Regression Coefficients and Statistical Parameters for the MLR Models.

Analgesic activity (*An*) model								
$An = 51.762(\pm 2.155) + 8.333(\pm 0.957) \cdot C - 001 - 6.929(\pm 1.534) \cdot C - 037$								
N	R	R^2	R^2Adj.	Q^2	SPRESS	ρ	F	p
15	0.967	0.935	0.923	0.905	3.143	5.000	85.15699	0.000000
Antiinflammatory activity (*Aa*) model								
$Aa = 36.708(\pm 1.789) + 5.527(\pm 1.232) \cdot C - 001 + 1.475(\pm 0.430) \cdot H - 046$								
N	R	R^2	R^2Adj.	Q^2	SPRESS	ρ	F	p
15	0.942	0.887	0.869	0.827	3.526	5.000	47.46719	0.000002
Ulcerogenic index (*U*) model								
$U = 0.718(\pm 0.044) - 0.056(\pm 0.020) \cdot C - 001 + 0.137(\pm 0.032) \cdot C - 037$								
N	R	R^2	R^2 Adj.	Q^2	SPRESS	ρ	F	p
15	0.896	0.803	0.771	0.713	0.065	5.000	24.56766	0.000057

specify desirability function(s) for the response(s), and search for the levels of the independent variables that simultaneously produce the most desirable response or the best possible compromise among responses leading to the most desirable solution (candidate molecule).

In the present work, the optimization of the overall desirability was carried on by the Optimum desirability at exact grid points option of the general regression module of STATISTICA.[55] Three desirability functions, one for each response, were fitted. Specifically, the analgesic and antiinflammatory activities ought to be maximized [eq. (3)]. For estimating their d_i's, the lower value L_i was set to 25%, and the upper value U_i, made equal to the target value T_i, was set to 100% for both responses. In contrast, the ulcerogenic index must be minimized where $L_i = T_i = 0$ and $U_i = 1.73$ [eq. (4)]. The value of $U_i = 1.73$ corresponds to the ulcerogenic index of aspirin (measured with the same protocol used for the training set[35]), a NSAID with a recognized ulcerogenic ability. Furthermore, the spline method[64,65] was used for fitting the desirability function and surface/contours maps, and the current level of each independent variable was set equal to its optimum value. As to the s and t parameters, these were fixed at 1.00 by assuming that the desirability functions increase linearly towards T_i on the three responses.

Results and Discussion

MOOP–DESIRE-Based Optimization

Following the strategy outlined previously, we began by seeking the best linear models relating each property to the ACF 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 two ACF[44] variables previously selected by GA was developed for each property. The resulting best-fit models are given in Table 2 together with the statistical regression parameters, whereas the computed ACF molecular descriptors along with the measured and predicted values of the analgesic activity, antiinflammatory activity, and the ulcerogenic index for the 15 training compounds are shown in Table 3.

As can be noticed, the models are good in both statistical significance and predictive ability (see Table 2). Good overall quality of the models is revealed by the large F and small p values, satisfactory ρ values ($\rho = 5$), along with R^2 and Adj.R^2 (goodness of fit) values ranging from 0.803 to 0.935 and 0.771 to 0.923, respectively; as well as Q^2_{LOO} (predictivity) values between 0.713 and 0.905.

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 SMI).

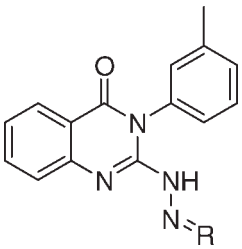
Another aspect deserving special attention is the applicability domain of the several PMs. The leverage values (h) and standardized residuals (Std. Res.) related to three PMs for the 15 training compounds are shown in Table 4, whereas Figure 3 shows the corresponding leverage plots. From these plots, the applicability domain is established inside a squared area within $\pm$ 2 standard deviations and a leverage threshold h^* of 0.6 (Notice that each model was fitted using 15 training compounds and included 3 adjustable parameters: two ACF descriptors plus the intercept.). As seen in Figure 3, only one compound of the training set has a leverage greater than h^* for *Aa*, but shows standard deviation values within the limits, which implies that it should not be considered an outlier but instead as an influential compound.

So far, we have demonstrated the satisfactory accuracy and the predictive ability of the developed PMs. We may now thus proceed with an adequate level of confidence to the simultaneous optimization of the analgesic, antiinflammatory and ulcerogenic properties for the set of compounds. Here it is important to remark that, since D is maximized directly over the independent variables domain, and at the same time, the predicted D values depend on the initial set of PMs, one should consider the applicability domain of each PM to determine the optimum level of each independent variable as well as for the selection of the optimal solution(s).

First, the predicted values for each property were used to fit a model containing all the independent variables (C-001, C-037, and H-046) applied in modeling the original properties (*An*, *Aa* and *U*). So, for the *An* and *U* properties, the original values of

Table 3. Computed ACF Descriptors (C-001, C-037, and H-046), Measured and Predicted Values for the Analgesic (*An*) and Antiinflammatory (*Aa*) Activities, Plus the Ulcerogenic Index (*U*) of the Training Set Compounds.

3-(3-Methylphenyl)-2-substituted amino-3*H*-quinazoline-4-one



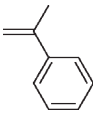
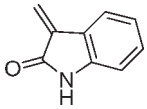
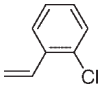
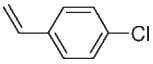
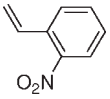
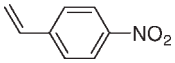
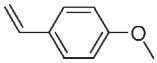
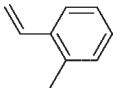
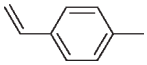
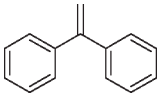
Compound	R	C-001	C-037	H-046	An_{meas} (%)	Aa_{meas} (%)	U_{meas} (%)	An_{pred} (%)	Aa_{pred} (%)	U_{pred} (%)
AS1		3	0	6	76	59	0.53	77	62	0.55
AS2		3	0	9	79	68	0.59	77	67	0.55
AS3		3	0	8	78	69	0.56	77	65	0.55
AS4		1	0	9	59	56	0.60	60	56	0.66
AS5		2	0	3	68	55	0.63	68	52	0.61
AS6		1	0	3	60	45	0.65	60	47	0.66
AS7		1	1	3	58	50	0.69	53	47	0.80
AS8		1	1	3	50	43	0.89	53	47	0.80
AS9		1	1	3	53	47	0.83	53	47	0.80
AS10		1	1	3	58	46	0.85	53	47	0.80
AS11		1	1	3	52	48	0.82	53	47	0.80
AS12		1	1	3	53	47	0.80	53	47	0.80
AS13		2	1	6	58	53	0.69	62	57	0.74
AS14		2	1	6	60	53	0.71	62	57	0.74
AS15		1	0	3	59	49	0.68	60	47	0.66

Table 4. Leverages (h) and Standardized Residuals (Std. Res.) for the Analgesic (An) and Antiinflammatory (Aa) Activities, Plus the Ulcerogenic Index (U) Prediction Models.

Compound	$h(An)$	Std. Res. (An)	$h(Aa)$	Std. Res. (Aa)	$h(U)$	Std. Res. (U)
AS1	0.276	−0.29	0.317	−1.11	0.276	−0.36
AS2	0.276	0.85	0.328	0.51	0.276	0.75
AS3	0.276	0.47	0.279	1.38	0.276	0.19
AS4	0.276	−0.42	0.776	0.17	0.276	−1.14
AS5	0.143	−0.16	0.226	0.99	0.143	0.45
AS6	0.276	−0.04	0.112	−0.58	0.276	−0.22
AS7	0.133	1.84	0.112	1.18	0.133	−2.02
AS8	0.133	−1.21	0.112	−1.29	0.133	1.68
AS9	0.133	−0.06	0.112	0.12	0.133	0.57
AS10	0.133	1.84	0.112	−0.23	0.133	0.94
AS11	0.133	−0.45	0.112	0.47	0.133	0.39
AS12	0.133	−0.06	0.112	0.12	0.133	0.02
AS13	0.200	−1.34	0.089	−1.27	0.200	−0.98
AS14	0.200	−0.57	0.089	−1.27	0.200	−0.61
AS15	0.276	−0.42	0.112	0.82	0.276	0.34

C-001 and C-037 were used (H-046 values were set to zero), and for Aa, the original values of C-001 and H-046 (C-037 values were set to zero). In so doing, one is able to discriminate opposite objectives like efficacy (analgesic and antiinflammatory activities) and toxicity (ulcerogenic ability) with total or partial overlap of the descriptors set used to built the PMs (Notice that the An and U models both contain the C-001 and C-037 descriptors, and the An, Aa, and U models share a common descriptor, i.e. C-001; see Table 2.). Once the model has been set up, the desirability functions for each property (d_i's) might be specified. In order to obtain candidate(s) with high analgesic and antiinflammatory activities as well as low ulcerogenic index, An and Aa should be maximized [eq. (3)] and U minimized [eq. (4)]. In addition, the individual d_i values for the An, Aa, and U properties were determined by setting the L_i, U_i and T_i values as referred previously. Then, the three d_is were combined into the single overall desirability D by means of eq. (1).

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^2_D and Adj.R^2_D values (∼1). Moreover, the high Q^2_D value (0.905) 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 ACF descriptors that simultaneously produce the most desirable combination of all properties. Figure 4 shows the multiple response overall desirability, as well as the individual desirability functions determined by the respective pairs of predictor variables included on the three MLR models.

By inspecting the form of each individual desirability function, it is possible to know the influence of a certain variable over each individual objective. In so doing, one can conclude that C-001 has a significant influence over the three properties, while H-046 has only a remarkable influence on the Aa activity. Here, one should note that the form of the An individual desirability function is similar to that obtained for the Aa activity (for these noncompeting objectives, both curves show a positive slope). However, opposite individual desirability function forms were obtained for competing objectives like Aa and U (i.e. the curve related to the ulcerogenic index has a negative slope).

Moreover, the data reveal that a 3-(3-methylphenyl)-2-substituted amino-3*H*-quinazolin-4-one optimized candidate must have analgesic and anti-inflammatory activities of 93.43% and 82.04%, respectively, plus an ulcerogenic index of 0.44. This represents an overall desirability of 0.8; that can be attained if the candidate has C-001, C-037 and H-046 values equal to 5, 0, and 12, respectively (see Fig. 4), being C-001 the most influencing variable. The significant slope of the C-001 curve suggests

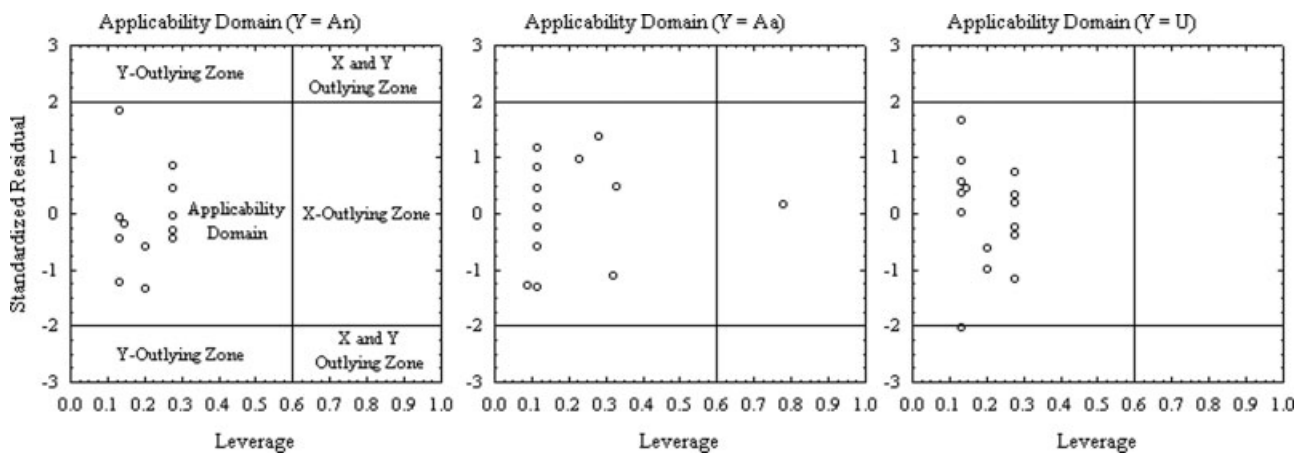


Figure 3. Leverage plots based on the three MLR models; i.e. plots of the standardized residuals.vs. leverage values for the training compounds, with a warning leverage of 0.6.

Table 5. Expected and Predicted Values for the Desirability Due to the Analgesic Activity [$d(An)$], Antiinflammatory Activity [$d(Aa)$], Ulcerogenic Index [$d(U)$], and Overall Desirability [$D(An\text{-}Aa\text{-}U)$].

Compound	$d(An)$	$d(An)_{pred}$	$d(Aa)$	$d(Aa)_{pred}$	$d(U)$	$d(U)_{pred}$	$D(An\text{-}Aa\text{-}U)$	$D(An\text{-}Aa\text{-}U)_{pred}$
AS1	0.68	0.69	0.45	0.49	0.69	0.68	0.60	0.62
AS2	0.72	0.69	0.57	0.56	0.66	0.68	0.65	0.64
AS3	0.71	0.69	0.59	0.53	0.68	0.68	0.65	0.63
AS4	0.45	0.47	0.41	0.41	0.65	0.62	0.50	0.49
AS5	0.57	0.57	0.40	0.36	0.64	0.65	0.53	0.51
AS6	0.47	0.47	0.27	0.29	0.62	0.62	0.43	0.44
AS7	0.44	0.37	0.33	0.29	0.60	0.54	0.45	0.39
AS8	0.33	0.37	0.24	0.29	0.49	0.54	0.34	0.39
AS9	0.37	0.37	0.29	0.29	0.52	0.54	0.38	0.39
AS10	0.44	0.37	0.28	0.29	0.51	0.54	0.40	0.39
AS11	0.36	0.37	0.31	0.29	0.53	0.54	0.39	0.39
AS12	0.37	0.37	0.29	0.29	0.54	0.54	0.39	0.39
AS13	0.44	0.49	0.37	0.43	0.60	0.57	0.46	0.49
AS14	0.47	0.49	0.37	0.43	0.59	0.57	0.47	0.49
AS15	0.45	0.47	0.32	0.29	0.61	0.62	0.44	0.44
Overall desirability function [$D(An\text{-}Aa\text{-}U)$] statistics[a]				$R^2_{D(An\text{-}Aa\text{-}U)} = 0.934$			Adj.$R^2_{D(An\text{-}Aa\text{-}U)} = 0.929$	

[a]Statistical quality of the overall desirability function estimated by the overall desirability determination coefficient (R^2_D) and the adjusted determination coefficient (Adj. R^2_D).

that more attractive candidates could be designed if its values are greater than 5. However, due to the high influence of C-001 over the overall desirability, the optimal range for this variable should be close to 5. But one must also consider the applicability domain of the original PMs. In fact, the training set show C-001 values up to 3 and thus, if the new candidate has a C-001 value extremely far from 3, it might be out of the applicability domain of the original PMs. On the other hand, as the shape of the H-046 desirability function reveals no significant influence (slope near zero), the overall desirability could be increased by large departures from its optimum value (=12). But again the applicability domain of the original PMs should be taken into account.

Figure 5 shows the contour plots of the overall desirability D for two independent variables with the third one kept fixed at its optimum value. An analysis of the plot pertaining to C-037 vs. H-046, allow us to conclude that when C-001 is held at its optimum value, the range of desirability is narrow ($0.62 \leq D \leq 0.78$). This confirms the high influence of the variable C-001 over the overall desirability. On the contrary, when C-037 or H-046 are held at their optimum values, the resultant desirability range is wider ($0.40 \leq D \leq 0.80$).

Table 6. Leave-One-Out Cross-Validation (LOO-CV) Results.

Compound	An_{pred}	Aa_{pred}	U_{pred}	$d(An)_{pred}$	$d(Aa)_{pred}$	$d(U)_{pred}$	$D(An\text{-}Aa\text{-}U)_{pred}$
AS1	77	64	0.56	0.69	0.52	0.69	0.63
AS2	76	66	0.53	0.68	0.55	0.68	0.63
AS3	76	64	0.55	0.68	0.52	0.68	0.62
AS4	61	54	0.69	0.48	0.39	0.60	0.48
AS5	69	51	0.60	0.59	0.35	0.65	0.51
AS6	60	47	0.67	0.47	0.29	0.61	0.44
AS7	52	46	0.82	0.36	0.28	0.53	0.38
AS8	54	47	0.79	0.39	0.29	0.54	0.39
AS9	53	47	0.79	0.37	0.29	0.54	0.39
AS10	52	47	0.79	0.36	0.29	0.54	0.39
AS11	53	46	0.80	0.37	0.28	0.54	0.38
AS12	53	47	0.80	0.37	0.29	0.54	0.39
AS13	62	57	0.76	0.49	0.43	0.56	0.49
AS14	62	57	0.75	0.49	0.43	0.57	0.49
AS15	61	46	0.65	0.48	0.28	0.62	0.44
Overall desirability's LOO-CV determination coefficient				$Q^2_{D(An\text{-}Aa\text{-}U)} = 0.905$			

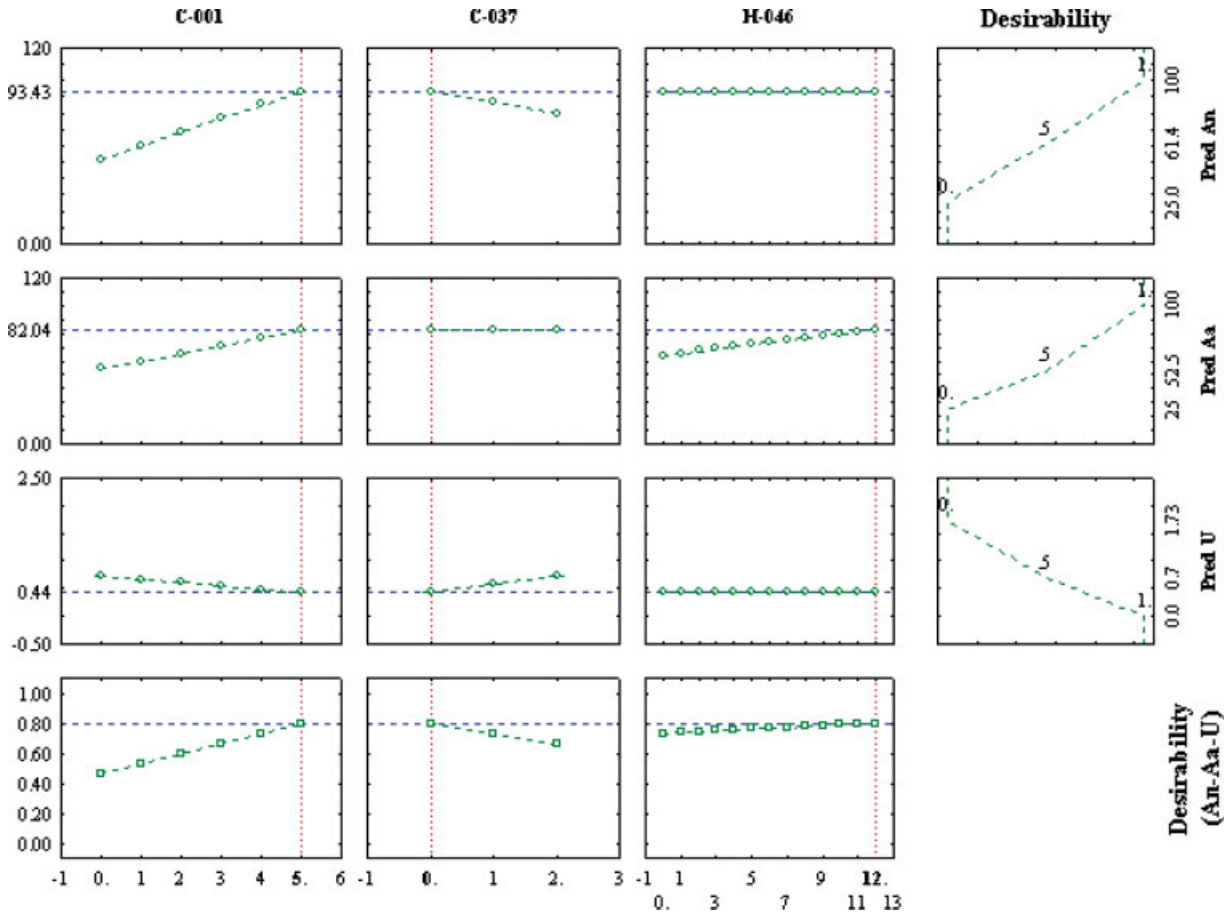


Figure 4. Multiple response desirability function due to the analgesic activity, anti-inflammatory activity and ulcerogenic index (*D*(*An-Aa-U*) (last row), along with the individual desirability functions coming from the pairs of predictor variables included on the three MLR models (first three rows). [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

Design of New Drug Candidates

According to the previous results, the most important variable was found to be descriptor C-001 and the second one descriptor C-037. These two ACF descriptors represent, respectively, the number of methyl groups and heteroatoms attached to a sp_2 carbon atom linked to the aromatic side ring in the drug candidates (see Figure 1). On the other hand, the less influencing ACF descriptor, H-046, represents the number of hydrogen atoms attached to a sp_3 carbon no heteroatom attached to another carbon (see Figure 1).

This information allows one to guess the most important chemical modifications needed to improve the overall desirability of the present compounds. Considering the positive/negative influence of C-001/C-037 a different number vs. type of alkyl groups on the C-2 position of the quinazoline ring should be introduced. In fact, the introduction of branched alkyl substituents might lead to a positive role due to the bulkiness of the substituents.

So, a new set of nine compounds was designed in which several different alkyl substituents were linked to the C-2 position of the quinazoline ring. The chemical modifications and the predicted values of the expected pharmaceutical properties are shown in Table 7. The leverage values obtained for each new designed candidate were also considered to check whether or not each new candidate falls within the applicability domain of the original PMs (see Table 7).

After a comprehensive data analysis, compound **ASNEW8** can be claimed to be the most desirable and reliable candidate designed in this study, displaying predicted percentages of analgesic and antiinflammatory activities of 93 and 82, respectively, plus a predicted ulcerogenic index of 0.44. Further, an excellent predicted overall desirability (0.8) is obtained. The data acquired allow us to propose also compounds **ASNEW4**, **ASNEW5**, **ASNEW6**, and **ASNEW9**, though having leverage values higher than h^*, i.e. out of the applicability domain of the original PMs. Interestingly, they possess the highest overall desirability and predictor variables values, significantly separated from those of the training compounds (see Table 8).

A noticeable profile improvement can be observed between the predicted properties displayed by compound **ASNEW8** and the most promising compound reported by Alagarsamy et al. (**AS3**).[35] Explicitly, **ASNEW8** displays analgesic and antiinflam-

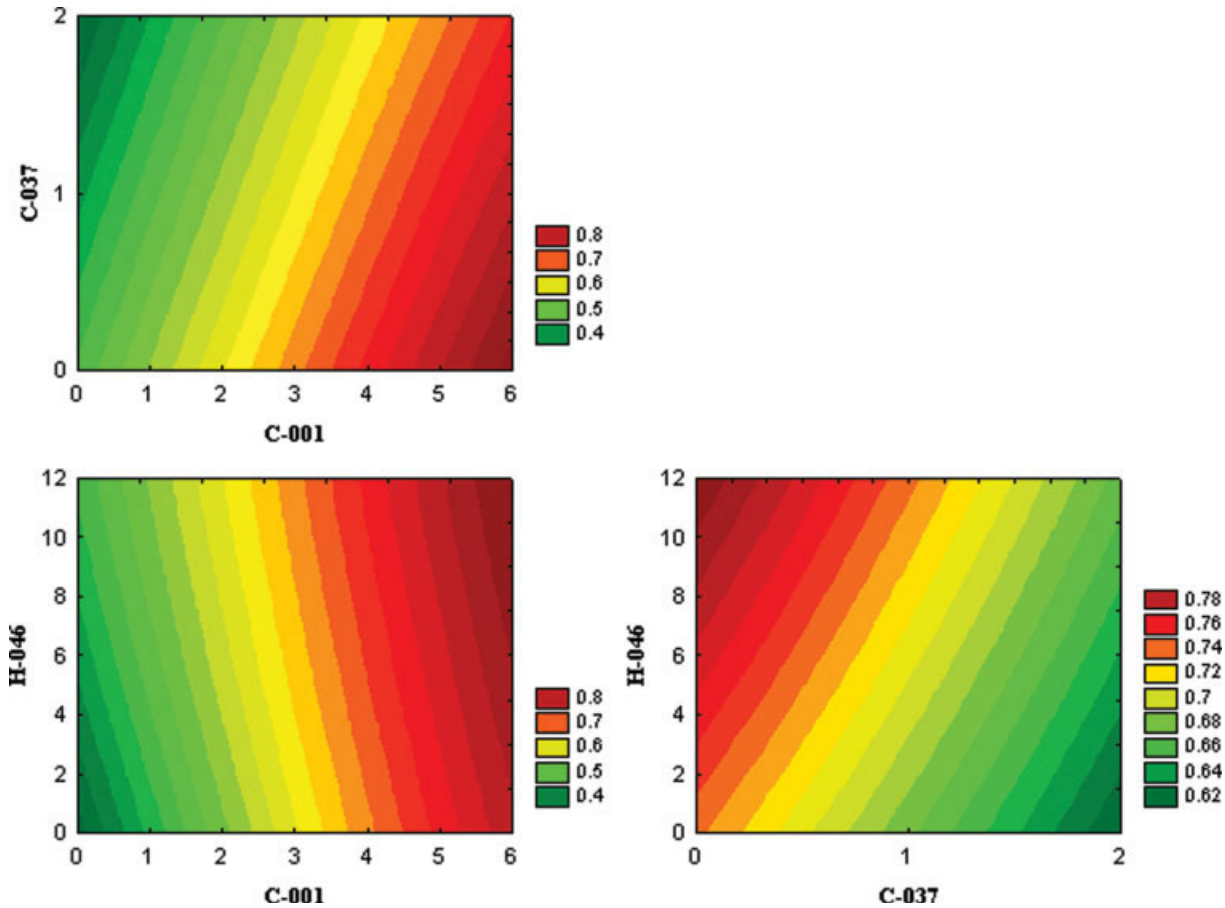


Figure 5. Contour plots of the overall desirability for the analgesic activity, antiinflammatory activity and ulcerogenic index *D*(*An-Aa-U*). Red corresponds to high zones (near to 1) of *D* and green to low zones (near to zero).

matory activities 15 and 13% higher, respectively. At the same time, **ASNEW8** shows only the 78.6% of the ulcerogenic ability of **AS3**. On the other hand, if we compare the performance of **ASNEW8** with diclofenac (a known NSAIDs used as reference compound[35]), one can easily notice its enhanced predicted pharmaceutical properties. In effect, **ASNEW8** displays analgesic and antiinflammatory activities 31% and 22% higher than diclofenac, respectively. In addition, the ulcerogenic index is extensively reduced (**ASNEW8** has almost a quarter (3.75 times lower) of the ulcerogenic ability of diclofenac).

In summary, a remarkable simultaneous improvement on the analgesic and antiinflammatory activities plus ulcerogenic profile

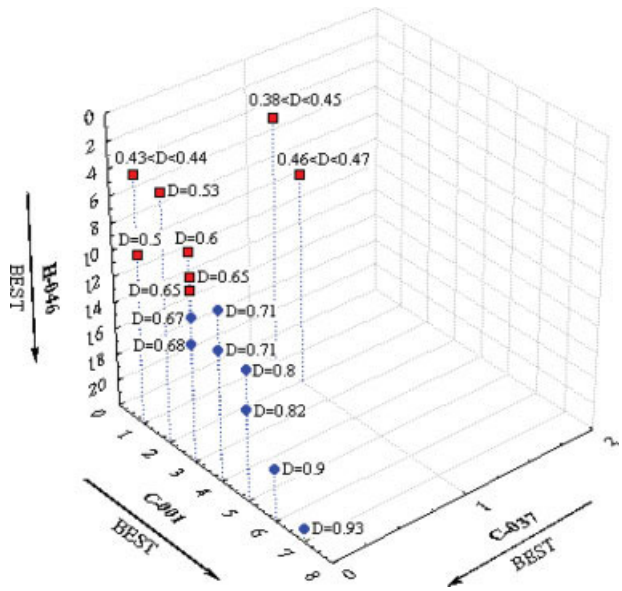
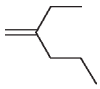
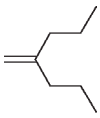
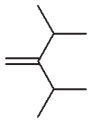
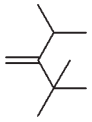
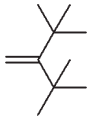


Figure 6. Pareto front of solutions directly optimized over the independent variables domain showing the corresponding overall desirability *D*(*An-Aa-U*) values for each compound. Training compounds are depicted in red squares and new designed compounds in blue dots.

Table 7. Computed ACF Descriptors (C-001, C-037, and H-046), Predicted and Leverage Values for the Analgesic (*An*) and Antiinflammatory (*Aa*) activities, Plus the Ulcerogenic Index (*U*) of the Nine New Designed Compounds.

Compound	R	C-001	C-037	H-046	An_{pred} (%)	Aa_{pred} (%)	U_{pred} (%)	*h*(*An*)	*h*(*Aa*)	*h*(*U*)
ASNEW1		3	0	11	77	70	0.55	0.216	0.361	0.216
ASNEW2		3	0	13	77	72	0.55	0.216	0.496	0.216
ASNEW3		4	0	12	85	77	0.49	0.403	0.453	0.403
ASNEW4[a]		5	0	15	93	86	0.44	0.573	**0.614**	0.573
ASNEW5[a]		6	0	18	102	96	0.38	**0.695**	**0.724**	**0.695**
ASNEW6[a]		7	0	21	110	106	0.33	**0.777**	**0.796**	**0.777**
ASNEW7		4	0	9	85	72	0.49	0.403	0.401	0.403
ASNEW8		5	0	12	93	82	0.44	0.573	0.562	0.573
ASNEW9[a]		5	0	15	93	86	0.44	0.573	**0.614**	0.573

[a]Compounds out of the predictions model's applicability domain; leverage values greater than *h** are marked in bold.

of the new designed candidates was obtained through MOOP-DESIRE-based methods combined with human expert interpretation and use of the results. The data suggest a positive role of the bulkiness of the alkyl substituents on the C-2 position of the quinazoline ring on the ulcerogenic properties. Anyhow, in the future, an experimental study of the analgesic, antiinflammatory

Table 8. Predicted Values for the Desirability Due to the Analgesic Activity [$d(An)$], Antiinflammatory Activity [$d(Aa)$], Ulcerogenic Index [d(U)], and Overall Desirability [$D(An\text{-}Aa\text{-}U)$] of the Nine New Designed Compounds.

Compound	$d(An)_{pred}$	$d(Aa)_{pred}$	$d(U)_{pred}$	$D(An\text{-}Aa\text{-}U)_{pred}$
ASNEW1	0.69	0.60	0.73	0.67
ASNEW2	0.69	0.63	0.73	0.68
ASNEW3	0.80	0.69	0.65	0.71
ASNEW4[a]	0.91	0.81	0.75	0.82
ASNEW5[a]	1.00	0.95	0.78	0.90
ASNEW6[a]	1.00	1.00	0.81	0.93
ASNEW7	0.80	0.63	0.72	0.71
ASNEW8	0.91	0.76	0.75	0.80
ASNEW9[a]	0.91	0.81	0.75	0.82

[a]Compounds out of the predictions model's applicability domain.

and ulcerogenic properties of the designed candidates should be carried out to validate the process.

Despite the limited size and homogeneity of our data set, this work offers the possibility of a deeper and case by case analysis of the results obtained by using the MOOP-DESIRE methodology. The use of small and homogeneous data set is more suitable for later stages of the drug development process once identified a lead rather than for early stages. Actually, the results of the optimization process can be used to perform specific structural modifications over the lead. For this, the use of clearly defined structural or physicochemical descriptors can lead to interpretable structure–desirability relationships which can be used to design new candidates with an improved pharmaceutical profile.

The MOOP-DESIRE methodology can also be applied to handle larger and/or more diverse data sets, such as those frequently obtained in High-Throughput Screening processes, being there more appropriate for early stages of the drug development process. That is, molecules coming from large and heterogeneous data sets can be filtered and ranked according to a certain criterion rather than applying the results of the optimization process to design new candidates. To accomplish that, one can resort to the overall desirability of each molecule as a ranking criterion or to several distance measures between the optimal values of the descriptors determined by MOOP-DESIRE and the computed values of the descriptors. In this case, it is advisable to use descriptors leading to highly predictive structure–desirability relationships rather than interpretable descriptors in order to ensure the accuracy of the predictions and therefore, an accurate assessment of the molecule's overall desirability which will then be the ranking criterion.

Comparison with Other MOOP Approaches

Finally, some considerations can be drawn about the desirability-based MOOP method proposed here and the presently most used MOOP methods. The desirability-based MOOP method, like the WSOF-based MOOP methods, (re)formulates a multiobjective problem into a single one (the overall desirability). The rationale is to find a single "*best*" solution overlooking however the presence of the paretofront of the objectives, which represents the main drawback of both methods when compared with pareto-based methods.

As the single "best" solution is directly found over the independent variables domain, one can effectively generalize to other solutions (candidates) with similar or improved compromise between the k objectives. It is worth noting that the "best" solution depends on the independent variables used to fit the PMs for each objective. So, in one run, the method will retrieve only one "best" solution. To obtain more information and other solutions, it must be run several times with different selections of independent variables and/or different weightings on the overall desirability formula. Thus, the desirability-based MOOP method can be placed somewhere between the WSOF- and pareto-based MOOP methods.

Actually, the major drawback of WSOF-based methods is the selection of the most appropriate weightings because it is often not clear how the different objectives should be ranked. In addition, the method is limited in its ability to find solutions to problems involving competing objectives.[22] But the MOOP-DESIRE method has the advantage of transforming the responses (objectives) to desirability d_i values, which are then combined into the single overall desirability D. So, competing objectives like potency and toxicity can be successfully handled by this method because the use of weights is avoided in the multi- to single objective problem reformulation. Furthermore, by changing the s and t parameters on the establishment of the individual d_i's [see eqs. (2–4)], one can nevertheless alter the objectives' weightings, if one has prior preferences or knowledge of the objectives importance.[34]

As regards pareto-based methods, although they are important for the simultaneous optimization of multiple objectives they still have some limitations. Specifically, the pareto-front may be vast, particularly in circumstances with large numbers of objectives.[22] One should remark here that in the presently proposed desirability-based MOOP method, the single "best" solution is achieved directly over the independent variables domain, making the solution independent from the number of objectives to optimize. Moreover, by analyzing the profile and contour plots of the overall desirability D (i.e. by looking at their shapes and slopes), one is able to establish the best departures from the X optimum values to further increase D. The optimum range of independent variables, in an analogy with pareto-based methods, can work as a pareto-front of independent variables leading to a set of optimal (desirable) solutions (candidates) which are ranked according to the overall desirability. Figure 6 shows such kind of pareto-front of solutions directly optimized over the independent variables domain. These solutions were obtained by interrelating the 15 training molecules, which were used to fit the desirability functions, and the nine new designed molecules. The approximated region of the best pareto-front solutions can be found at values ranging from 4 to 6, from 0 to 1, and from 8 to 14 for the predictor variables C-001, C-037, and C-046, respectively.

An additional drawback of the pareto-based methods is that the distribution of the pareto-front may lead solutions to drift to more densely distributed regions of the surface and, in more extreme circumstances, lead to dictatorship conditions where a single objective dominates.[22] The use of the overall desirability

values in the present MOOP method avoids this problem since they provide the overall assessment of the combined response (objective) levels.

Finally, the main drawback of the proposed MOOP-DESIRE method is related to the modeling technique used to fit the initial set of PMs. Since the optimization process over the independent variables domain is based on a MLR approach, neither the predicted responses nor the optimum levels of each independent variable that determines the predicted overall desirability will be reliable if the parametric assumptions inherent to regression techniques are not fulfilled.[56,57] Specifically, the effect of potential nonlinear relations between descriptors and objectives could lead to very poor predictions and consequently to very unreliable structure–desirability relationships. The combination of nonlinear modeling techniques such as machine learning algorithms with optimization methods can be a solution to this bottleneck on the application of desirability based-MOOP methods.

Conclusions

In this work, a novel MOOP method sustained on the desirability estimation of several interrelated responses is proposed. The MOOP-DESIRE methodology based on Derringer's desirability function enables one to perform global QSAR studies, considering simultaneously the pharmacological, pharmacokinetic and toxicological profiles of a set of molecule candidates. The usefulness of the methodology, placed between WSOF- and pareto-based MOOP methods, was demonstrated by applying it to the simultaneous optimization of the analgesic, antiinflammatory and ulcerogenic properties of a library of fifteen 3-(3-methylphenyl)-2-substituted amino-3*H*-quinazolin-4-one compounds. The best compromise between the mentioned properties was established and new drug candidates with the highest overall desirability then designed. In particular, one of the designed candidates (compound **ASNEW8**) is predicted to have 93% of analgesic activity, 82% of inflammatory inhibition and an ulcerogenic index of 0.44, which represents an excellent overall desirability (=0.8), being this accomplished by modifying the compounds' structure in such a way that pushed the values of the C-001, C-037, and H-046 predictor variables to 5, 0, and 12, respectively. Furthermore, it was observed that the presence of bulky alkyl substituents at the C-2 position of the quinazoline ring displayed a positive role on the ulcerogenic ability without a negative influence in the other properties. Yet, further experimental corroboration is still needed to validate the model.

In conclusion, the desirability-based MOOP method herein proposed is regarded as a valuable tool and shall aid in the future rational design of novel successful drugs.

References

1. Ekins, S.; Boulanger, B.; Swaan, P. W.; Hupcey, M. A. J Comput Aided Mol Des 2002, 16, 381.
2. Jorgensen, W. L. Science 2004, 303, 1813.
3. Seifert, M. H. J.; Wolf, K.; Vitt, D. Drug Discov Today: Biosilico 2003, 1, 143.
4. Brown, N.; Lewis, R. A. Curr Opin Drug Discov Devel 2006, 9, 419.
5. Hansch, C. J Med Chem 1976, 19, 1.
6. Fukunaga, J. Y.; Hansch, C.; Steller, E. E. J Med Chem 1976, 19, 605.
7. Mayer, J. M.; van de Waterbeemd, H. Environ Health Perspect 1985, 61, 295.
8. Moriguchi, I.; Hirano, H.; Hirono, S. Environ Health Perspect 1996, 104, 1051.
9. Estrada, E. SAR QSAR Environ Res 2000, 11, 55.
10. Vilar, S.; Estrada, E.; Uriarte, E.; Santana, L.; Gutierrez, Y. J Chem Inf Model 2005, 45, 502.
11. 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 (Online) 2006, 12, 255.
12. Helguera, A. M.; Cabrera Perez, M. A.; Gonzalez, M. P. J Mol Model (Online) 2006, 12, 769.
13. Gonzalez-Diaz, H.; Cruz-Monteagudo, M.; Molina, R.; Tenorio, E.; Uriarte, E. Bioorg Med Chem 2005, 13, 1119.
14. Gonzalez-Diaz, H.; Cruz-Monteagudo, M.; Vina, D.; Santana, L.; Uriarte, E.; De Clercq, E. Bioorg Med Chem Lett 2005, 15, 1651.
15. Cruz-Monteagudo, M.; Gonzalez-Diaz, H.; Borges, F.; Gonzalez-Diaz, Y. Bull Math Biol 2006, 68, 1555.
16. Cruz-Monteagudo, M.; Cordeiro, M. N.; Borges, F. J Comput Chem 2008, 29, 533.
17. Cruz-Monteagudo, M.; Borges, F.; Perez Gonzalez, M.; Cordeiro, M. N. Bioorg Med Chem 2007, 15, 5322.
18. 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, 1909.
19. Gonzalez-Diaz, H.; Aguero, G.; Cabrera, M. A.; Molina, R.; Santana, L.; Uriarte, E.; Delogu, G.; Castanedo, N. Bioorg Med Chem Lett 2005, 15, 551.
20. Prado-Prado, F. J.; Gonzalez-Diaz, H.; Santana, L.; Uriarte, E. Bioorg Med Chem 2007, 15, 897.
21. Gonzalez-Diaz, H.; Prado-Prado, F. J.; Santana, L.; Uriarte, E. Bioorg Med Chem 2006, 14, 5973.
22. Nicolaou, A. C.; Brown, N.; Pattichis, C. S. Curr Opin Drug Discov Devel 2007, 10, 316.
23. Yann, C.; Siarry, P., Eds. Multiobjective Optimization: Principles and Case Studies; Springer-Verlag: Berlin, Germany, 2004.
24. Jones, G.; Willett, P.; Glen, R. C. J Comput Aided Mol Des 1995, 9, 532.
25. Handschuh, S.; Wagener, M.; Gasteier, J. J Chem Inf Comput Sci 1998, 38, 220.
26. Shepphird, J. K.; Clarck, R. D. J Comput Aided Mol Des 2006, 20, 735.
27. Janson, S.; Merkle, D. In Hybrid Metaheuristics Second International Workshop, HM 2005; Blesa, M. J.; Blum, C.; Roli, A.; Sampels, M., Eds.; Springer-Verlag: Barcelona, Spain, 2005, p. 128.
28. Zoete, V.; Grosdidier, A.; Michielin, O. High Performance Computing for the Life Sciences Symposium, Lausanne, Switzerland, 2005.
29. Brown, N.; Mckay, B.; Gasteiger, J. J Comput Aided Mol Des 2006, 20, 333.
30. Lameijer, E. W.; Kok, J. N.; Back, T.; Ijerman, A. P. J Chem Inf Model 2006, 46, 545.
31. Nicolotti, O.; Gillet, V. J.; Fleming, P. J.; Green, D. V. J Med Chem 2002, 45, 5069.
32. Stockfisch, T. P. J Chem Inf Comput Sci 2003, 43, 1608.
33. Rao, S. N.; Stockfisch, T. P. J Chem Inf Comput Sci 2003, 43, 1614.
34. Derringer, G.; Suich, R. J Quality Technol 1980, 12, 214.
35. Alagarsamy, V.; Dhanabal, K.; Parthiban, P.; Anjana, G.; Deepa, G.; Murugesan, B.; Rajkumar, S.; Beevi, A. J. J Pharm Pharmacol 2007, 59, 669.
36. Arulmozhi, D. K.; Veeranjaneyulu, A.; Bodhankar, S. L.; Arora, S. K. J Pharm Pharmacol 2004, 56, 655.

37. Ganguly, A. K.; Bhatnagar, O. P. Can J Physiol Pharmacol 1973, 51, 748.
38. Goyal, R. K.; Chakrabarti, A.; Sanyal, A. K. Planta Med 1985, 29, 85.
39. ChemDrawn Ultra 9.0. CambridgeSoft. 2004.
40. Burkert, U.; Allinger, N. L. Molecular Mechanics; ACS: Washington, D.C., 1982.
41. Clark, T. Computational Chemistry; Wiley: N.Y., 1985.
42. Frank, J. MOPAC 2.0; Seiler Research Laboratory, US Air Force Academy, Colorado Springs: CO, 1993.
43. Todeschini, R.; Consonni, V.; Pavan, M. DRAGON 2.1; Milano Chemometrics: Milano, 2002.
44. Viswanadhan, V. N.; Ghose, A. K.; Revankar, G. R.; Robins, R. K. J Chem Inf Comput Sci 1989, 29, 163.
45. Caballero, J.; Fernandez, L.; Abreu, J. I.; Fernandez, M. J Chem Inf Model 2006, 46, 1255.
46. Fernandez, M.; Caballero, J. J Mol Graph Model 2006, 25, 410.
47. Fernandez, L.; Caballero, J.; Abreu, J. I.; Fernandez, M. Proteins 2007, 67, 834.
48. Caballero, J.; Fernandez, L.; Garriga, M.; Abreu, J. I.; Collina, S.; Fernandez, M. J Mol Graph Model 2007, 26, 166.
49. Leardi, R.; Boggia, R.; Terrile, M. J Chemom 1992, 6, 267.
50. Yasri, A.; Hartsough, D. J Chem Inf Comput Sci 2001, 41, 1218.
51. Hou, T. J.; Wang, J. M.; Liao, N.; Xu, X. J. J Chem Inf Comput Sci 1999, 39, 775.
52. Hasegawa, K.; Kimura, T.; Funatsu, K. J Chem Inf Comput Sci 1999, 39, 112.
53. Barbosa de Oliveira, D.; Gaudio, A. C.; BuildQSAR; University of Espírito Santo: Vitória ES, Brasil, 2000.
54. Barbosa de Oliveira, D.; Gaudio, A. C. Quant Struct-Act Relat 2000, 19, 599.
55. STATISTICA 6.0. Statsoft_Inc., 2001.
56. Eriksson, L.; Jaworska, J.; Worth, A. P.; Cronin, M. T.; McDowell, R. M.; Gramatica, P. Environ Health Perspect 2003, 111, 1361.
57. Stewart, J.; Gill, L. Econometrics; Prentice Hall: London, 1998.
58. Kutner, M. H.; Nachtsheim, C. J.; Neter, J.; Li, W., Eds. In Applied Linear Statistical Models; McGraw Hill: New York, 2005, p 278.
59. Atkinson, A. C. Plots, Transformations and Regression; Clarendon Press: Oxford 1985.
60. Kutner, M. H.; Nachtsheim, C. J.; Neter, J.; Li, W. Applied Linear Statistical Models; McGraw Hill: New York, 2005.
61. Nelder, J. A.; Mead, R. Comput J 1965, 7, 308.
62. Fletcher, R.; Reeves, C. M. Comput J 1964, 7, 149.
63. Hooke, R.; Jeeves, T. A. J Assoc Comp Machin 1961, 8, 212.
64. De Boor, C. A Practical Guide to Splines; Springer-Verlag: New York, 1978.
65. Gerald, C. F.; Wheatley, P. O. Applied Numerical Analysis; Addison Wesley Reading: MA, 1989.

SUPPORTING INFORMATION

DESIRABILITY-BASED MULTI-OBJECTIVE OPTIMIZATION FOR GLOBAL QSAR STUDIES. APPLICATION TO THE DESIGN OF NOVEL NSAIDS WITH IMPROVED ANALGESIC, ANTI-INFLAMMATORY AND ULCEROGENIC PROFILES

Maykel Cruz-Monteagudo, Fernanda Borges, M. Natália D.S. Cordeiro

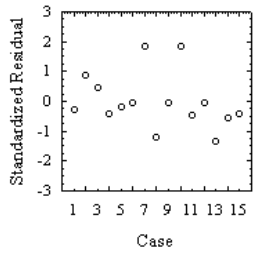
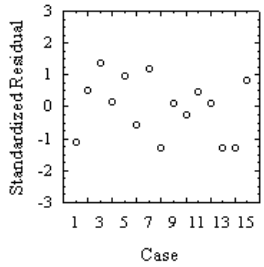
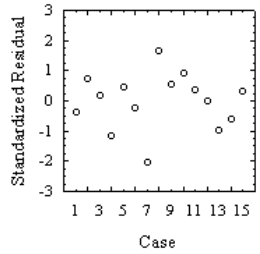
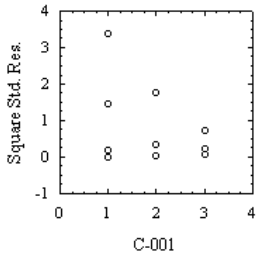
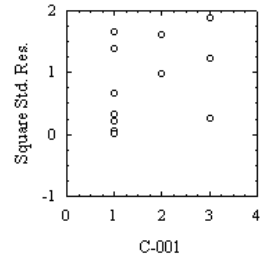
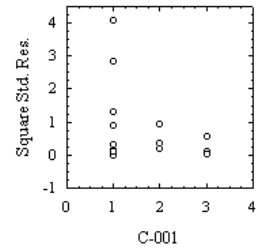
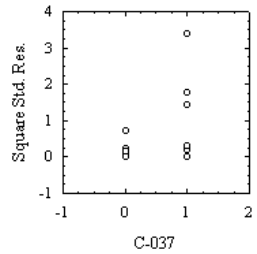
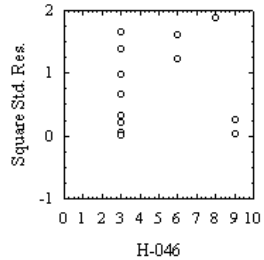
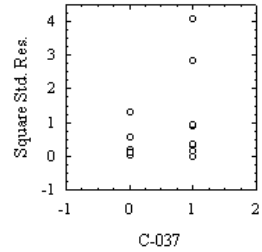
CONTENTS

- Checking the main parametric assumptions related to the three MLR models used to fit the desirability functions.

This section provides details about the checking of the pre-adopted parametric assumptions, a very important aspect in the application of linear multivariate statistical-based approaches (MLR techniques) (1). In fact, once the linear regression model has been set up, it is very important to check the parametric assumptions to assure the validity of extrapolation from the sample to the population. These include the linearity of the modeled property, normal distribution as well as the homoscedasticity and non-multicollinearity descriptors. Notice that severe violations of one or various of these assumptions can markedly compromise the reliability of the predictions resulting from our MLR models (1).

We first check the linearity hypothesis by looking at the distribution of the standardized residuals for all cases. Indeed the plots in Table SMI (1st row) do not show any specific pattern, reinforcing the idea that our models do not exhibit a non-linear dependence (1). Next, we check the hypothesis of homoscedasticity (*i.e.*: homogeneity of variance of the variables), which can be confirmed by simply plotting the square of standardized residuals for each predictor variable (1) (2nd row of plots in Table SMI). These plots reveal significant scatter of points, without any systematic pattern, *post-mortem* validating the pre-adopted assumption of homoscedasticity for all the PMs. They also provide a check for the no auto-correlation of the residuals. Moving on to the hypothesis of normally distributed residuals, one can easily confirm that the residuals follow a normal distribution by applying the Kolmogorov-Smirnov statistical test (3rd row of Table SMI). In addition, as the term related to the error (represented by residuals) is not included in the MLR equations, the mean must be zero what actually occurs (check 4th row of Table SMI). The last aspect deserving special attention is the degree of multicollinearity among the variables. Highly collinear variables may be identified by examining their pair-correlations (R). As can be seen (5th row of Table SMI), the variables included in the models exhibit a low collinearity among them as the Rs are always lower than 0.7. One should emphasize here that the common interpretation of a regression coefficient as measuring the change in the expected value of the response variable, when the given predictor variable is increased by one unit while all other predictor variables are held constant, is not fully applicable when multicollinearity exists ($R \geq 0.7$) (2).

Table SM I. Checking the main parametric assumptions related to the three MLR models used to fit the desirability functions.

	***An* MLR Model**	***Aa* MLR Model**	***U* MLR Model**
Linearity			
Homoscedasticity			
			
Normality of Residuals	K-S d = 0.24899, p > 0.20	K-S d = 0.15142, p > 0.20	K-S d = 0.11477, p > 0.20
Res. Mean = 0	-0.000000	0.000000	-0.000000
Non Multi-Collinearity	*R* (C-001/C-0037) = -0.47; *R* (C-001/H-046) = 0.67; *R* (C-037/H-046) = -0.46		

References

1. Stewart J, Gill L. Econometrics. 2nd edition ed. Allan P, editor. London: Prentice Hall; 1998.
2. Kutner MH, Nachtsheim CJ, Neter J, Li W. Multicollinearity and its effects. Applied Linear Statistical Models. Fifth ed. New York: McGraw Hill; 2005. p. 278-89.

ANNEX II

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$_{(PHARM\text{-}TOX)}$ 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.

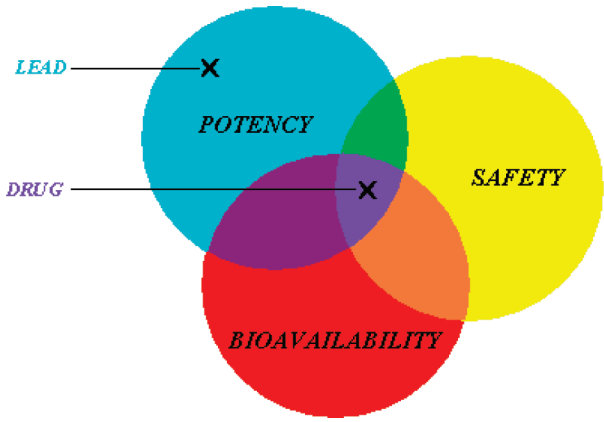


Figure 1. Graphic representation of the compromise between therapeutic efficacy (potency), bioavailability (ADME properties), and toxicity (safety) required to reach a successful drug.

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, 1] 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 3	a_2 6	a_3 2	a_4 4	a_5 5	a_6 8	a_7 1	a_8 7	a_9 10	a_{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

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

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 ($\Delta_{i\mathrm{MIN}}$) and $U_i = \Delta_{i\mathrm{MAX}}$.

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

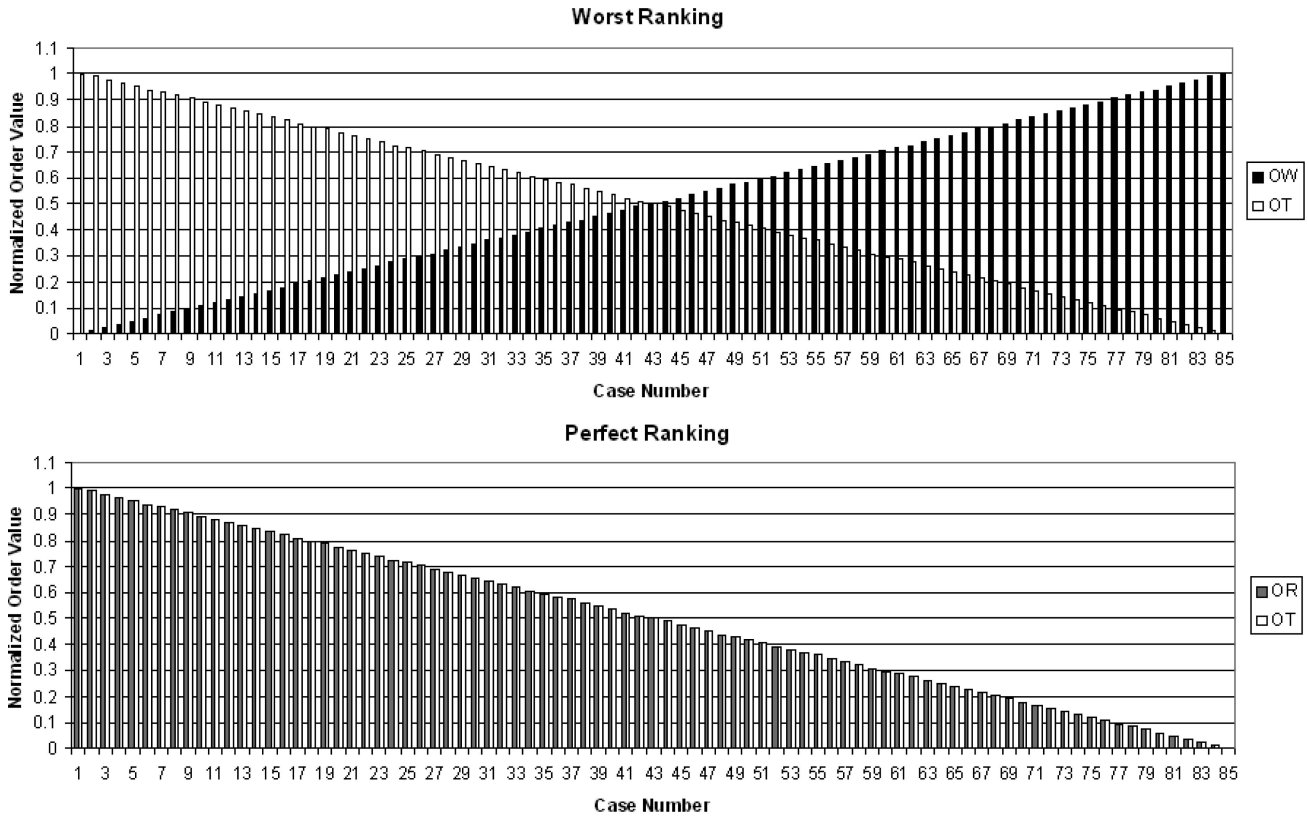


Figure 2. Worst (top) and perfect (bottom) ranking.

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

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

$1/1 + \mathrm{MIC} = 27.127(\pm 3.925) - 1.573(\pm 0.170)\cdot \mathrm{H4M} - 13.504(\pm 1.969)\cdot \mathrm{BELp1} + 0.071(\pm 0.012)\cdot \mathrm{RDF020e} - 0.130(\pm 0.024)\cdot \mathrm{Mor05m} - 0.006(\pm 0.001)\cdot \mathrm{G(F\cdots F)} + 5.670(\pm 1.097)\cdot \mathrm{HATS3m} + 0.002(\pm 0.000)\cdot \mathrm{D/Dr06} - 0.234(\pm 0.064)\cdot \mathrm{Mor14v} + 1.449(\pm 0.423)\cdot \mathrm{HATS3e} + 0.011(\pm 0.003)\cdot \mathrm{RDF050e}$

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 ($\mathrm{IC_{50}} = 1/(1 + \mathrm{IC_{50}})$)

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

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 $O_\mathrm{T} = 1, 2, 3, ..., n$. 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 ith 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

$R_{D(MIC\text{-}IC_{50})}{}^2 = 0.702$ Adj.$R_{D(MIC\text{-}IC_{50})}{}^2 = 0.698$

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

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

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^*)

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

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$_{(PHARM\text{-}TOX)}$-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).

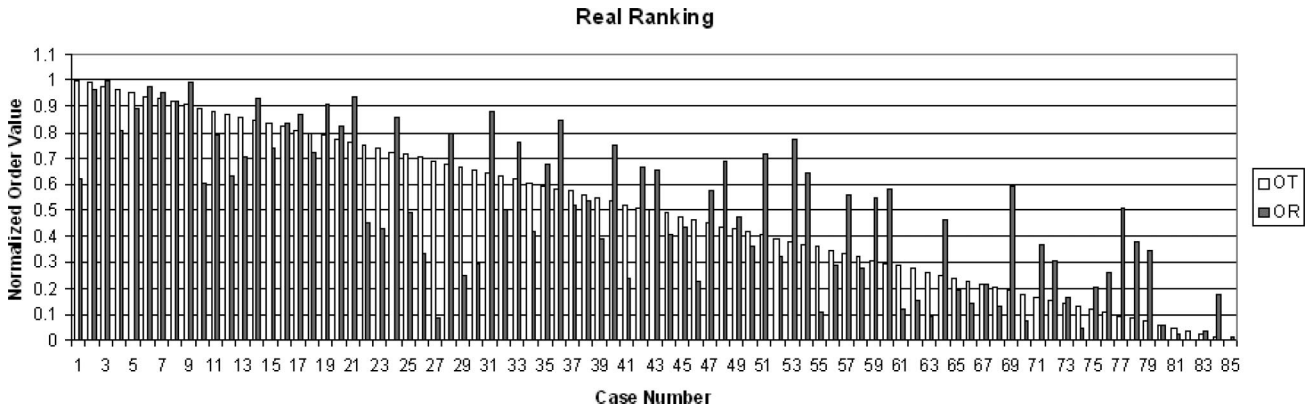


Figure 3. Δ_i-Based ranking of the fluoroquinolone library.

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).

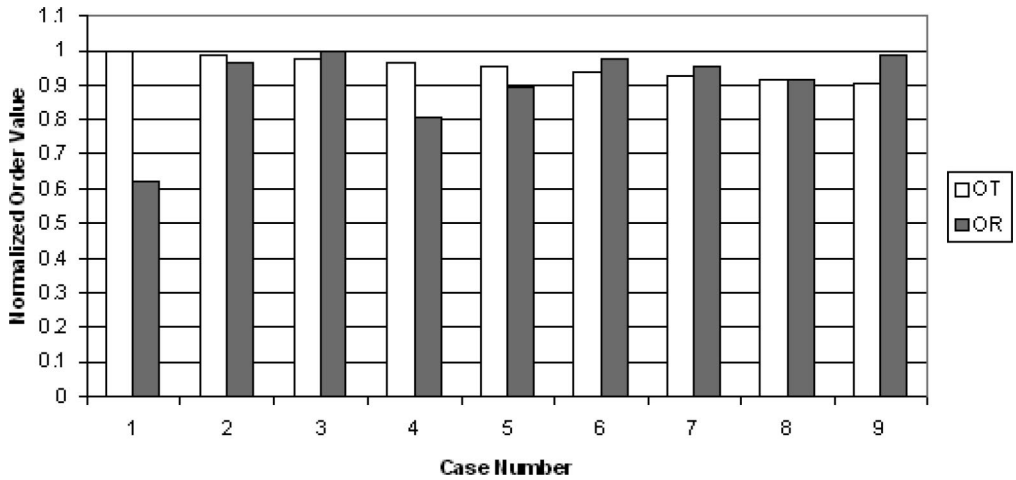


Figure 4. Ranking attained for the 10% of the library of compounds.

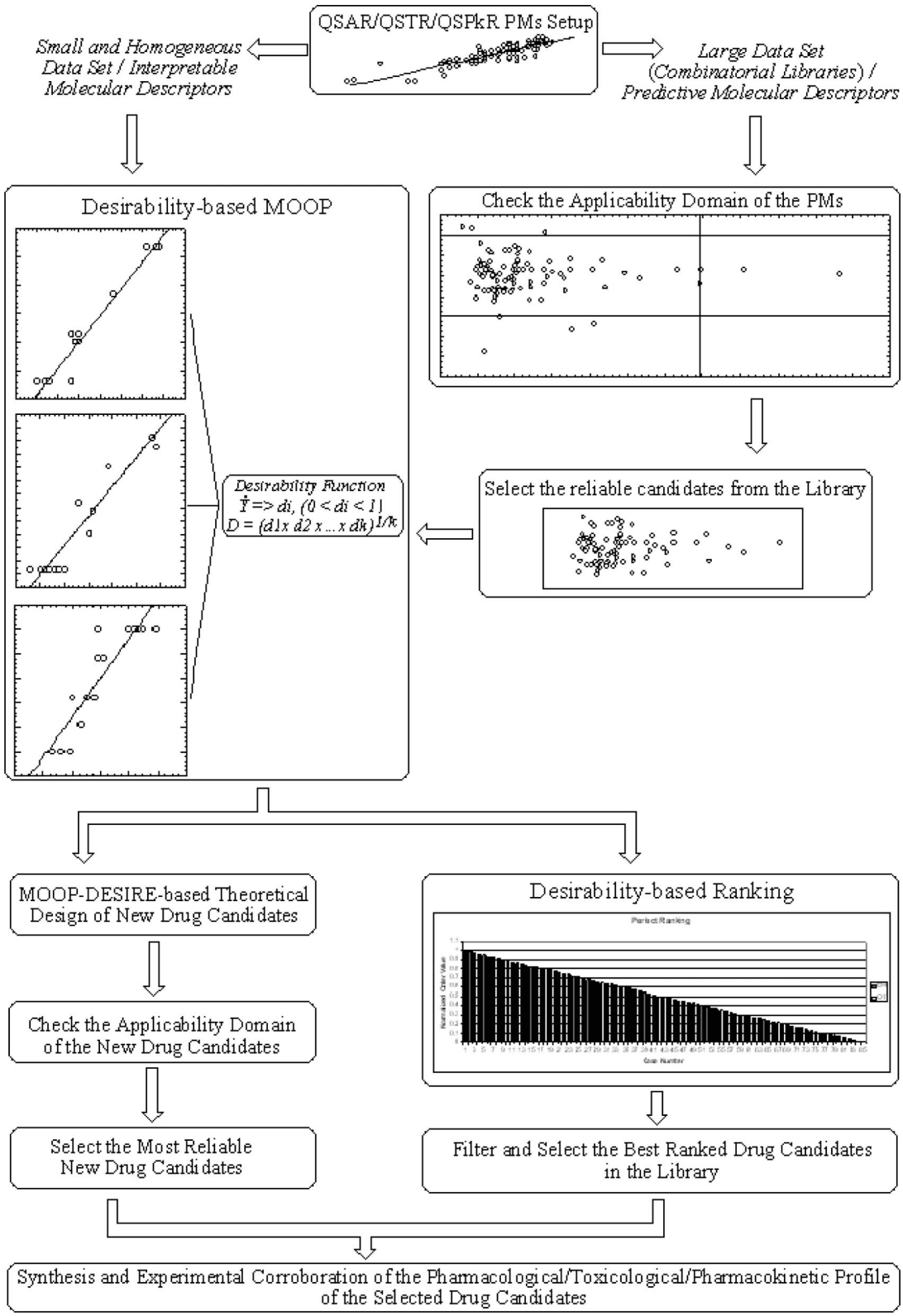


Figure 5. MOOP-DESIRE-based rational drug discovery and development.

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

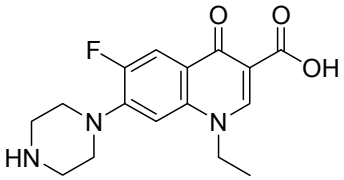
SUPPORTING INFORMATION

DESIRABILITY-BASED METHODS OF MULTI-OBJECTIVE OPTIMIZATION AND RANKING FOR GLOBAL QSAR STUDIES. FILTERING SAFE AND POTENT DRUG CANDIDATES FROM COMBINATORIAL LIBRARIES

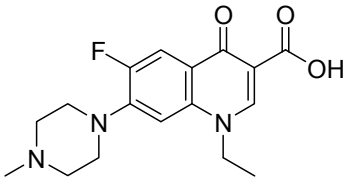
Maykel Cruz-Monteagudo, Fernanda Borges, M. Natália D.S. Cordeiro, J. Luis Cagide Fajin, Carlos Morell, Reinaldo Molina Ruiz, Yudith Cañizares-Carmenate, Elena Rosa Dominguez

CONTENTS

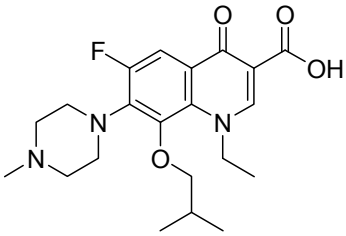
Chemical Structures of the Library of Fluoroquinolones



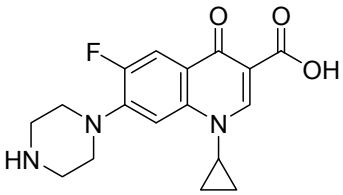
001-1-Norfloxacin



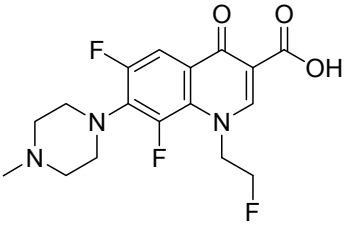
002-2-Pefloxacin



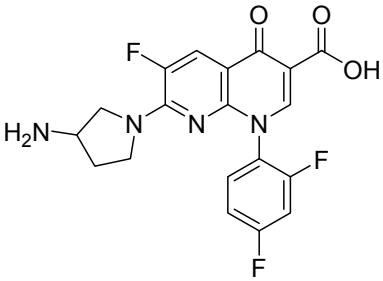
003-3-Ofloxacin



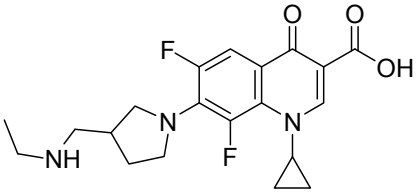
004-4-Ciprofloxacin



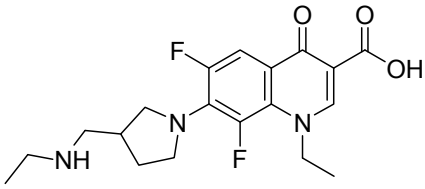
005-5-Fleroxacin



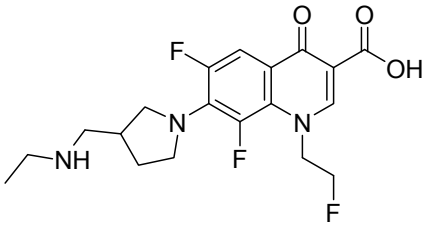
006-6-Tosufloxacin



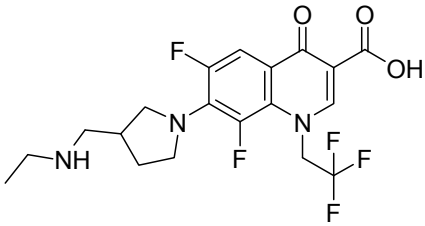
007-7-PD117558



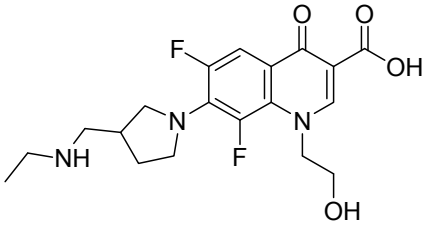
008-8



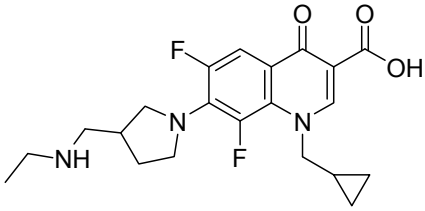
009-9



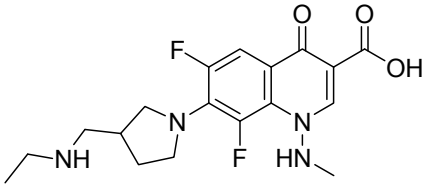
010-10

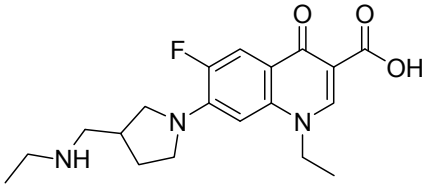


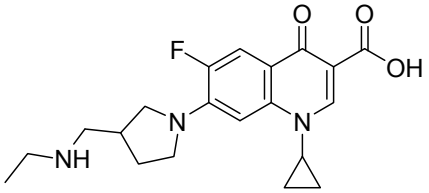
011-11

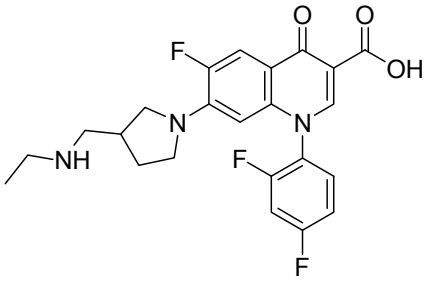


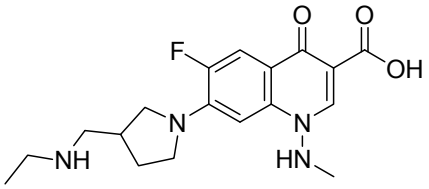
012-13

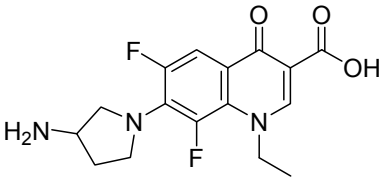

013-14

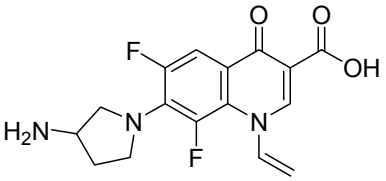

014-15

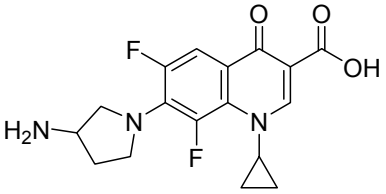

015-16

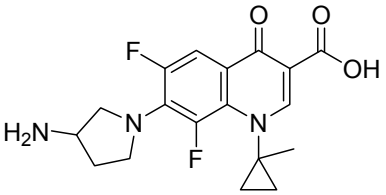

016-17

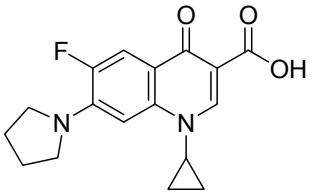

017-18

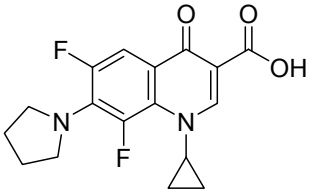

018-19

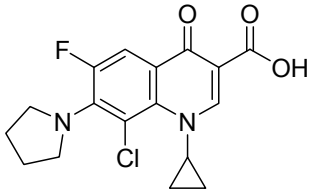

019-20

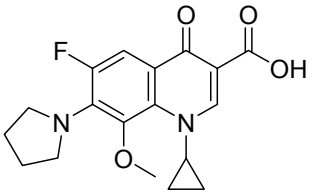

020-21

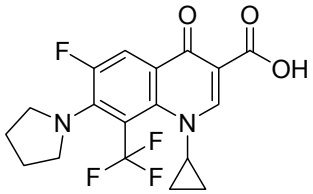

021-22

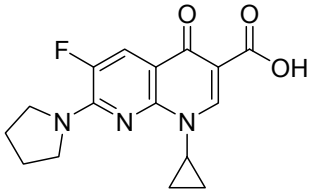

022-23A


023-23B

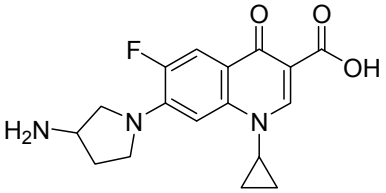

024-23C


025-23D

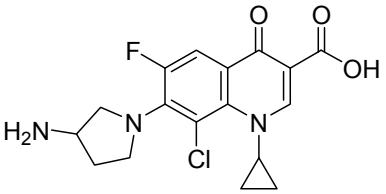

026-23E



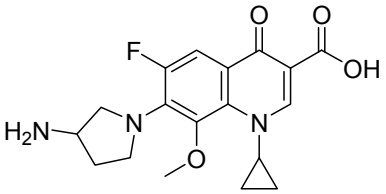
027-23F



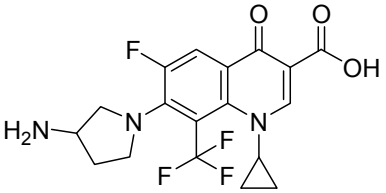
028-24A



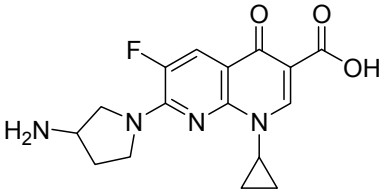
029-24C



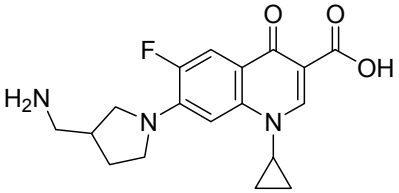
030-24D



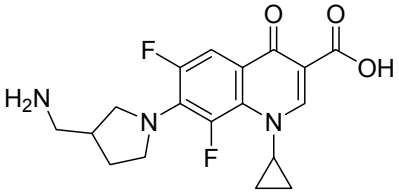
031-24E



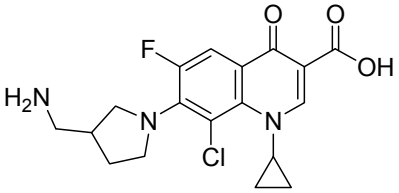
032-24F



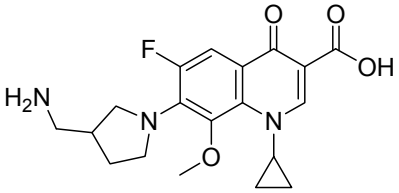
033-25A



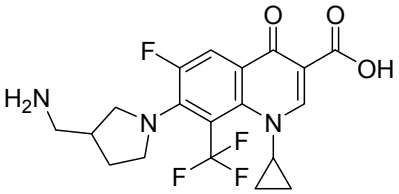
034-25B



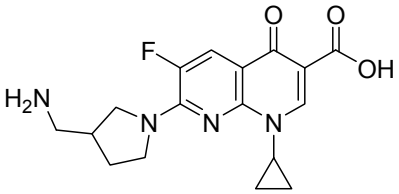
035-25C



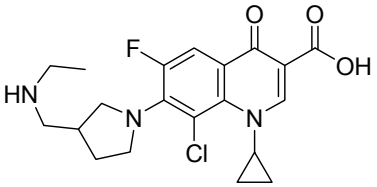
036-25D



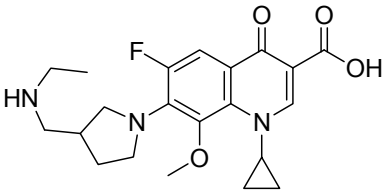
037-25E



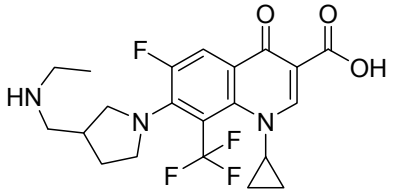
038-25F

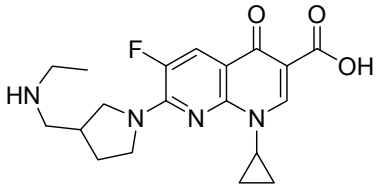


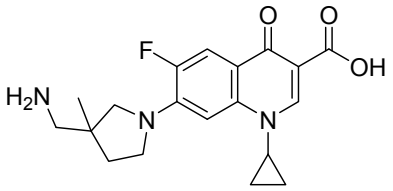
039-26C

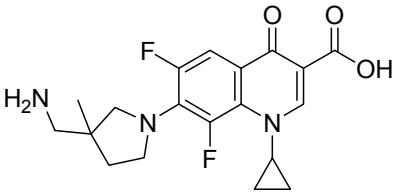


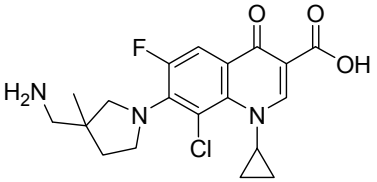
040-26D

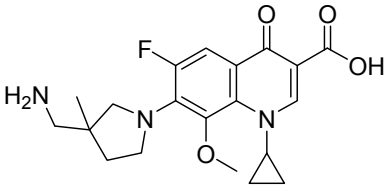

041-26E

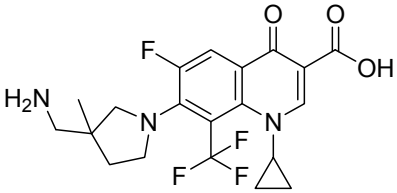

042-26F

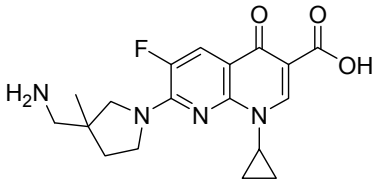

043-27A

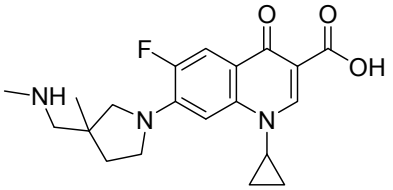

044-27B

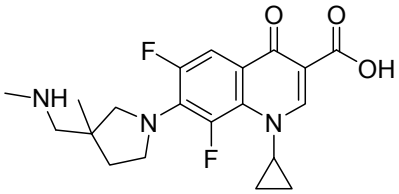

045-27C

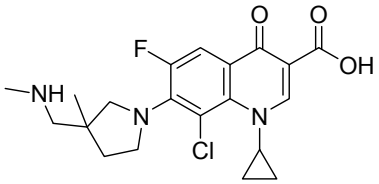

046-27D

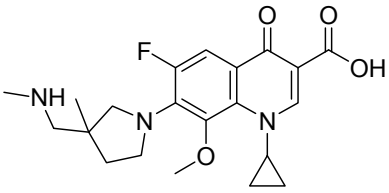

047-27E

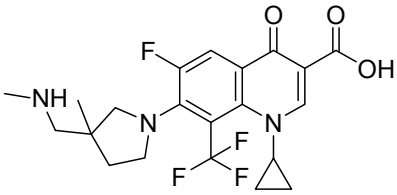

048-27F

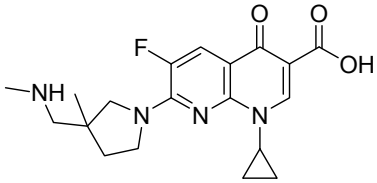

049-28A

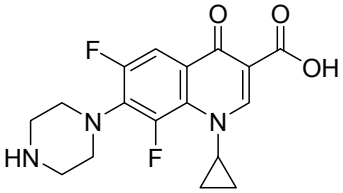

050-28B


051-28C

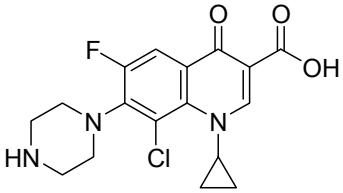

052-28D


053-28E

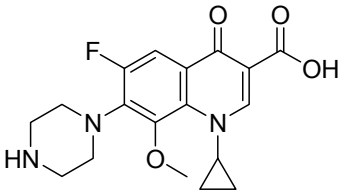

054-28F



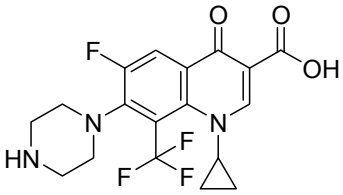
055-29B



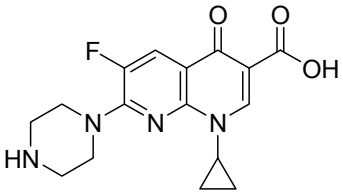
056-29C



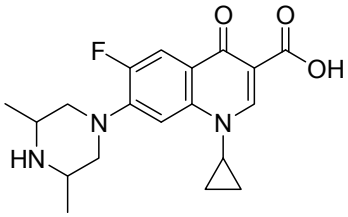
057-29D



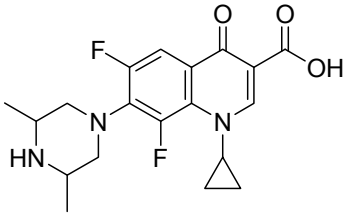
058-29E



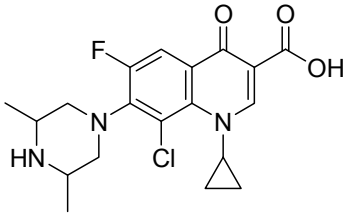
059-29F



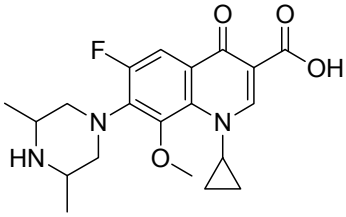
060-30A



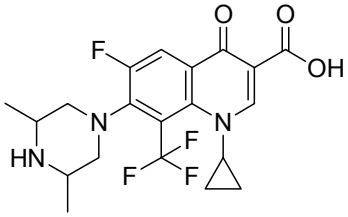
061-30B



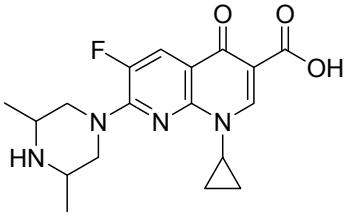
062-30C



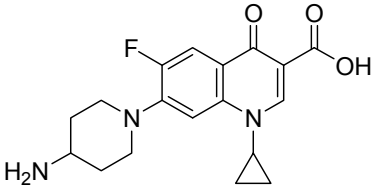
063-30D



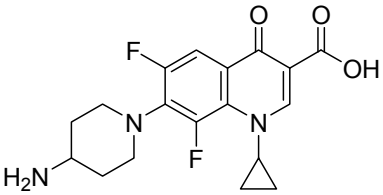
064-30E



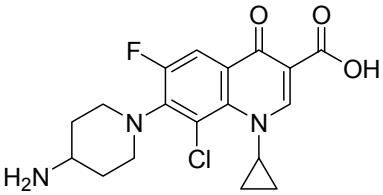
065-30F



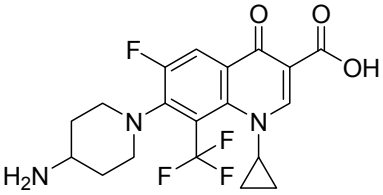
066-31A



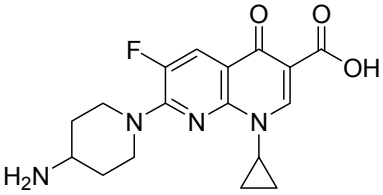
067-31B



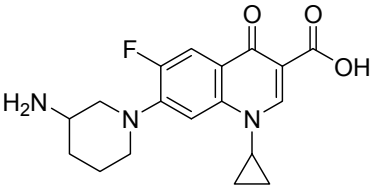
068-31C



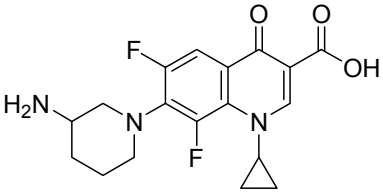
070-31E



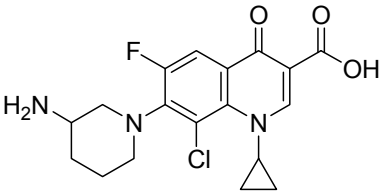
071-31F



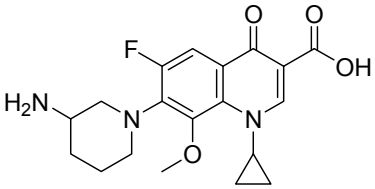
072-32A



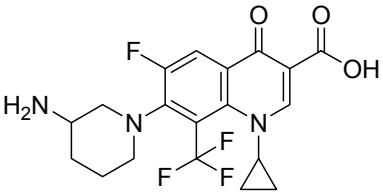
073-32B



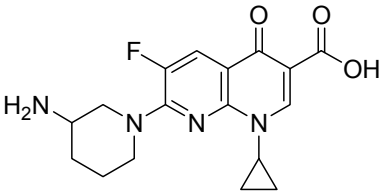
074-32C



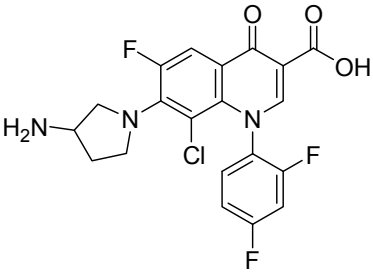
075-32D



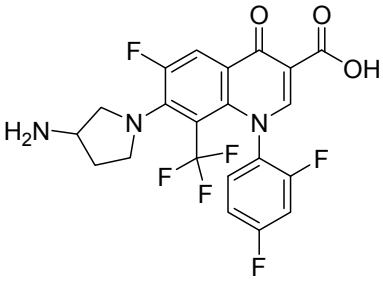
076-32E



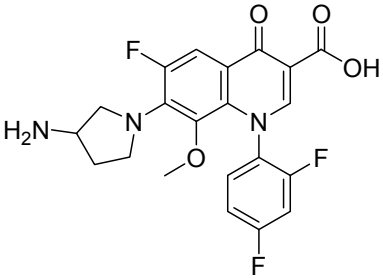
077-32F



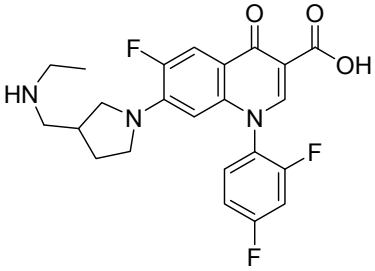
078-33B



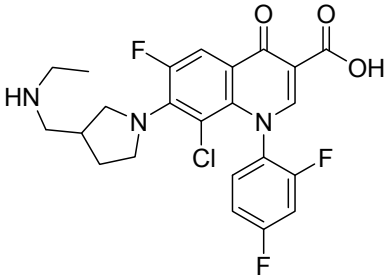
079-34B



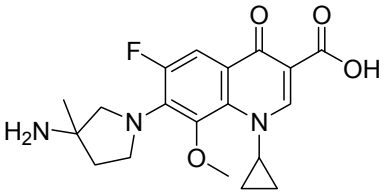
080-35B



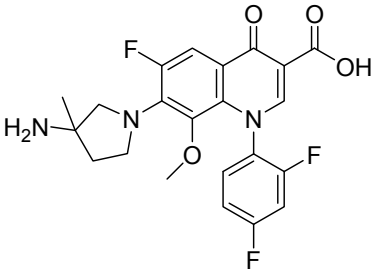
081-36B



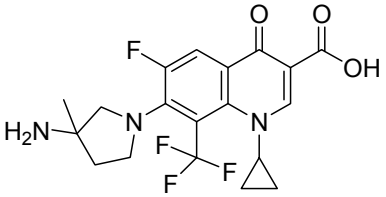
082-37B



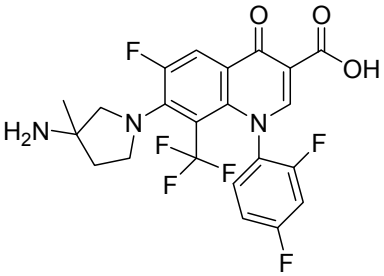
083-38A



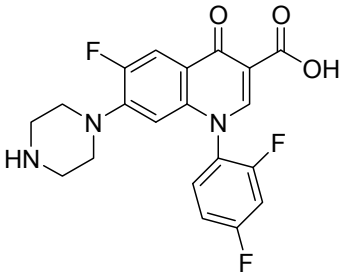
084-38B



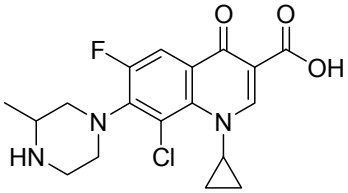
085-39A



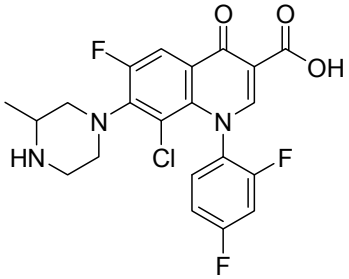
086-39B



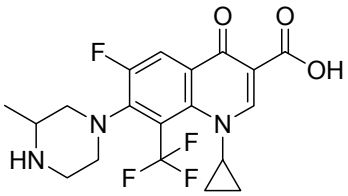
087-40B



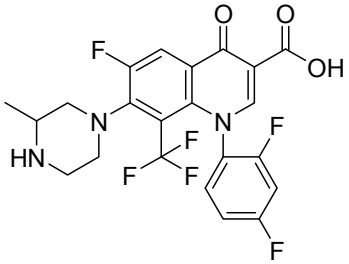
088-41A



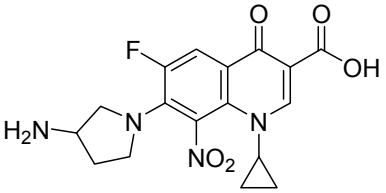
089-41B

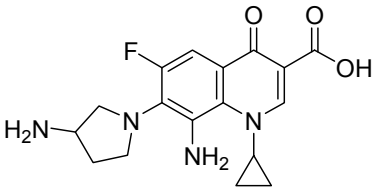


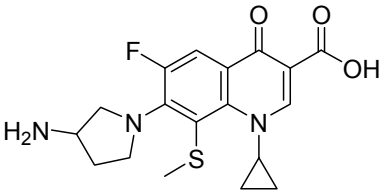
090-42A

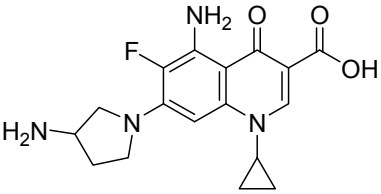


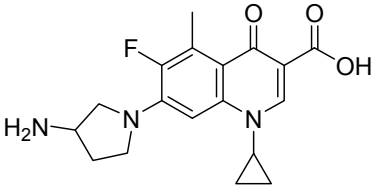
091-42B

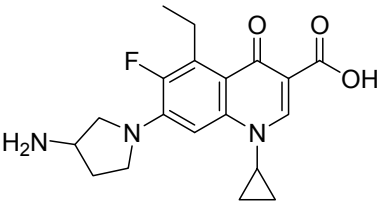

092-48

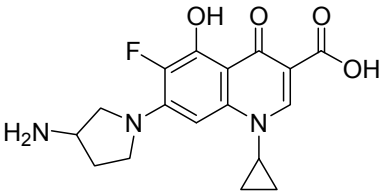

093-49

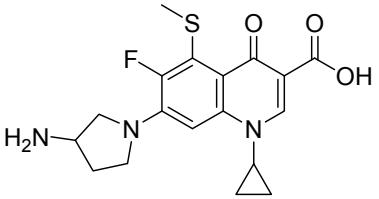

094-50

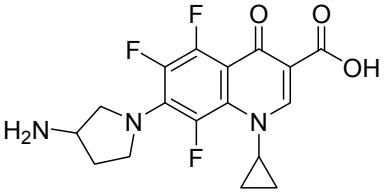

095-51

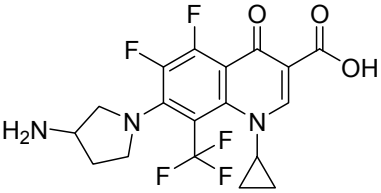

096-52

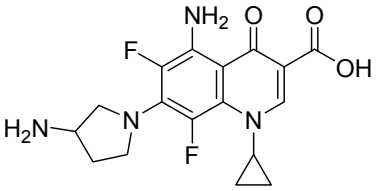

097-53

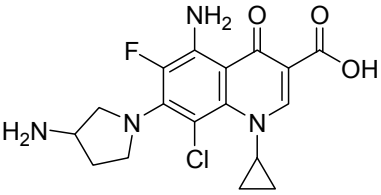

098-54

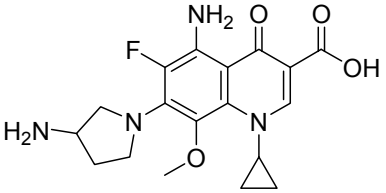

099-55

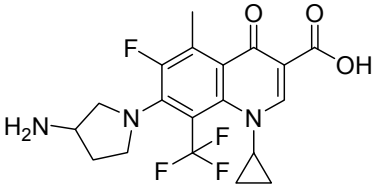

100-56

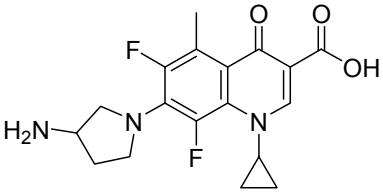

101-57


102-58

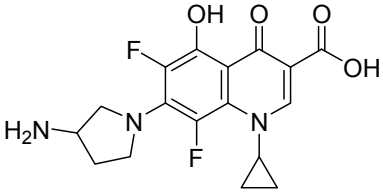

103-59


104-60

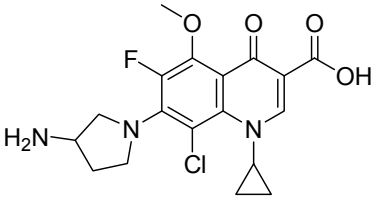

105-61



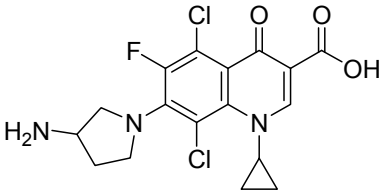
106-62



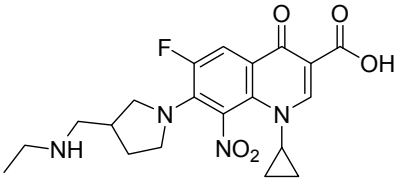
107-63



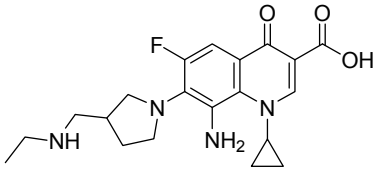
108-64



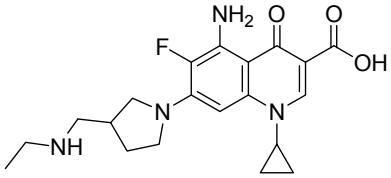
109-65



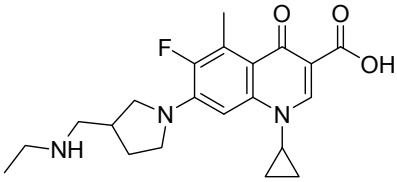
110-70



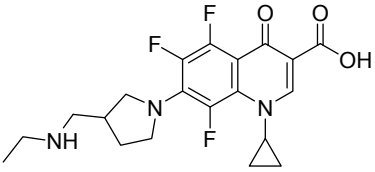
111-71



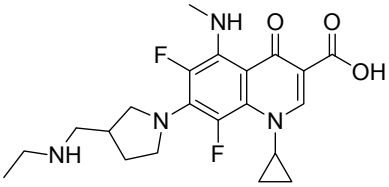
112-72



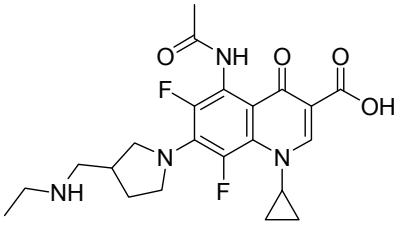
113-73



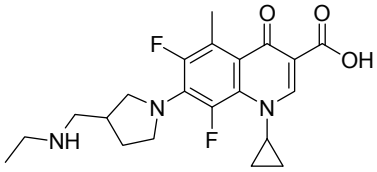
114-74



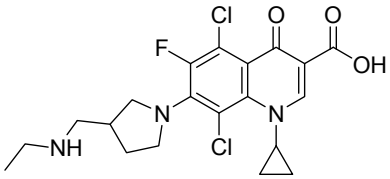
115-75



116-76



117-77



118-78

Table SI1. Compound ID, values of IC_{50} and MIC of the 117 fluoroquinolones used in this work.

Compound ID	IC_{50} (µg/ml)	$MIC_{Gram-neg.}$ (µg/ml)	Inaccurate Values (<, > o ≤)
001-1-Norfloxacin	500	0.06	>
002-2-Pefloxacin	500	0.07	>
003-3-Ofloxacin	500	0.2	>
004-4-Ciprofloxacin	380	0.1	
005-5-Fleroxacin	500	0.35	>
006-6-Tosufloxacin	128	0.09	
007-7-PD117558	11	0.09	
008-8	160	0.2	
009-9	500	0.4	>
010-10	58	1.82	
011-11	500	50	>
012-13	240	4.17	
013-14	500	0.26	>
014-15	310	0.56	
015-16	160	0.46	
016-17	190	0.8	
017-18	500	12.5	>
018-19	300	0.12	
019-20	310	0.13	
020-21	30	0.04	
021-22	160	0.3	
022-23A	38	0.2	
023-23B	120	0.1	
024-23C	150	0.1	
025-23D	150	0.3	
026-23E	280	12.5	
027-23F	58	0.26	
028-24A	72	0.09	
029-24C	26	0.03	
030-24D	45	0.07	
031-24E	300	0.2	
032-24F	98	0.03	
033-25A	81	0.2	
034-25B	11	0.05	
035-25C	8	0.06	≤
036-25D	10	0.11	
037-25E	38	0.52	
038-25F	51	0.14	
039-26C	11	0.14	<
040-26D	22	0.26	
041-26E	120	0.6	
042-26F	160	0.21	
043-27A	23	0.52	
044-27B	8	0.13	
045-27C	8	0.07	
046-27D	38	0.26	
047-27E	110	1	
048-27F	25	0.35	
049-28A	205	0.4	
050-28B	8	0.23	
051-28C	23	0.26	
052-28D	120	0.52	
053-28E	230	30	>
054-28F	58	0.6	
055-29B	47	0.07	
056-29C	43	0.07	
057-29D	82	0.07	
058-29E	160	0.15	
059-29F	120	0.09	
060-30A	500	0.08	>
061-30B	150	0.05	
062-30C	140	0.23	
063-30D	500	0.34	
064-30E	500	0.91	
065-30F	250	0.17	
066-31A	230	0.26	
067-31B	23	0.2	
068-31C	18	0.08	

Table SI1. (*Continued…*)

Compound ID	IC_{50} (µg/ml)	$MIC_{Gram\text{-}neg.}$ (µg/ml)	Inaccurate Values (<, > o ≤)
070-31E	20	0.26	
071-31F	100	0.23	
072-32A	500	0.91	>
073-32B	53	0.13	
074-32C	24	0.07	
075-32D	69	0.23	
076-32E	500	0.52	>
077-32F	100	0.4	
078-33B	89	0.23	
079-34B	99	0.52	
080-35B	290	0.35	
081-36B	190	0.8	
082-37B	130	1.05	
083-38A	38	0.26	
084-38B	250	0.46	
085-39A	110	1	
086-39B	18	2.07	
087-40B	500	0.11	>
088-41A	45	0.08	
089-41B	500	2.07	>
090-42A	220	0.46	
091-42B	500	4.14	>
092-48	69	0.46	
093-49	260	0.53	
094-50	31	0.2	
095-51	54	0.04	
096-52	14	0.09	
097-53	46	1.81	>
098-54	72	0.04	
099-55	92	1.58	>
100-56	100	0.08	
101-57	190	25	
102-58	15	0.01	
103-59	59	0.08	
104-60	100	0.11	
105-61	290	0.91	
106-62	11	0.02	
107-63	43	0.03	
108-64	500	2.08	>
109-65	500	0.13	>
110-70	67	1.05	
111-71	290	0.91	
112-72	61	0.35	
113-73	42	0.6	
114-74	46	0.56	
115-75	53	0.69	
116-76	500	25	>
117-77	9	0.28	
118-78	270	0.6	

Results, fitting algorithm and functions employed in the MATLAB data-fitting process for the library of fluoroquinolones:

Fitting Algorithm

```
-------------
diary resultados.txt

XYZ=load('datos.d')

W=load('Param.d');
x0(:,1)=W(:,1);


RT(:,1)=XYZ(:,1);
RT(:,2)=XYZ(:,2);
RT(:,3)=XYZ(:,3);
RT(:,4)=XYZ(:,4);
RT(:,5)=XYZ(:,5);
RT(:,6)=XYZ(:,6);
RT(:,7)=XYZ(:,7);
RT(:,8)=XYZ(:,8);
RT(:,9)=XYZ(:,9);
RT(:,10)=XYZ(:,10);
RT(:,11)=XYZ(:,11);
RT(:,12)=XYZ(:,12);
RT(:,13)=XYZ(:,13);
RT(:,14)=XYZ(:,14);
RT(:,15)=XYZ(:,15);
RT(:,16)=XYZ(:,16);
RT(:,17)=XYZ(:,17);
RT(:,18)=XYZ(:,18);
ee=XYZ(:,19);

q=85;

t=sort(ee);

ll=t(q,1)
lll=t(1,1)
llll=t(q,1)-t(1,1)


for i=1:q
 S(i,1)=((ee(i,1)-lll)/llll);
end


lb=[]'
ub=[]'


options=optimset('Jacobian','off','DiffMaxChange',5e30,'DiffMinChange',5e-
30,'LargeScale','on','LevenbergMarquardt','off','TolFun',1e-9,'TolX',1e-
9,'MaxIter',3000000,'MaxFunEvals',996000000);
[u,resnorm,residual,exitflag,output]=lsqcurvefit(@funcion,x0,RT,S,lb,ub,options)


Param(:,1)=x0;
Param(:,2)=u;

%save Param_finales.txt Param -ASCII

Param
resnorm
Residual_absolute=max(abs(residual))

diary off
-------------
```

Functions

```
-------------
function [V] = funcion(u,RT,V)
q=85;

Par=u;
h1=Par(1);
h2=Par(2);
h3=Par(3);
h4=Par(4);
h5=Par(5);
h6=Par(6);
h7=Par(7);
h8=Par(8);
h9=Par(9);
h10=Par(10);
h11=Par(11);
h12=Par(12);
h13=Par(13);
h14=Par(14);
h15=Par(15);
h16=Par(16);
h17=Par(17);
h18=Par(18);

for n=1:q;
opt(n,1)=0.058539d0;
opt(n,2)=0.097922d0;
opt(n,3)=2.716396d0;
opt(n,4)=0.996478d0;
opt(n,5)=0.095266d0;
opt(n,6)=0.266749d0;
opt(n,7)=0.215403d0;
opt(n,8)=0.560622d0;
opt(n,9)=-5.39527d0;
opt(n,10)=0.836179d0;
opt(n,11)=202.3135d0;
opt(n,12)=2.022805d0;
opt(n,13)=6.533513d0;
opt(n,14)=21.75996d0;
opt(n,15)=-6.61889d0;
opt(n,16)=-0.049637d0;
opt(n,17)=0.049289d0;
opt(n,18)=0.242573d0;
end

q1=sqrt([RT(:,1)-opt(:,1)].^2);
q2=sqrt([RT(:,2)-opt(:,2)].^2);
q3=sqrt([RT(:,3)-opt(:,3)].^2);
q4=sqrt([RT(:,4)-opt(:,4)].^2);
q5=sqrt([RT(:,5)-opt(:,5)].^2);
q6=sqrt([RT(:,6)-opt(:,6)].^2);
q7=sqrt([RT(:,7)-opt(:,7)].^2);
q8=sqrt([RT(:,8)-opt(:,8)].^2);
q9=sqrt([RT(:,9)-opt(:,9)].^2);
q10=sqrt([RT(:,10)-opt(:,10)].^2);
q11=sqrt([RT(:,11)-opt(:,11)].^2);
q12=sqrt([RT(:,12)-opt(:,12)].^2);
q13=sqrt([RT(:,13)-opt(:,13)].^2);
q14=sqrt([RT(:,14)-opt(:,14)].^2);
q15=sqrt([RT(:,15)-opt(:,15)].^2);
q16=sqrt([RT(:,16)-opt(:,16)].^2);
q17=sqrt([RT(:,17)-opt(:,17)].^2);
q18=sqrt([RT(:,18)-opt(:,18)].^2);

r1=h1*q1;
r2=h2*q2;
r3=h3*q3;
r4=h4*q4;
r5=h5*q5;
r6=h6*q6;
r7=h7*q7;
r8=h8*q8;
r9=h9*q9;
r10=h10*q10;
r11=h11*q11;
```

```
r12=h12*q12;
r13=h13*q13;
r14=h14*q14;
r15=h15*q15;
r16=h16*q16;
r17=h17*q17;
r18=h18*q18;

s=r1+r2+r3+r4+r5+r6+r7+r8+r9+r10+r11+r12+r13+r14+r15+r16+r17+r18;


t=sort(s);

ll=t(q,1);
lll=t(1,1);
llll=t(q,1)-t(1,1);


for i=1:q
 z(i,1)=1-((s(i,1)-lll)/llll);
end

V=z;
-------------
```

Results

```
XYZ =

   1.0e+02 *

  Columns 1 through 5

   0.00037000000000     -0.00027000000000       0.02289000000000       0.00995000000000
0.00137000000000
   0.00045000000000     -0.00040000000000       0.02002000000000       0.00978000000000
0.00227000000000
   0.00048000000000       0.00037000000000       0.02321000000000       0.00981000000000
0.00271000000000
   0.00027000000000     -0.00004000000000       0.02304000000000       0.00967000000000
0.00228000000000
   0.00034000000000     -0.00005000000000       0.02216000000000       0.00984000000000
0.00225000000000
   0.00044000000000     -0.00026000000000       0.01985000000000       0.00955000000000
0.00285000000000
   0.00028000000000       0.00039000000000       0.02486000000000       0.00996000000000
0.00131000000000
   0.00028000000000       0.00028000000000       0.02377000000000       0.01000000000000
0.00182000000000
   0.00036000000000       0.00039000000000       0.02377000000000       0.01000000000000
0.00333000000000
   0.00041000000000       0.00016000000000       0.02362000000000       0.00980000000000
0.00203000000000
   0.00034000000000     -0.00012000000000       0.02394000000000       0.01000000000000
0.00222000000000
   0.00036000000000       0.00054000000000       0.02417000000000       0.01000000000000
0.00332000000000
   0.00036000000000       0.00037000000000       0.02665000000000       0.00997000000000
0.00196000000000
   0.00041000000000       0.00046000000000       0.02297000000000       0.00966000000000
0.00196000000000
   0.00034000000000       0.00034000000000       0.02267000000000       0.00998000000000
0.00255000000000
   0.00034000000000     -0.00025000000000       0.02355000000000       0.00989000000000
0.00304000000000
   0.00036000000000       0.00021000000000       0.02610000000000       0.01000000000000
0.00270000000000
   0.00043000000000       0.00031000000000       0.02273000000000       0.00966000000000
0.00162000000000
   0.00055000000000     -0.00015000000000       0.02723000000000       0.00990000000000
0.00189000000000
   0.00034000000000       0.00017000000000       0.02242000000000       0.00994000000000
0.00273000000000
   0.00035000000000     -0.00003000000000       0.02144000000000       0.00980000000000
0.00260000000000
   0.00038000000000       0.00061000000000       0.02165000000000       0.00997000000000
0.00316000000000
   0.00055000000000       0.00006000000000       0.02509000000000       0.00979000000000
0.00233000000000
   0.00035000000000       0.00041000000000       0.02051000000000       0.00988000000000
0.00216000000000
   0.00041000000000       0.00052000000000       0.02170000000000       0.00959000000000
0.00268000000000
   0.00053000000000       0.00004000000000       0.02532000000000       0.00976000000000
0.00180000000000
   0.00034000000000       0.00041000000000       0.02147000000000       0.00980000000000
0.00261000000000
   0.00036000000000     -0.00004000000000       0.01940000000000       0.00962000000000
0.00362000000000
   0.00040000000000       0.00059000000000       0.01960000000000       0.00985000000000
0.00400000000000
   0.00045000000000       0.00051000000000       0.01925000000000       0.00949000000000
0.00182000000000
   0.00060000000000       0.00003000000000       0.02302000000000       0.00967000000000
0.00268000000000
   0.00036000000000       0.00040000000000       0.01879000000000       0.00973000000000
0.00253000000000
   0.00035000000000     -0.00001000000000       0.02273000000000       0.00958000000000
0.00425000000000
```

0.00044000000000	0.00056000000000	0.02225000000000	0.00951000000000	0.00306000000000
0.00035000000000	0.00045000000000	0.02259000000000	0.00966000000000	0.00400000000000
0.00040000000000	0.00037000000000	0.02310000000000	0.01000000000000	0.00282000000000
0.00040000000000	0.00019000000000	0.02530000000000	0.01000000000000	0.00251000000000
0.00043000000000	0.00029000000000	0.02219000000000	0.00969000000000	0.00204000000000
0.00052000000000	-0.00017000000000	0.02653000000000	0.00955000000000	0.00255000000000
0.00037000000000	0.00015000000000	0.02136000000000	0.00991000000000	0.00222000000000
0.00039000000000	0.00020000000000	0.02473000000000	0.00996000000000	0.00334000000000
0.00039000000000	0	0.02631000000000	0.00999000000000	0.00237000000000
0.00045000000000	0.00012000000000	0.02384000000000	0.00967000000000	0.00249000000000
0.00058000000000	-0.00032000000000	0.02742000000000	0.00980000000000	0.00273000000000
0.00035000000000	-0.00005000000000	0.02452000000000	0.00984000000000	0.00263000000000
0.00040000000000	-0.00030000000000	0.02080000000000	0.00990000000000	0.00158000000000
0.00045000000000	0.00036000000000	0.02101000000000	0.00997000000000	0.00265000000000
0.00045000000000	0.00017000000000	0.02305000000000	0.01000000000000	0.00186000000000
0.00054000000000	-0.00013000000000	0.02453000000000	0.00982000000000	0.00072000000000
0.00040000000000	0.00012000000000	0.01945000000000	0.00986000000000	0.00209000000000
0.00039000000000	0.00052000000000	0.02256000000000	0.00994000000000	0.00262000000000
0.00039000000000	0.00036000000000	0.02457000000000	0.01000000000000	0.00223000000000
0.00044000000000	0.00045000000000	0.02179000000000	0.00963000000000	0.00147000000000
0.00036000000000	0.00033000000000	0.02140000000000	0.00983000000000	0.00201000000000
0.00047000000000	-0.00033000000000	0.02276000000000	0.00998000000000	0.00333000000000
0.00066000000000	-0.00049000000000	0.02417000000000	0.00987000000000	0.00053000000000
0.00051000000000	-0.00025000000000	0.02013000000000	0.00973000000000	0.00166000000000
0.00044000000000	-0.00026000000000	0.01985000000000	0.00950000000000	0.00306000000000
0.00047000000000	0.00002000000000	0.02134000000000	0.00986000000000	0.00302000000000
0.00045000000000	0.00008000000000	0.02152000000000	0.00956000000000	0.00224000000000
0.00053000000000	-0.00040000000000	0.01913000000000	0.00962000000000	0.00354000000000
0.00060000000000	-0.00034000000000	0.02580000000000	0.00987000000000	0.00392000000000
0.00070000000000	-0.00063000000000	0.02297000000000	0.00970000000000	0.00378000000000
0.00039000000000	0.00009000000000	0.02486000000000	0.01000000000000	0.00240000000000
0.00055000000000	-0.00025000000000	0.02613000000000	0.00976000000000	0.00287000000000
0.00048000000000	0	0.02458000000000	0.00984000000000	0.00251000000000
0.00036000000000	0.00011000000000	0.02511000000000	0.00995000000000	0.00243000000000
0.00043000000000	-0.00002000000000	0.02804000000000	0.00980000000000	0.00390000000000
0.00039000000000	0.00028000000000	0.02277000000000	0.00988000000000	0.00288000000000
0.00039000000000	-0.00008000000000	0.02170000000000	0.00989000000000	0.00233000000000
0.00039000000000	0.00057000000000	0.02328000000000	0.00991000000000	0.00255000000000

0.00041000000000	0.00100000000000	0.02395000000000	0.01000000000000	0.00209000000000
0.00041000000000	0.00077000000000	0.02301000000000	0.00998000000000	0.00274000000000
0.00041000000000	0.00067000000000	0.02469000000000	0.01000000000000	0.00190000000000
0.00047000000000	0.00076000000000	0.02213000000000	0.00968000000000	0.00111000000000
0.00059000000000	-0.00007000000000	0.02471000000000	0.00981000000000	0.00083000000000
0.00041000000000	0.00053000000000	0.02198000000000	0.00996000000000	0.00345000000000
0.00041000000000	0.00094000000000	0.02351000000000	0.00998000000000	0.00359000000000
0.00047000000000	0.00022000000000	0.02309000000000	0.00972000000000	0.00328000000000
0.00037000000000	0.00031000000000	0.02353000000000	0.00986000000000	0.00235000000000
0.00038000000000	0.00047000000000	0.02167000000000	0.00966000000000	0.00341000000000
0.00038000000000	0.00010000000000	0.02085000000000	0.00967000000000	0.00279000000000
0.00041000000000	0.00123000000000	0.02247000000000	0.00979000000000	0.00334000000000
0.00044000000000	0.00105000000000	0.02248000000000	0.00975000000000	0.00298000000000
0.00041000000000	0.00103000000000	0.02109000000000	0.00979000000000	0.00213000000000

Columns 6 through 10

0.00128000000000	0.00194000000000	0.00409000000000	0	0.00691000000000
0.00116000000000	0.00204000000000	0.00396000000000	0.20920000000000	0.00974000000000
0.00174000000000	0.00177000000000	0.00439000000000	0.39050000000000	0.01032000000000
0.00200000000000	0.00178000000000	0.00472000000000	0	0.00846000000000
0.00241000000000	0.00182000000000	0.00490000000000	0	0.00817000000000
0.00256000000000	0.00160000000000	0.00481000000000	0.22080000000000	0.01088000000000
0.00096000000000	0.00204000000000	0.00425000000000	0.04780000000000	0.00720000000000
0.00093000000000	0.00212000000000	0.00401000000000	0.04760000000000	0.00744000000000
0.00101000000000	0.00201000000000	0.00420000000000	0.04760000000000	0.00733000000000
0.00118000000000	0.00194000000000	0.00460000000000	0.04780000000000	0.00723000000000
0.00071000000000	0.00229000000000	0.00442000000000	0	0.00654000000000
0.00090000000000	0.00211000000000	0.00416000000000	0.04760000000000	0.00726000000000
0.00125000000000	0.00221000000000	0.00525000000000	0	0.00749000000000
0.00179000000000	0.00206000000000	0.00536000000000	0	0.00735000000000
0.00083000000000	0.00208000000000	0.00392000000000	0	0.00663000000000
0.00113000000000	0.00217000000000	0.00453000000000	0	0.00722000000000
0.00180000000000	0.00215000000000	0.00533000000000	0	0.00864000000000
0.00177000000000	0.00192000000000	0.00523000000000	0	0.00758000000000
0.00290000000000	0.00193000000000	0.00566000000000	0.16420000000000	0.00776000000000
0.00101000000000	0.00199000000000	0.00397000000000	0	0.00692000000000
0.00131000000000	0.00230000000000	0.00476000000000	0	0.00786000000000
0.00117000000000	0.00200000000000	0.00455000000000	0.04760000000000	0.00735000000000
0.00321000000000	0.00187000000000	0.00582000000000	0.16430000000000	0.00844000000000

0.00133000000000	0.00189000000000	0.00432000000000	0	0.00763000000000
0.00333000000000	0.00177000000000	0.00569000000000	0	0.00880000000000
0.00405000000000	0.00173000000000	0.00605000000000	0.16420000000000	0.00915000000000
0.00247000000000	0.00170000000000	0.00459000000000	0	0.00832000000000
0.00153000000000	0.00268000000000	0.00514000000000	0	0.00869000000000
0.00123000000000	0.00195000000000	0.00500000000000	0.04760000000000	0.00806000000000
0.00177000000000	0.00188000000000	0.00555000000000	0	0.00848000000000
0.00315000000000	0.00203000000000	0.00583000000000	0.16520000000000	0.00862000000000
0.00157000000000	0.00226000000000	0.00450000000000	0	0.00819000000000
0.00209000000000	0.00219000000000	0.00551000000000	0	0.00915000000000
0.00266000000000	0.00179000000000	0.00584000000000	0	0.00877000000000
0.00204000000000	0.00233000000000	0.00505000000000	0	0.00856000000000
0.00119000000000	0.00215000000000	0.00429000000000	0.04780000000000	0.00705000000000
0.00157000000000	0.00217000000000	0.00539000000000	0	0.00759000000000
0.00189000000000	0.00189000000000	0.00512000000000	0	0.00741000000000
0.00209000000000	0.00200000000000	0.00646000000000	0.16600000000000	0.01040000000000
0.00118000000000	0.00193000000000	0.00391000000000	0	0.00696000000000
0.00204000000000	0.00195000000000	0.00464000000000	0.04780000000000	0.00716000000000
0.00202000000000	0.00198000000000	0.00563000000000	0	0.00824000000000
0.00258000000000	0.00175000000000	0.00533000000000	0	0.00770000000000
0.00305000000000	0.00196000000000	0.00591000000000	0.16660000000000	0.00827000000000
0.00188000000000	0.00180000000000	0.00436000000000	0	0.00726000000000
0.00170000000000	0.00192000000000	0.00422000000000	0	0.00734000000000
0.00136000000000	0.00206000000000	0.00429000000000	0.04780000000000	0.00738000000000
0.00171000000000	0.00207000000000	0.00536000000000	0	0.00847000000000
0.00259000000000	0.00212000000000	0.00570000000000	0.16640000000000	0.00800000000000
0.00139000000000	0.00188000000000	0.00399000000000	0	0.00749000000000
0.00148000000000	0.00203000000000	0.00436000000000	0.04780000000000	0.00714000000000
0.00196000000000	0.00210000000000	0.00543000000000	0	0.00789000000000
0.00205000000000	0.00182000000000	0.00520000000000	0	0.00762000000000
0.00135000000000	0.00180000000000	0.00405000000000	0	0.00726000000000
0.00162000000000	0.00180000000000	0.00439000000000	0.20730000000000	0.01035000000000
0.00292000000000	0.00184000000000	0.00516000000000	0.66290000000000	0.01088000000000
0.00228000000000	0.00192000000000	0.00479000000000	0.20120000000000	0.01057000000000
0.00284000000000	0.00166000000000	0.00495000000000	0.21580000000000	0.01109000000000
0.00280000000000	0.00165000000000	0.00494000000000	0.20730000000000	0.01192000000000
0.00163000000000	0.00189000000000	0.00518000000000	0	0.00805000000000
0.00214000000000	0.00173000000000	0.00470000000000	0.20340000000000	0.01095000000000

0.00300000000000	0.00205000000000	0.00554000000000	0.16690000000000	0.00883000000000
0.00296000000000	0.00173000000000	0.00519000000000	0.69030000000000	0.01114000000000
0.00181000000000	0.00192000000000	0.00532000000000	0	0.00782000000000
0.00298000000000	0.00190000000000	0.00608000000000	0.16490000000000	0.00885000000000
0.00253000000000	0.00202000000000	0.00563000000000	0	0.00807000000000
0.00222000000000	0.00206000000000	0.00473000000000	0	0.00685000000000
0.00191000000000	0.00192000000000	0.00611000000000	0	0.00750000000000
0.00155000000000	0.00215000000000	0.00438000000000	0	0.00717000000000
0.00138000000000	0.00175000000000	0.00442000000000	0	0.00573000000000
0.00110000000000	0.00217000000000	0.00430000000000	0	0.00772000000000
0.00106000000000	0.00234000000000	0.00432000000000	0.12870000000000	0.00923000000000
0.00132000000000	0.00174000000000	0.00423000000000	0.04770000000000	0.00709000000000
0.00162000000000	0.00180000000000	0.00501000000000	0	0.00805000000000
0.00228000000000	0.00171000000000	0.00501000000000	0	0.00709000000000
0.00280000000000	0.00169000000000	0.00593000000000	0.16280000000000	0.00893000000000
0.00110000000000	0.00188000000000	0.00433000000000	0.04750000000000	0.00607000000000
0.00108000000000	0.00223000000000	0.00425000000000	0.04760000000000	0.00764000000000
0.00337000000000	0.00199000000000	0.00603000000000	0	0.01035000000000
0.00258000000000	0.00204000000000	0.00525000000000	0	0.00851000000000
0.00259000000000	0.00165000000000	0.00483000000000	0	0.00899000000000
0.00277000000000	0.00164000000000	0.00497000000000	0	0.00822000000000
0.00231000000000	0.00211000000000	0.00485000000000	0.12890000000000	0.01059000000000
0.00238000000000	0.00130000000000	0.00458000000000	0.04760000000000	0.00828000000000
0.00297000000000	0.00224000000000	0.00637000000000	0	0.01506000000000

Columns 11 through 15

1.82617000000000	0.02048000000000	0.04730000000000	0.32161000000000	-0.04719000000000
2.20699000000000	0.02059000000000	0.03874000000000	0.28138000000000	-0.05259000000000
1.56391000000000	0.02044000000000	0.06272000000000	0.33234000000000	-0.05141000000000
1.32242000000000	0.02043000000000	0.05807000000000	0.36196000000000	-0.04530000000000
1.36448000000000	0.02035000000000	0.06964000000000	0.32073000000000	-0.04825000000000
2.54343000000000	0.02062000000000	0.05472000000000	0.37253000000000	-0.06206000000000
1.14921000000000	0.02044000000000	0.04608000000000	0.28955000000000	-0.04754000000000
1.14921000000000	0.02042000000000	0.04526000000000	0.26332000000000	-0.05070000000000
1.19127000000000	0.02036000000000	0.04190000000000	0.25320000000000	-0.05217000000000
1.24911000000000	0.02038000000000	0.04374000000000	0.29643000000000	-0.05191000000000
1.07306000000000	0.02035000000000	0.02495000000000	0.21298000000000	-0.03203000000000
1.12386000000000	0.02036000000000	0.02580000000000	0.26162000000000	-0.05086000000000
1.12386000000000	0.02025000000000	0.02706000000000	0.25300000000000	-0.04462000000000

1.18169000000000	0.02043000000000	0.05121000000000	0.31500000000000	-0.04819000000000
1.07306000000000	0.02030000000000	0.02115000000000	0.26728000000000	-0.04550000000000
1.14048000000000	0.02035000000000	0.04177000000000	0.28364000000000	-0.04586000000000
1.19127000000000	0.02025000000000	0.04163000000000	0.24311000000000	-0.04906000000000
1.24911000000000	0.02043000000000	0.06977000000000	0.34572000000000	-0.05034000000000
1.36479000000000	0.02049000000000	0.05074000000000	0.35073000000000	-0.05781000000000
1.14048000000000	0.02031000000000	0.03724000000000	0.26583000000000	-0.04504000000000
1.21189000000000	0.02035000000000	0.04599000000000	0.29257000000000	-0.04736000000000
1.26269000000000	0.02036000000000	0.06110000000000	0.33345000000000	-0.05166000000000
1.43620000000000	0.02049000000000	0.05610000000000	0.36310000000000	-0.05991000000000
1.21189000000000	0.02031000000000	0.04472000000000	0.28013000000000	-0.04630000000000
1.47312000000000	0.02044000000000	0.08087000000000	0.43557000000000	-0.05079000000000
1.58879000000000	0.02049000000000	0.08018000000000	0.39534000000000	-0.06002000000000
1.36448000000000	0.02031000000000	0.05370000000000	0.35911000000000	-0.05037000000000
1.27931000000000	0.02035000000000	0.05324000000000	0.31793000000000	-0.05040000000000
1.33010000000000	0.02036000000000	0.04987000000000	0.31548000000000	-0.05716000000000
1.38794000000000	0.02044000000000	0.07645000000000	0.43187000000000	-0.05360000000000
1.50362000000000	0.02049000000000	0.06183000000000	0.40217000000000	-0.05132000000000
1.27931000000000	0.02031000000000	0.04823000000000	0.26637000000000	-0.04866000000000
1.35413000000000	0.02035000000000	0.05623000000000	0.38917000000000	-0.05125000000000
1.46276000000000	0.02044000000000	0.08216000000000	0.41155000000000	-0.05779000000000
1.35413000000000	0.02031000000000	0.05206000000000	0.35823000000000	-0.05152000000000
1.89565000000000	0.02049000000000	0.04676000000000	0.33442000000000	-0.05389000000000
1.89565000000000	0.02040000000000	0.04580000000000	0.25428000000000	-0.05467000000000
1.97487000000000	0.02056000000000	0.06911000000000	0.40126000000000	-0.04825000000000
2.13331000000000	0.02061000000000	0.06230000000000	0.45272000000000	-0.05408000000000
1.82617000000000	0.02045000000000	0.04608000000000	0.28411000000000	-0.04773000000000
2.09628000000000	0.02054000000000	0.04008000000000	0.43856000000000	-0.05687000000000
2.09628000000000	0.02045000000000	0.04234000000000	0.34783000000000	-0.06320000000000
2.17550000000000	0.02061000000000	0.06284000000000	0.46849000000000	-0.05038000000000
2.33394000000000	0.02065000000000	0.05631000000000	0.48951000000000	-0.05630000000000
2.02680000000000	0.02050000000000	0.03903000000000	0.39993000000000	-0.04984000000000
1.94300000000000	0.02051000000000	0.04223000000000	0.29342000000000	-0.05033000000000
2.01248000000000	0.02052000000000	0.04278000000000	0.33244000000000	-0.05551000000000
2.01248000000000	0.02042000000000	0.04325000000000	0.27359000000000	-0.06145000000000
2.25014000000000	0.02063000000000	0.05595000000000	0.37893000000000	-0.05451000000000
1.94300000000000	0.02047000000000	0.04034000000000	0.28410000000000	-0.04913000000000
1.99596000000000	0.02052000000000	0.04191000000000	0.36580000000000	-0.05517000000000

1.99596000000000	0.02043000000000	0.04008000000000	0.26713000000000	-0.05745000000000
2.07518000000000	0.02059000000000	0.06652000000000	0.44928000000000	-0.05331000000000
1.92649000000000	0.02047000000000	0.03998000000000	0.31684000000000	-0.04750000000000
2.28241000000000	0.02055000000000	0.04126000000000	0.21551000000000	-0.05751000000000
2.53720000000000	0.02072000000000	0.05412000000000	0.43502000000000	-0.07324000000000
2.36734000000000	0.02068000000000	0.06985000000000	0.31577000000000	-0.06383000000000
2.54343000000000	0.02062000000000	0.05652000000000	0.32319000000000	-0.06013000000000
2.61884000000000	0.02055000000000	0.05692000000000	0.31796000000000	-0.05988000000000
1.31652000000000	0.02044000000000	0.06670000000000	0.35691000000000	-0.05324000000000
2.46955000000000	0.02068000000000	0.05989000000000	0.42901000000000	-0.06381000000000
1.43220000000000	0.02049000000000	0.05016000000000	0.37750000000000	-0.04892000000000
2.63941000000000	0.02072000000000	0.05267000000000	0.42242000000000	-0.06619000000000
1.99596000000000	0.02042000000000	0.04319000000000	0.29217000000000	-0.05956000000000
2.23362000000000	0.02063000000000	0.05788000000000	0.43829000000000	-0.06136000000000
1.30695000000000	0.02061000000000	0.04191000000000	0.36259000000000	-0.06211000000000
1.19127000000000	0.02046000000000	0.05613000000000	0.28141000000000	-0.04321000000000
1.24911000000000	0.02026000000000	0.04092000000000	0.31555000000000	-0.04600000000000
1.19127000000000	0.02046000000000	0.05838000000000	0.29409000000000	-0.04888000000000
1.19127000000000	0.02047000000000	0.04192000000000	0.31748000000000	-0.04788000000000
1.19127000000000	0.02040000000000	0.04255000000000	0.27626000000000	-0.04817000000000
1.24206000000000	0.02037000000000	0.04038000000000	0.28204000000000	-0.05414000000000
1.24206000000000	0.02047000000000	0.05893000000000	0.30322000000000	-0.05447000000000
1.24206000000000	0.02037000000000	0.05893000000000	0.27406000000000	-0.05975000000000
1.29990000000000	0.02054000000000	0.08739000000000	0.37244000000000	-0.05087000000000
1.41558000000000	0.02060000000000	0.05284000000000	0.42772000000000	-0.06601000000000
1.24206000000000	0.02048000000000	0.04561000000000	0.31088000000000	-0.05478000000000
1.24206000000000	0.02041000000000	0.04357000000000	0.29974000000000	-0.05475000000000
1.53096000000000	0.02061000000000	0.05461000000000	0.44661000000000	-0.06821000000000
1.41528000000000	0.02046000000000	0.06752000000000	0.38115000000000	-0.04354000000000
1.41528000000000	0.02046000000000	0.06961000000000	0.37863000000000	-0.05084000000000
1.41528000000000	0.02047000000000	0.05774000000000	0.39596000000000	-0.05054000000000
1.46607000000000	0.02037000000000	0.05118000000000	0.38075000000000	-0.05986000000000
1.52391000000000	0.02051000000000	0.07321000000000	0.41061000000000	-0.05920000000000
1.46607000000000	0.02017000000000	0.04969000000000	0.33240000000000	-0.05725000000000

Columns 16 through 19

0.00234000000000	0.00075000000000	0.00390000000000	0.00956023808000
0.00012000000000	0.00170000000000	0.00486000000000	0.00968283778000
0.00454000000000	0.00100000000000	0.00308000000000	0.00452294036000
0.00803000000000	0.00079000000000	0.00344000000000	0.00751452373000
0.00790000000000	0.00073000000000	0.00334000000000	0.00788136585000
0.00198000000000	0.00128000000000	0.00370000000000	0.00775569801000

0.00411000000000	0.00101000000000	0.00419000000000	0.00943365736000
0.00128000000000	0.00106000000000	0.00396000000000	0.00958561395000
0.00297000000000	0.00091000000000	0.00382000000000	0.00792695871000
0.00577000000000	0.00094000000000	0.00364000000000	0.00900967265000
0.00165000000000	0.00102000000000	0.00451000000000	0.00806225385000
0.00212000000000	0.00083000000000	0.00417000000000	0.00777236997000
0.00380000000000	0.00102000000000	0.00425000000000	0.00903740194000
0.00362000000000	0.00087000000000	0.00321000000000	0.00760931503000
0.00238000000000	0.00082000000000	0.00427000000000	0.00856381036000
0.00395000000000	0.00096000000000	0.00431000000000	0.00930338659000
0.00403000000000	0.00102000000000	0.00394000000000	0.00879139843000
0.00444000000000	0.00085000000000	0.00304000000000	0.00881999744000
0.00447000000000	0.00086000000000	0.00314000000000	0.00822975967000
0.00323000000000	0.00085000000000	0.00416000000000	0.00896161976000
0.00591000000000	0.00092000000000	0.00399000000000	0.00790285479000
0.00668000000000	0.00091000000000	0.00371000000000	0.00507807136000
0.00589000000000	0.00085000000000	0.00308000000000	0.00699329227000
0.00396000000000	0.00083000000000	0.00380000000000	0.00713444679000
0.00699000000000	0.00073000000000	0.00280000000000	0.00736875703000
0.00868000000000	0.00079000000000	0.00281000000000	0.00756194221000
0.00651000000000	0.00071000000000	0.00349000000000	0.00811468910000
0.00831000000000	0.00103000000000	0.00358000000000	0.00707031450000
0.00647000000000	0.00096000000000	0.00318000000000	0
0.00696000000000	0.00086000000000	0.00278000000000	0.00646828925000
0.00649000000000	0.00094000000000	0.00300000000000	0.00666718845000
0.00645000000000	0.00090000000000	0.00329000000000	0.00685292049000
0.00797000000000	0.00090000000000	0.00308000000000	0.00752817139000
0.00884000000000	0.00074000000000	0.00242000000000	0.00707227161000
0.00749000000000	0.00082000000000	0.00283000000000	0.00788583802000
0.00196000000000	0.00084000000000	0.00389000000000	0.00835182296000
0.00458000000000	0.00098000000000	0.00395000000000	0.00791052257000
0.00154000000000	0.00074000000000	0.00299000000000	0.00822231384000
0.00080000000000	0.00115000000000	0.00345000000000	0.00811035259000
0.00196000000000	0.00076000000000	0.00395000000000	0.00905308255000
0.00282000000000	0.00076000000000	0.00348000000000	0.00871746279000
0.00636000000000	0.00086000000000	0.00302000000000	0.00800375590000
0.00225000000000	0.00067000000000	0.00258000000000	0.00859591919000
0.00400000000000	0.00081000000000	0.00277000000000	0.00793202579000
0.00364000000000	0.00070000000000	0.00336000000000	0.00885000372000
0.00392000000000	0.00079000000000	0.00412000000000	0.00882246586000
0.00327000000000	0.00084000000000	0.00385000000000	0.00717045017000
0.00592000000000	0.00095000000000	0.00386000000000	0.00733222073000
0.00361000000000	0.00091000000000	0.00322000000000	0.00768894411000
0.00388000000000	0.00078000000000	0.00396000000000	0.00780156495000
0.00330000000000	0.00083000000000	0.00390000000000	0.00795764632000
0.00689000000000	0.00099000000000	0.00370000000000	0.00708640834000
0.00352000000000	0.00072000000000	0.00275000000000	0.00825623026000
0.00417000000000	0.00076000000000	0.00387000000000	0.00847957399000
-0.00099000000000	0.00149000000000	0.00433000000000	0.00820259960000
-0.00277000000000	0.00137000000000	0.00392000000000	0.00699294631000
-0.00184000000000	0.00135000000000	0.00374000000000	0.00893926673000
0.00361000000000	0.00130000000000	0.00374000000000	0.00715116772000
0.00111000000000	0.00132000000000	0.00369000000000	0.00695488358000
0.00561000000000	0.00088000000000	0.00298000000000	0.00857483935000
-0.00072000000000	0.00131000000000	0.00348000000000	0.00803047145000
0.00622000000000	0.00086000000000	0.00303000000000	0.00596073411000
0.00256000000000	0.00137000000000	0.00359000000000	0.00357578369000
0.00474000000000	0.00088000000000	0.00353000000000	0.00763287888000
0.00450000000000	0.00087000000000	0.00300000000000	0.00728741459000
0.00411000000000	0.00101000000000	0.00316000000000	0.00770094438000
0.00308000000000	0.00087000000000	0.00376000000000	0.00920308796000
0.00041000000000	0.00083000000000	0.00269000000000	0.00771323324000
0.00264000000000	0.00086000000000	0.00399000000000	0.00928644627000
0.00431000000000	0.00081000000000	0.00352000000000	0.00664475750000
0.00210000000000	0.00107000000000	0.00394000000000	0.00956693446000
0.00327000000000	0.00128000000000	0.00398000000000	0.00894731645000
0.00301000000000	0.00087000000000	0.00351000000000	0.00738327846000
0.00386000000000	0.00090000000000	0.00341000000000	0.00837675126000
0.00505000000000	0.00080000000000	0.00282000000000	0.00889883834000
0.00376000000000	0.00081000000000	0.00274000000000	0.00640697800000
0.00512000000000	0.00080000000000	0.00333000000000	0.00445779858000
0.00397000000000	0.00111000000000	0.00384000000000	0.00839600914000
0.00834000000000	0.00097000000000	0.00295000000000	0.00637166182000
0.00649000000000	0.00083000000000	0.00294000000000	0.00711946204000
0.00556000000000	0.00074000000000	0.00310000000000	0.00752853064000
0.00712000000000	0.00068000000000	0.00298000000000	0.00655293070000
0.00676000000000	0.00116000000000	0.00344000000000	0.00753751036000

```
   0.00809000000000   0.00062000000000   0.00260000000000   0.00674681071000
   0.00657000000000   0.00191000000000   0.00331000000000   0.00775209090000


ll =

   0.96828377800000


lll =

     0


llll =

   0.96828377800000


lb =

     []


ub =

     []

Optimization terminated: relative function value
 changing by less than OPTIONS.TolFun.

u =

  23.32337655232970
  -1.25936503981221
   1.19031263618303
  -9.77159165278830
   3.70980543978095
   4.90340625220429
  -1.05301468060386
  -6.97975645980404
   0.05225717396115
   1.57264571873312
  -0.00117462388857
  11.36454165681105
   0.02628214348797
  -0.01900506111183
   0.01252267186125
   0.56028855564767
  -9.24808297862453
  -5.81129581562833


resnorm =

   0.94055734963692


residual =

   0.00609387375019
  -0.01268473213479
  -0.08465463557102
   0.09415527339313
   0.02496705552289
  -0.08358845160060
  -0.02016293223171
  -0.04737579153633
  -0.06600938625237
  -0.08966309248955
   0.10537179416280
  -0.01478718809042
   0.03374491100404
   0.02501303764194
   0.01589998179080
  -0.07309767870972
   0.01822284164544
```

```
  -0.04775026029695
  -0.04411652904139
  -0.03051923358595
   0.00357284458239
   0.11908337059162
  -0.00718109336928
   0.11893519972093
  -0.10374750044163
  -0.18482692753230
   0.07438656499821
  -0.08903765307501
   0.42437855911845
   0.05196253426788
  -0.00989189131866
   0.02365478469217
  -0.18333350320319
  -0.03623494152043
  -0.18853273217238
  -0.07454391536612
   0.00086237172915
  -0.04354373216678
   0.16239920834448
  -0.04238573052003
  -0.12702571813465
  -0.07506563910853
  -0.05340187594706
  -0.05332204597203
  -0.02309882687308
   0.06896556993335
   0.02035859409890
   0.09549555064642
   0.03970889360856
   0.07348600829793
  -0.03340358638977
   0.05821064705870
  -0.02455307212971
   0.01744360586450
  -0.07743578926800
  -0.37427473189553
  -0.09945942738271
  -0.09891392735489
  -0.07233503882054
  -0.07620228527653
  -0.15498371786568
  -0.05422179392841
  -0.36929088055010
   0.00334854279089
   0.01494115844809
   0.08887554934827
  -0.00193609100698
  -0.00567903126387
  -0.05669874025119
   0.06509591849070
  -0.02914429947808
  -0.04020831041030
   0.01280490368444
   0.04121152587341
  -0.01875684277602
   0.20258803656317
   0.05729862864947
  -0.08580994850542
   0.00623483773136
   0.07293101951889
   0.00790014416087
   0.09290840614391
  -0.11451813827132
   0.00095762228734
   0.08491135159753


exitflag =

     3


output =
```

```
    firstorderopt: 3.253484379572846e-06
       iterations: 29
        funcCount: 570
     cgiterations: 243
        algorithm: 'large-scale: trust-region reflective Newton'
          message: [1x87 char]


Param =

  18.54647095176016  23.32337655232970
  -0.67312758845460  -1.25936503981221
   1.48491721689638   1.19031263618303
 -15.46732577325159  -9.77159165278830
   3.57843783810543   3.70980543978095
   4.55125379009545   4.90340625220429
  -0.99036067733401  -1.05301468060386
  -8.20131095208484  -6.97975645980404
   0.05712802821842   0.05225717396115
   1.68563900562438   1.57264571873312
  -0.00096502914294  -0.00117462388857
  -0.00576779400417  11.36454165681105
   0.03416326819399   0.02628214348797
  -0.00882706773381  -0.01900506111183
  -0.03675360885525   0.01252267186125
   0.30148575065305   0.56028855564767
 -10.65765272071565  -9.24808297862453
  -6.11294707198806  -5.81129581562833


resnorm =

   0.94055734963692


Residual_absolute =

   0.42437855911845
--------------------------------------------------------------------------------
```

Table SI2. DRAGON Molecular descriptors included on the MLR PMs and used in the MOOP process.

Compound ID	Molecular Descriptors								
	JGI6	MATS3e	GATS5p	FDI	Mor24v	H6v	R4e+	R5p	G(F..F)
004-4-Ciprofloxacin	0.037	-0.027	2.289	0.995	0.137	0.128	0.194	0.409	0.000
006-6-Tosufloxacin	0.045	-0.040	2.002	0.978	0.227	0.116	0.204	0.396	20.920
007-7-PD117558	0.037	0.060	2.233	0.987	0.315	0.172	0.184	0.491	4.760
008-8	0.030	0.061	2.316	0.981	0.230	0.168	0.185	0.466	4.760
010-10	0.048	0.037	2.321	0.981	0.271	0.174	0.177	0.439	39.050
012-13	0.036	0.080	2.287	0.968	0.280	0.190	0.171	0.464	4.760
014-15	0.027	-0.004	2.304	0.967	0.228	0.200	0.178	0.472	0.000
015-16	0.034	-0.005	2.216	0.984	0.225	0.241	0.182	0.490	0.000
016-17	0.044	-0.026	1.985	0.955	0.285	0.256	0.160	0.481	22.080
018-19	0.028	0.039	2.486	0.996	0.131	0.096	0.204	0.425	4.780
019-20	0.028	0.028	2.377	1.000	0.182	0.093	0.212	0.401	4.760
020-21	0.036	0.039	2.377	1.000	0.333	0.101	0.201	0.420	4.760
021-22	0.041	0.016	2.362	0.980	0.203	0.118	0.194	0.460	4.780
022-23A	0.034	-0.012	2.394	1.000	0.222	0.071	0.229	0.442	0.000
023-23B	0.036	0.054	2.417	1.000	0.332	0.090	0.211	0.416	4.760
024-23C	0.036	0.037	2.665	0.997	0.196	0.125	0.221	0.525	0.000
025-23D	0.041	0.046	2.297	0.966	0.196	0.179	0.206	0.536	0.000
026-23E	0.051	-0.002	2.753	0.994	0.096	0.227	0.207	0.586	16.720
027-23F	0.034	0.034	2.267	0.998	0.255	0.083	0.208	0.392	0.000
028-24A	0.034	-0.025	2.355	0.989	0.304	0.113	0.217	0.453	0.000
029-24C	0.036	0.021	2.610	1.000	0.270	0.180	0.215	0.533	0.000
030-24D	0.043	0.031	2.273	0.966	0.162	0.177	0.192	0.523	0.000
031-24E	0.055	-0.015	2.723	0.990	0.189	0.290	0.193	0.566	16.420
032-24F	0.034	0.017	2.242	0.994	0.273	0.101	0.199	0.397	0.000
033-25A	0.035	-0.003	2.144	0.980	0.260	0.131	0.230	0.476	0.000
034-25B	0.038	0.061	2.165	0.997	0.316	0.117	0.200	0.455	4.760
036-25D	0.042	0.052	2.100	0.960	0.158	0.176	0.178	0.537	0.000
037-25E	0.055	0.006	2.509	0.979	0.233	0.321	0.187	0.582	16.430
038-25F	0.035	0.041	2.051	0.988	0.216	0.133	0.189	0.432	0.000
040-26D	0.041	0.052	2.170	0.959	0.268	0.333	0.177	0.569	0.000
041-26E	0.053	0.004	2.532	0.976	0.180	0.405	0.173	0.605	16.420
042-26F	0.034	0.041	2.147	0.980	0.261	0.247	0.170	0.459	0.000
043-27A	0.036	-0.004	1.940	0.962	0.362	0.153	0.268	0.514	0.000
044-27B	0.040	0.059	1.960	0.985	0.400	0.123	0.195	0.500	4.760
045-27C	0.040	0.043	2.143	0.989	0.268	0.170	0.199	0.563	0.000
046-27D	0.045	0.051	1.925	0.949	0.182	0.177	0.188	0.555	0.000
047-27E	0.060	0.003	2.302	0.967	0.268	0.315	0.203	0.583	16.520
048-27F	0.036	0.040	1.879	0.973	0.253	0.157	0.226	0.450	0.000
049-28A	0.035	-0.001	2.273	0.958	0.425	0.209	0.219	0.551	0.000
050-28B	0.038	0.063	2.289	0.982	0.485	0.182	0.175	0.527	4.760
051-28C	0.038	0.047	2.441	0.986	0.402	0.222	0.173	0.575	0.000
052-28D	0.044	0.056	2.225	0.951	0.306	0.266	0.179	0.584	0.000
054-28F	0.035	0.045	2.259	0.966	0.400	0.204	0.233	0.505	0.000
055-29B	0.040	0.037	2.310	1.000	0.282	0.119	0.215	0.429	4.780
056-29C	0.040	0.019	2.530	1.000	0.251	0.157	0.217	0.539	0.000
057-29D	0.043	0.029	2.219	0.969	0.204	0.189	0.189	0.512	0.000
058-29E	0.052	-0.017	2.653	0.955	0.255	0.209	0.200	0.646	16.600
059-29F	0.037	0.015	2.136	0.991	0.222	0.118	0.193	0.391	0.000
061-30B	0.039	0.020	2.473	0.996	0.334	0.204	0.195	0.464	4.780
062-30C	0.039	0.000	2.631	0.999	0.237	0.202	0.198	0.563	0.000
063-30D	0.045	0.012	2.384	0.967	0.249	0.258	0.175	0.533	0.000
064-30E	0.058	-0.032	2.742	0.980	0.273	0.305	0.196	0.591	16.660
065-30F	0.035	-0.005	2.452	0.984	0.263	0.188	0.180	0.436	0.000
066-31A	0.040	-0.030	2.080	0.990	0.158	0.170	0.192	0.422	0.000
067-31B	0.045	0.036	2.101	0.997	0.265	0.136	0.206	0.429	4.780
068-31C	0.045	0.017	2.305	1.000	0.186	0.171	0.207	0.536	0.000
070-31E	0.054	-0.013	2.453	0.982	0.072	0.259	0.212	0.570	16.640
071-31F	0.040	0.012	1.945	0.986	0.209	0.139	0.188	0.399	0.000
073-32B	0.039	0.052	2.256	0.994	0.262	0.148	0.203	0.436	4.780
074-32C	0.039	0.036	2.457	1.000	0.223	0.196	0.210	0.543	0.000
075-32D	0.044	0.045	2.179	0.963	0.147	0.205	0.182	0.520	0.000
077-32F	0.036	0.033	2.140	0.983	0.201	0.135	0.180	0.405	0.000
078-33B	0.047	-0.033	2.276	0.998	0.333	0.162	0.180	0.439	20.730
079-34B	0.066	-0.049	2.417	0.987	0.053	0.292	0.184	0.516	66.290
080-35B	0.051	-0.025	2.013	0.973	0.166	0.228	0.192	0.479	20.120
081-36B	0.044	-0.026	1.985	0.950	0.306	0.284	0.166	0.495	21.580
082-37B	0.047	0.002	2.134	0.986	0.302	0.280	0.165	0.494	20.730

Table SI2. (*Continued…*)

Compound ID	Molecular Descriptors								
	JGI6	MATS3e	GATS5p	FDI	Mor24v	H6v	R4e+	R5p	G(F..F)
083-38A	0.045	0.008	2.152	0.956	0.224	0.163	0.189	0.518	0.000
084-38B	0.053	-0.040	1.913	0.962	0.354	0.214	0.173	0.470	20.340
085-39A	0.060	-0.034	2.580	0.987	0.392	0.300	0.205	0.554	16.690
086-39B	0.070	-0.063	2.297	0.970	0.378	0.296	0.173	0.519	69.030
088-41A	0.039	0.009	2.486	1.000	0.240	0.181	0.192	0.532	0.000
090-42A	0.055	-0.025	2.613	0.976	0.287	0.298	0.190	0.608	16.490
092-48	0.048	0.000	2.458	0.984	0.251	0.253	0.202	0.563	0.000
093-49	0.036	0.011	2.511	0.995	0.243	0.222	0.206	0.473	0.000
094-50	0.043	-0.002	2.804	0.980	0.390	0.191	0.192	0.611	0.000
095-51	0.039	0.028	2.277	0.988	0.288	0.155	0.215	0.438	0.000
096-52	0.039	-0.008	2.170	0.989	0.233	0.138	0.175	0.442	0.000
098-54	0.039	0.057	2.328	0.991	0.255	0.110	0.217	0.430	0.000
100-56	0.041	0.100	2.395	1.000	0.209	0.106	0.234	0.432	12.870
101-57	0.059	0.040	2.740	0.992	0.202	0.291	0.228	0.568	37.570
102-58	0.041	0.077	2.301	0.998	0.274	0.132	0.174	0.423	4.770
103-59	0.041	0.067	2.469	1.000	0.190	0.162	0.180	0.501	0.000
104-60	0.047	0.076	2.213	0.968	0.111	0.228	0.171	0.501	0.000
105-61	0.059	-0.007	2.471	0.981	0.083	0.280	0.169	0.593	16.280
106-62	0.041	0.053	2.198	0.996	0.345	0.110	0.188	0.433	4.750
107-63	0.041	0.094	2.351	0.998	0.359	0.108	0.223	0.425	4.760
110-70	0.047	0.022	2.309	0.972	0.328	0.337	0.199	0.603	0.000
111-71	0.037	0.031	2.353	0.986	0.235	0.258	0.204	0.525	0.000
112-72	0.038	0.047	2.167	0.966	0.341	0.259	0.165	0.483	0.000
113-73	0.038	0.010	2.085	0.967	0.279	0.277	0.164	0.497	0.000
114-74	0.041	0.123	2.247	0.979	0.334	0.231	0.211	0.485	12.890
115-75	0.044	0.105	2.248	0.975	0.298	0.238	0.130	0.458	4.760
117-77	0.041	0.072	2.107	0.992	0.355	0.277	0.160	0.475	4.740
118-78	0.041	0.103	2.109	0.979	0.213	0.297	0.224	0.637	0.000

Table SI2. (*Continued…*)

Compound ID	Molecular Descriptors								
	H4m	D/Dr06	BELp1	RDF020e	RDF050e	Mor05m	Mor14v	HATS3m	HATS3e
004-4-Ciprofloxacin	0.691	182.617	2.048	4.730	32.161	-4.719	0.234	0.075	0.390
006-6-Tosufloxacin	0.974	220.699	2.059	3.874	28.138	-5.259	0.012	0.170	0.486
007-7-PD117558	0.873	141.528	2.036	5.459	32.393	-5.570	0.653	0.083	0.314
008-8	0.874	137.322	2.044	5.757	35.175	-5.278	0.599	0.085	0.308
010-10	1.032	156.391	2.044	6.272	33.234	-5.141	0.454	0.100	0.308
012-13	0.931	148.606	2.042	5.644	37.942	-5.536	0.661	0.078	0.302
014-15	0.846	132.242	2.043	5.807	36.196	-4.530	0.803	0.079	0.344
015-16	0.817	136.448	2.035	6.964	32.073	-4.825	0.790	0.073	0.334
016-17	1.088	254.343	2.062	5.472	37.253	-6.206	0.198	0.128	0.370
018-19	0.720	114.921	2.044	4.608	28.955	-4.754	0.411	0.101	0.419
019-20	0.744	114.921	2.042	4.526	26.332	-5.070	0.128	0.106	0.396
020-21	0.733	119.127	2.036	4.190	25.320	-5.217	0.297	0.091	0.382
021-22	0.723	124.911	2.038	4.374	29.643	-5.191	0.577	0.094	0.364
022-23A	0.654	107.306	2.035	2.495	21.298	-3.203	0.165	0.102	0.451
023-23B	0.726	112.386	2.036	2.580	26.162	-5.086	0.212	0.083	0.417
024-23C	0.749	112.386	2.025	2.706	25.300	-4.462	0.380	0.102	0.425
025-23D	0.735	118.169	2.043	5.121	31.500	-4.819	0.362	0.087	0.321
026-23E	0.934	129.737	2.049	3.662	33.347	-5.246	0.242	0.086	0.330
027-23F	0.663	107.306	2.030	2.115	26.728	-4.550	0.238	0.082	0.427
028-24A	0.722	114.048	2.035	4.177	28.364	-4.586	0.395	0.096	0.431
029-24C	0.864	119.127	2.025	4.163	24.311	-4.906	0.403	0.102	0.394
030-24D	0.758	124.911	2.043	6.977	34.572	-5.034	0.444	0.085	0.304
031-24E	0.776	136.479	2.049	5.074	35.073	-5.781	0.447	0.086	0.314
032-24F	0.692	114.048	2.031	3.724	26.583	-4.504	0.323	0.085	0.416
033-25A	0.786	121.189	2.035	4.599	29.257	-4.736	0.591	0.092	0.399
034-25B	0.735	126.269	2.036	6.110	33.345	-5.166	0.668	0.091	0.371
036-25D	0.782	132.052	2.043	7.442	36.532	-5.172	0.480	0.080	0.288
037-25E	0.844	143.620	2.049	5.610	36.310	-5.991	0.589	0.085	0.308
038-25F	0.763	121.189	2.031	4.472	28.013	-4.630	0.396	0.083	0.380
040-26D	0.880	147.312	2.044	8.087	43.557	-5.079	0.699	0.073	0.280
041-26E	0.915	158.879	2.049	8.018	39.534	-6.002	0.868	0.079	0.281
042-26F	0.832	136.448	2.031	5.370	35.911	-5.037	0.651	0.071	0.349
043-27A	0.869	127.931	2.035	5.324	31.793	-5.040	0.831	0.103	0.358
044-27B	0.806	133.010	2.036	4.987	31.548	-5.716	0.647	0.096	0.318
045-27C	0.854	133.010	2.026	5.150	29.781	-5.948	0.604	0.100	0.319
046-27D	0.848	138.794	2.044	7.645	43.187	-5.360	0.696	0.086	0.278
047-27E	0.862	150.362	2.049	6.183	40.217	-5.132	0.649	0.094	0.300
048-27F	0.819	127.931	2.031	4.823	26.637	-4.866	0.645	0.090	0.329
049-28A	0.915	135.413	2.035	5.623	38.917	-5.125	0.797	0.090	0.308
050-28B	0.844	140.493	2.036	5.651	38.251	-5.946	0.707	0.084	0.289
051-28C	0.894	140.493	2.026	5.502	31.990	-4.735	0.741	0.082	0.294
052-28D	0.877	146.276	2.044	8.216	41.155	-5.779	0.884	0.074	0.242
054-28F	0.856	135.413	2.031	5.206	35.823	-5.152	0.749	0.082	0.283
055-29B	0.705	189.565	2.049	4.676	33.442	-5.389	0.196	0.084	0.389
056-29C	0.759	189.565	2.040	4.580	25.428	-5.467	0.458	0.098	0.395
057-29D	0.741	197.487	2.056	6.911	40.126	-4.825	0.154	0.074	0.299
058-29E	1.040	213.331	2.061	6.230	45.272	-5.408	0.080	0.115	0.345
059-29F	0.696	182.617	2.045	4.608	28.411	-4.773	0.196	0.076	0.395
061-30B	0.716	209.628	2.054	4.008	43.856	-5.687	0.282	0.076	0.348
062-30C	0.824	209.628	2.045	4.234	34.783	-6.320	0.636	0.086	0.302
063-30D	0.770	217.550	2.061	6.284	46.849	-5.038	0.225	0.067	0.258
064-30E	0.827	233.394	2.065	5.631	48.951	-5.630	0.400	0.081	0.277
065-30F	0.726	202.680	2.050	3.903	39.993	-4.984	0.364	0.070	0.336
066-31A	0.734	194.300	2.051	4.223	29.342	-5.033	0.392	0.079	0.412
067-31B	0.738	201.248	2.052	4.278	33.244	-5.551	0.327	0.084	0.385
068-31C	0.847	201.248	2.042	4.325	27.359	-6.145	0.592	0.095	0.386
070-31E	0.800	225.014	2.063	5.595	37.893	-5.451	0.361	0.091	0.322
071-31F	0.749	194.300	2.047	4.034	28.410	-4.913	0.388	0.078	0.396
073-32B	0.714	199.596	2.052	4.191	36.580	-5.517	0.330	0.083	0.390
074-32C	0.789	199.596	2.043	4.008	26.713	-5.745	0.689	0.099	0.370
075-32D	0.762	207.518	2.059	6.652	44.928	-5.331	0.352	0.072	0.275
077-32F	0.726	192.649	2.047	3.998	31.684	-4.750	0.417	0.076	0.387
078-33B	1.035	228.241	2.055	4.126	21.551	-5.751	-0.099	0.149	0.433
079-34B	1.088	253.720	2.072	5.412	43.502	-7.324	-0.277	0.137	0.392
080-35B	1.057	236.734	2.068	6.985	31.577	-6.383	-0.184	0.135	0.374
081-36B	1.109	254.343	2.062	5.652	32.319	-6.013	0.361	0.130	0.374
082-37B	1.192	261.884	2.055	5.692	31.796	-5.988	0.111	0.132	0.369

Table SI2. (*Continued…*)

Compound ID	Molecular Descriptors								
	H4m	D/Dr06	BELp1	RDF020e	RDF050e	Mor05m	Mor14v	HATS3m	HATS3e
083-38A	0.805	131.652	2.044	6.670	35.691	-5.324	0.561	0.088	0.298
084-38B	1.095	246.955	2.068	5.989	42.901	-6.381	-0.072	0.131	0.348
085-39A	0.883	143.220	2.049	5.016	37.750	-4.892	0.622	0.086	0.303
086-39B	1.114	263.941	2.072	5.267	42.242	-6.619	0.256	0.137	0.359
088-41A	0.782	199.596	2.042	4.319	29.217	-5.956	0.474	0.088	0.353
090-42A	0.885	223.362	2.063	5.788	43.829	-6.136	0.450	0.087	0.300
092-48	0.807	130.695	2.061	4.191	36.259	-6.211	0.411	0.101	0.316
093-49	0.685	119.127	2.046	5.613	28.141	-4.321	0.308	0.087	0.376
094-50	0.750	124.911	2.026	4.092	31.555	-4.600	0.041	0.083	0.269
095-51	0.717	119.127	2.046	5.838	29.409	-4.888	0.264	0.086	0.399
096-52	0.573	119.127	2.047	4.192	31.748	-4.788	0.431	0.081	0.352
098-54	0.772	119.127	2.040	4.255	27.626	-4.817	0.210	0.107	0.394
100-56	0.923	124.206	2.037	4.038	28.204	-5.414	0.327	0.128	0.398
101-57	0.999	141.558	2.050	4.902	31.645	-5.750	0.346	0.110	0.314
102-58	0.709	124.206	2.047	5.893	30.322	-5.447	0.301	0.087	0.351
103-59	0.805	124.206	2.037	5.893	27.406	-5.975	0.386	0.090	0.341
104-60	0.709	129.990	2.054	8.739	37.244	-5.087	0.505	0.080	0.282
105-61	0.893	141.558	2.060	5.284	42.772	-6.601	0.376	0.081	0.274
106-62	0.607	124.206	2.048	4.561	31.088	-5.478	0.512	0.080	0.333
107-63	0.764	124.206	2.041	4.357	29.974	-5.475	0.397	0.111	0.384
110-70	1.035	153.096	2.061	5.461	44.661	-6.821	0.834	0.097	0.295
111-71	0.851	141.528	2.046	6.752	38.115	-4.354	0.649	0.083	0.294
112-72	0.899	141.528	2.046	6.961	37.863	-5.084	0.556	0.074	0.310
113-73	0.822	141.528	2.047	5.774	39.596	-5.054	0.712	0.068	0.298
114-74	1.059	146.607	2.037	5.118	38.075	-5.986	0.676	0.116	0.344
115-75	0.828	152.391	2.051	7.321	41.061	-5.920	0.809	0.062	0.260
117-77	0.716	146.607	2.048	7.214	39.830	-5.712	0.909	0.067	0.285
118-78	1.506	146.607	2.017	4.969	33.240	-5.725	0.657	0.191	0.331

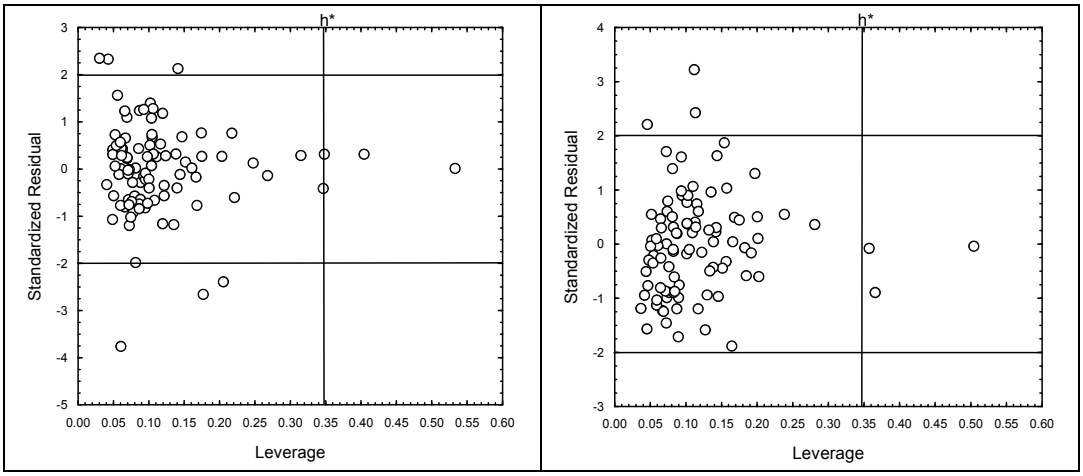


Figure SI1. Applicability domain of the respective MLR PMs

Table SI3. Observed and predicted values of **1/1+IC$_{50}$** and **1/1+MIC**, standardized residual and leverage values of the 95 fluoroquinolones used in this work.

Compound ID	Obs. and Pred. Properties				Std. Residual and Leverage			
	1/1+IC$_{50}$	Pred. 1/1+IC$_{50}$	1/1+MIC	Pred. 1/1+MIC	Std. Res. 1/1+IC$_{50}$	Leverage 1/1+IC$_{50}$	Std. Res. 1/1+MIC	Leverage 1/1+MIC
004-4-Ciprofloxacin	0.003	-0.010	0.909	0.908	0.899	0.095	0.010	0.060
006-6-Tosufloxacin	0.008	-0.006	0.917	0.931	0.960	0.135	-0.136	0.268
007-7-PD117558	0.083	0.052	0.917	0.693	2.212	0.046	2.330	0.043
008-8	0.006	0.017	0.833	0.607	-0.756	0.090	2.352	0.030
010-10	0.017	0.021	0.355	0.281	-0.320	0.156	0.764	0.217
012-13	0.004	-0.002	0.193	0.555	0.407	0.113	-3.758	0.060
014-15	0.003	-0.011	0.641	0.576	1.037	0.157	0.680	0.104
015-16	0.006	0.020	0.685	0.764	-0.988	0.089	-0.822	0.095
016-17	0.005	0.007	0.556	0.644	-0.150	0.122	-0.916	0.078
018-19	0.003	0.003	0.893	0.891	0.047	0.138	0.020	0.081
019-20	0.003	0.006	0.885	0.947	-0.178	0.101	-0.648	0.071
020-21	0.032	0.031	0.962	0.891	0.070	0.052	0.732	0.052
021-22	0.006	0.009	0.769	0.872	-0.212	0.055	-1.069	0.049
022-23A	0.026	0.021	0.833	0.807	0.317	0.082	0.272	0.203
023-23B	0.008	0.026	0.909	0.795	-1.226	0.067	1.180	0.120
024-23C	0.007	0.015	0.909	0.936	-0.606	0.084	-0.285	0.088
025-23D	0.007	0.027	0.769	0.780	-1.457	0.072	-0.115	0.058
026-23E	0.004	-0.007	0.074	0.304	0.747	0.116	-2.388	0.205
027-23F	0.017	0.021	0.794	0.905	-0.293	0.048	-1.157	0.120
028-24A	0.014	0.014	0.917	0.973	0.007	0.073	-0.573	0.080
029-24C	0.037	0.013	0.971	0.865	1.711	0.072	1.096	0.069
030-24D	0.022	0.015	0.935	0.893	0.466	0.064	0.429	0.063
031-24E	0.003	0.001	0.833	0.683	0.196	0.087	1.564	0.056
032-24F	0.010	0.017	0.971	0.944	-0.507	0.044	0.282	0.065
033-25A	0.012	0.026	0.833	0.827	-0.989	0.073	0.063	0.053
034-25B	0.083	0.075	0.952	1.016	0.609	0.074	-0.659	0.108
036-25D	0.091	0.045	0.901	0.879	3.222	0.112	0.223	0.069
037-25E	0.026	0.021	0.658	0.618	0.303	0.065	0.412	0.049
038-25F	0.019	0.041	0.877	0.848	-1.565	0.045	0.306	0.049
040-26D	0.043	0.028	0.794	0.745	1.069	0.110	0.504	0.102
041-26E	0.008	0.005	0.625	0.600	0.230	0.142	0.264	0.111
042-26F	0.006	0.019	0.826	0.795	-0.896	0.077	0.323	0.106
043-27A	0.042	0.037	0.658	0.773	0.361	0.281	-1.198	0.072
044-27B	0.111	0.112	0.885	0.843	-0.070	0.183	0.440	0.061
045-27C	0.111	0.088	0.935	0.989	1.632	0.144	-0.564	0.122
046-27D	0.026	0.052	0.794	0.855	-1.881	0.164	-0.640	0.088
047-27E	0.009	0.026	0.500	0.598	-1.195	0.117	-1.015	0.074
048-27F	0.038	0.036	0.741	0.717	0.209	0.109	0.243	0.069
049-28A	0.005	0.018	0.714	0.687	-0.964	0.145	0.287	0.061
050-28B	0.111	0.085	0.813	0.823	1.870	0.154	-0.100	0.070
051-28C	0.042	0.064	0.794	0.659	-1.581	0.127	1.401	0.102
052-28D	0.008	0.032	0.658	0.729	-1.710	0.089	-0.736	0.086
054-28F	0.017	0.013	0.625	0.702	0.304	0.142	-0.802	0.066
055-29B	0.021	0.032	0.935	1.009	-0.764	0.047	-0.768	0.060
056-29C	0.023	0.039	0.935	1.016	-1.128	0.059	-0.842	0.086
057-29D	0.012	0.025	0.935	0.883	-0.944	0.042	0.532	0.117
058-29E	0.006	-0.002	0.870	0.664	0.552	0.238	2.133	0.141
059-29F	0.008	0.013	0.917	0.919	-0.347	0.054	-0.013	0.072
061-30B	0.007	0.021	0.952	0.938	-1.034	0.059	0.148	0.152
062-30C	0.007	0.024	0.813	0.824	-1.193	0.087	-0.113	0.144
063-30D	0.002	0.002	0.746	0.744	-0.026	0.061	0.023	0.161
064-30E	0.002	-0.007	0.524	0.637	0.608	0.117	-1.178	0.135
065-30F	0.004	-0.019	0.855	0.784	1.610	0.094	0.738	0.104
066-31A	0.004	0.006	0.794	0.808	-0.099	0.105	-0.152	0.093
067-31B	0.042	0.044	0.833	0.888	-0.134	0.082	-0.567	0.050
068-31C	0.053	0.042	0.926	0.898	0.772	0.101	0.285	0.124
070-31E	0.048	0.013	0.794	0.674	2.426	0.113	1.240	0.087
071-31F	0.010	0.023	0.813	0.771	-0.895	0.083	0.433	0.086
073-32B	0.019	0.035	0.885	0.951	-1.187	0.037	-0.688	0.076
074-32C	0.040	0.044	0.935	0.869	-0.254	0.065	0.685	0.147
075-32D	0.014	0.020	0.813	0.834	-0.413	0.076	-0.221	0.095
077-32F	0.010	0.010	0.714	0.787	-0.038	0.050	-0.755	0.072
078-33B	0.011	0.010	0.813	0.739	0.047	0.166	0.769	0.174
079-34B	0.010	0.023	0.658	0.628	-0.895	0.366	0.312	0.404
080-35B	0.003	-0.002	0.741	0.799	0.374	0.102	-0.603	0.221
081-36B	0.005	0.002	0.556	0.525	0.259	0.132	0.319	0.138
082-37B	0.008	0.014	0.488	0.562	-0.442	0.151	-0.771	0.168

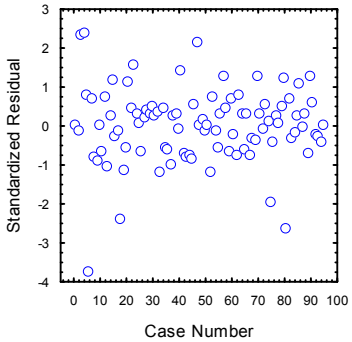
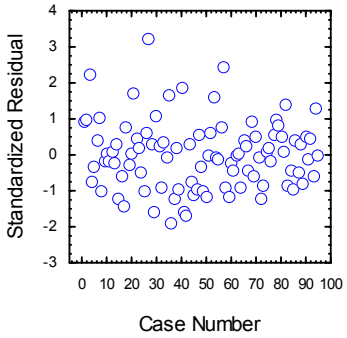
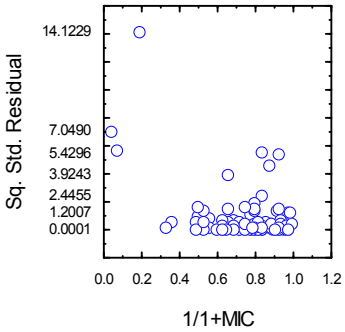
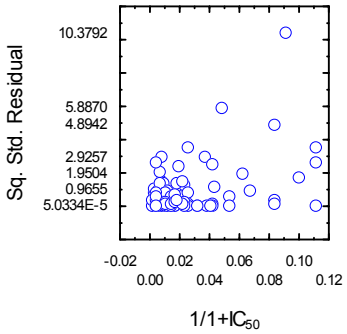
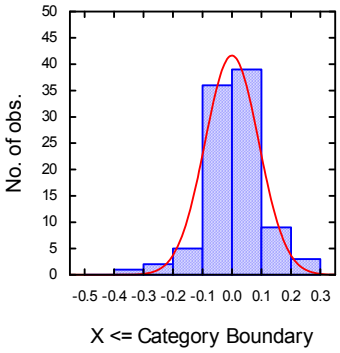
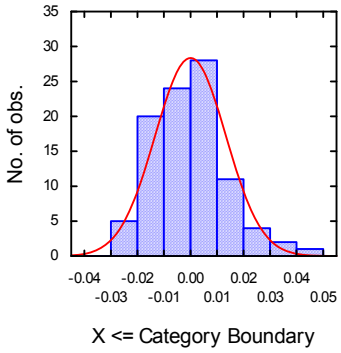
Table SI3. (*Continued…*)

Compound ID	Obs. and Pred. Properties				Std. Residual and Leverage			
	1/1+IC_{50}	Pred. 1/1+IC_{50}	1/1+MIC	Pred. 1/1+MIC	Std. Res. 1/1+IC_{50}	Leverage 1/1+IC_{50}	Std. Res. 1/1+MIC	Leverage 1/1+MIC
083-38A	0.026	0.013	0.794	0.826	0.901	0.103	-0.332	0.041
084-38B	0.004	0.012	0.685	0.723	-0.582	0.185	-0.399	0.140
085-39A	0.009	0.002	0.500	0.376	0.505	0.200	1.285	0.106
086-39B	0.053	0.054	0.326	0.296	-0.075	0.357	0.313	0.348
088-41A	0.022	0.039	0.926	0.934	-1.246	0.068	-0.087	0.095
090-42A	0.005	0.017	0.685	0.634	-0.868	0.071	0.534	0.057
092-48	0.014	0.013	0.685	0.673	0.104	0.058	0.128	0.248
093-49	0.004	0.001	0.654	0.844	0.204	0.087	-1.981	0.081
094-50	0.031	0.034	0.833	0.873	-0.160	0.192	-0.409	0.346
095-51	0.018	0.010	0.962	0.936	0.550	0.052	0.265	0.098
096-52	0.067	0.053	0.917	0.910	0.983	0.093	0.075	0.104
098-54	0.014	0.002	0.962	0.913	0.795	0.075	0.499	0.054
100-56	0.010	0.003	0.926	0.807	0.495	0.168	1.233	0.066
101-57	0.005	0.004	0.038	0.294	0.105	0.201	-2.655	0.177
102-58	0.063	0.043	0.990	0.926	1.397	0.081	0.661	0.067
103-59	0.017	0.029	0.926	0.960	-0.871	0.084	-0.352	0.122
104-60	0.010	0.016	0.901	0.917	-0.428	0.138	-0.167	0.167
105-61	0.003	0.017	0.524	0.498	-0.937	0.130	0.269	0.175
106-62	0.083	0.078	0.980	0.877	0.385	0.101	1.078	0.104
107-63	0.023	0.030	0.971	0.973	-0.492	0.134	-0.025	0.071
110-70	0.015	0.010	0.488	0.460	0.316	0.114	0.289	0.315
111-71	0.003	0.015	0.524	0.593	-0.806	0.064	-0.724	0.098
112-72	0.016	0.009	0.741	0.619	0.508	0.081	1.267	0.093
113-73	0.023	0.025	0.625	0.570	-0.101	0.082	0.572	0.060
114-74	0.021	0.015	0.641	0.661	0.446	0.175	-0.211	0.100
115-75	0.019	0.027	0.592	0.619	-0.596	0.203	-0.286	0.077
117-77	0.100	0.082	0.781	0.820	1.305	0.197	-0.402	0.100
118-78	0.004	0.004	0.625	0.623	-0.037	0.504	0.017	0.533

This section provides details about the checking of the pre-adopted parametric assumptions, a very important aspect in the application of linear multivariate statistical-based approaches (MLR techniques) (1). In fact, once the linear regression model has been set up, it is very important to check the parametric assumptions to assure the validity of extrapolation from the sample to the population. These include the linearity of the modeled property, normal distribution as well as the homoscedasticity and non-multicollinearity descriptors. Notice that severe violations of one or various of these assumptions can markedly compromise the reliability of the predictions resulting from our MLR models (1).

We first check the linearity hypothesis by looking at the distribution of the standardized residuals for all cases. Indeed the plots in Table SI4 (1st row) do not show any specific pattern, reinforcing the idea that our models do not exhibit a non-linear dependence (1). Next, we check the hypothesis of homoscedasticity (*i.e.*: homogeneity of variance of the variables), which can be confirmed by simply plotting the square of standardized residuals related to each dependent variable (1) (2nd row of plots in Table SI4). These plots reveal significant scatter of points, without any systematic pattern, *post-mortem* validating the pre-adopted assumption of homoscedasticity for all the PMs. They also provide a check for the no auto-correlation of the residuals. Moving on to the hypothesis of normally distributed residuals, one can easily confirm that the residuals follow a normal distribution by applying the Kolmogorov-Smirnov and Lilliefors statistical test (3rd row of Table SI4). In addition, as the term related to the error (represented by residuals) is not included in the MLR equations, the mean must be zero what actually occurs (check 4th row of Table SI4). The last aspect deserving special attention is the degree of multicollinearity among the variables. Highly collinear variables may be identified by examining their pair-correlations (R_{ij}). Most of the predictors included in the models exhibit a value of R_{ij} lower than 0.7. Only a few pair of variables (two in the IC_{50} PM and three in the MIC PM) have values of R_{ij} over 0.7, but no one higher than 0.75 suggesting that the problem of the collinearity is not serious. One should emphasize here that the common interpretation of a regression coefficient as measuring the change in the expected value of the response variable, when the given predictor variable is increased by one unit while all other predictor variables are held constant, is not fully applicable when multicollinearity exists ($R \geq 0.7$) (2).

Table SI4. Checking the main parametric assumptions related to the MLR models used to fit the desirability functions.

	***MIC* MLR Model**	***IC*$_{50}$ MLR Model**
Linearity		
Homoscedasticity		
Normality of Residuals	K-S d = 0.08168, p > 0.20 Lilliefors p> 0.15 	K-S d = 0.07039, p > 0.20 Lilliefors p> .20 
Res. Mean = 0	0.000000	0.000000

References

1. Stewart J, Gill L. Econometrics. 2nd edition ed. Allan P, editor. London: Prentice Hall; 1998.
2. Kutner MH, Nachtsheim CJ, Neter J, Li W. Multicollinearity and its effects. Applied Linear Statistical Models. Fifth ed. New York: McGraw Hill; 2005. p. 278-89.

ANNEX III

molecular informatics

DOI: 10.1002/minf.200900047

Prioritizing Hits with Appropriate Trade-Offs Between HIV-1 Reverse Transcriptase Inhibitory Efficacy and MT4 Blood Cells Toxicity Through Desirability-Based Multiobjective Optimization and Ranking

Maykel Cruz-Monteagudo,*[a, b, c, d] Hai PhamThe,[d] M. Natalia D. S. Cordeiro,*[e] and Fernanda Borges[a]

Abstract: Nonnucleoside reverse transcriptase (RT) inhibitors (NNRTIs) constitute a promising therapeutic option for AIDS. However, the emergence of virus-NNRTIs resistance was found to be a major problem in the field. Toward that goal, a "knock-out" strategy stands out between the several options to circumvent the problem. However the high drug or drug-drug concentrations often used generate additional safety concerns. The need for approaches able to early integrate drug- or lead-likeness, toxicity and bioavailability criteria in the drug discovery phase is an emergent issue. Given that, we propose a combined strategy based on desirability-based multiobjective optimization (MOOP) and ranking for the prioritization of HIV-1 NNRTIs hits with appropriate trade-offs between inhibitory efficacy over the HIV-1 RT and toxic effects over MT4 blood cells. Through the MOOP process, the theoretical levels of the predictive variables required to reach a desirable RT inhibitor candidate with the best possible compromise between efficacy and safety were found. This information is used as a pattern to rank the library of compounds according to a similarity-based structural criterion, providing a ranking quality of 64%/71%/73% in training/validation/test set. A comparative study between the sequential, parallel and multiobjective virtual screening revealed that the multiobjective approach can outperform the other approaches. These results suggest that the identification of NNRTIs hits with appropriate trade-offs between potency and safety, rather than fully optimized hits solely based on potency, can facilitate the hit to lead transition and increase the likelihood of the candidate to evolve into a successful antiretroviral drug.

Keywords: Drug discovery · HIV-1 · NNRTI · Multiobjective optimization · Virtual screening

1 Introduction

Reverse transcriptase (RT) is a key enzyme which plays an essential role in the replication of the human immunodeficiency virus type-1 (HIV-1). RT represents an attractive target for the development of new drugs useful in acquired immune deficiency syndrome (AIDS) therapy.[1] Toward that goal, nonnucleoside RT inhibitors (NNRTIs) are at present a promising option in HIV-1 drug discovery due to their low toxicity profile when compared to the nucleoside analogues.[2] Unlike nucleoside analogues, NNRTIs bind in a noncompetitive manner to a specific 'pocket' of the HIV-1 RT altering its ability to function.[3] NNRTIs selectively inhibit HIV-1 RT replication in cell culture at a concentration notably lower than the required concentration to affect normal cell viability.[4]

Nowadays NNRTIs are considered a promising scaffold for the discovery of a new medicine for the treatment of HIV-1 infections. Even though a great number of the candidates of drug discovery programs present a high selectivity and potency towards HIV-1 replication only four NNRTIs (nevirapine, delavirdine, efavirenz and etravirine) have at

[a] *M. Cruz-Monteagudo, F. Borges*
Department of Chemistry, Faculty of Sciences, University of Porto
4169-007 Porto, Portugal
**e-mail: maikelcm@uclv.edu.cu*
gmailkelcm@yahoo.es

[b] *M. Cruz-Monteagudo*
Department of Organic Chemistry, Faculty of Pharmacy, University of Porto
4150-047 Porto, Portugal

[c] *M. Cruz-Monteagudo*
Applied Chemistry Research Center (CEQA), Faculty of Chemistry and Pharmacy, Central University of "Las Villas"
Santa Clara, 54830, Cuba

[d] *M. Cruz-Monteagudo, H. PhamThe*
Molecular Simulation and Drug Design Group, Chemical Bioactive Center (CBQ), Central University of "Las Villas"
Santa Clara, 54830, Cuba

[e] *M. N. D. S. Cordeiro*
REQUIMTE, Department of Chemistry, Faculty of Sciences, University of Porto
4169-007 Porto, Portugal
**e-mail: ncordeir@fc.up.pt*

Supporting information for this article is available on the WWW under http://dx.doi.org/10.1002/minf.20090047

present been approved for clinical use in the treatment of AIDS.[5]

The virus-drug resistance is considered one of the major drawbacks that compromise the therapeutic usefulness of the NNRTIs.[6] In fact, HIV virus rapidly develops resistance to NNRTIs due to mutagenic processes, mainly located at positions that surround the binding region of NNRTIs to the HIV-1 RT pocket.[6] So, the potential NNRTIs therapeutic value may be assisted by the development of strategic approaches suitable to prevent, circumvent or overcome drug resistance process. Among the different approaches pointed out in the literature the "knock-out" strategy stands out as a very promising one.[4] However, as the strategy involves the administration of high drug or drug-drug concentrations other problems related with their safety profile must be considered in addition to drug selectivity and efficacy requirements.

In fact, this is a particular case of one of the major problems found in drug discovery and development. Really, the need for approaches able to early integrate drug- or lead-likeness, toxicity and bioavailability criteria in the drug discovery phase is an emergent issue.[7] That is, methods that can handle additional criteria for the early simultaneous treatment of the most important properties, potency, safety, and bioavailability, determining the pharmaceutical profile of a drug candidate.[8]

Although "Costs of Goods" has been claimed as one of the major reasons for the end of a R&D project[9] one cannot disregard the idea that toxicity and/or pharmacokinetics profiles of the clinical candidates are still decisive causes of failure in drug development process.[7,10] In fact, the ability to improve the pharmaceutical profile of candidates in lead optimization process on the sole basis of their activity has been often overestimated.[7] The adjustment of the multiple criteria in hit-to-lead identification and lead optimization is considered to be a major advance in the rational drug discovery process. The aim of this paradigm shift is the prompt identification and elimination of candidate molecules that are unlikely to survive later stages of discovery and development. In turn, this new approach will reduce clinical attrition, and as a consequence, the overall cost of the process.[7b, 11]

The virtual screening (VS) techniques currently employed in early stages of drug discovery do not involve multiple criteria assessment (one by one, starting with potency) of the properties that can modulate the success of a drug candidate (potency, safety and bioavailability). Accordingly, numerous failures of the R&D projects have been described and attributed to the reduced toxicological and/or pharmacokinetic outline of drug candidates. Thus, the employment of multiobjective approaches allowing to obtain candidates with acceptable trade-offs between potency, safety and/or bioavailability is an emerging issue in drug discovery and development.[8]

In this paper, we describe the application of multiobjective optimization (MOOP) and ranking methods[8b,d] for simultaneously probe the inhibitory efficacy towards HIV-1 RT, and the toxic effects towards MT4 blood cells, of a diverse set of HIV-1 NNRTIs reported in the literature.[12] This methodology is proposed as a rational strategy of multiobjective VS to identify new HIV-1 NNRTIs hits with acceptable trade-offs between the above mentioned properties. Finally, a retrospective analysis of the training set, based on well-known enrichment measures,[13] will be done allowing to compare the performance of several approaches (sequential, parallel and multiobjective) as VS strategies. The performance of the multiobjective VS strategy to retrieve pharmaceutically acceptable NNRTI candidates from a pool of NNRTI decoys is also tested.

2 Material and Methods

2.1 Data Set

The prediction models (PMs) for inhibitory efficacy over the HIV-1 RT and toxicity over MT4 blood cells, as well as the desirability-based MOOP and ranking process were performed using a library of NNRTIs collected from previous literature reports.[12]

To collect a representative set of NNRTIs, we collect a data base containing four of the most studied chemical families of this class of HIV-1 RT inhibitors. Thus, in the initial pool we have included: 39 1-[(2-Hydroxyethoxy)methyl]-6-(phenylthio) thymine (HEPT) analogues, 25 diaryltriazine (DATA) analogues, 62 acylthiocarbamate (ATC) analogues, and 36 2-alkoxy-3,4-dihydro-6-benzyl-4(3*H*)-pyrimidin-4-ones (DABO) analogues. From the initial pool of 162 compounds, 53 were removed since their property values were inaccurately reported ($<$, $>$ or $\leq$ values). This was done considering that the use of these values can reduce significantly the goodness of fit of a multiple linear regression (MLR) model.

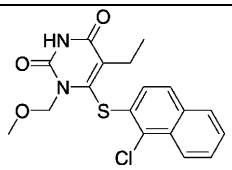
In Table 1, the chemical families included in the data set here employed are depicted. The structural diversity of this set can also be checked in this table.

The remaining set of 109 compounds was randomly split up into training and validation subsets. Approximately 80% of the compounds (88) were used for training whereas the remaining 20% (21) were reserved for validation purposes. Additionally, to test the predictive ability of the trained models on a true test set, we select a subset of the 53 compounds initially excluded from the training or validation sets. Only 18 of such compounds, which were within the applicability domain of both models, were selected for this test set.

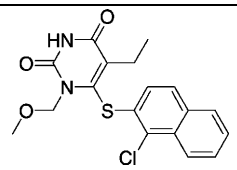
According to the literature,[12] the concentration of a compound required to protect the cell against viral cytopathogenicity by 50% (IC_{50}; measured in μM), as well as its concentration that reduces the normal uninfected cell viability by 50% (CC_{50}; measured in μM) were evaluated against wild-type HIV-1 strain IIIB in MT-4 cells using the 3-(4,5-dimethylthiazol-2-yl)-2, 5-diphenyltetrazolium bromide

Table 1. Classes of NNRTIs included in the data set.

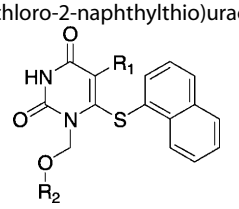
HEPT Analogues



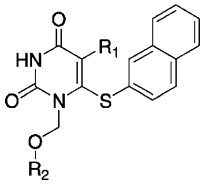
1-Methoxymethyl-5-ethyl-6-(1-chloro-2-naphthylthio)uracil



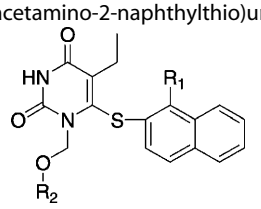
1-[(Benzyloxy)methyl]-5-ethyl-6-(1-acetamino-2-naphthylthio)uracil



A

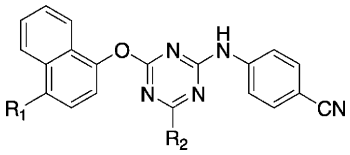


B

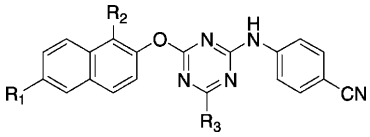


C

DATA Analogues

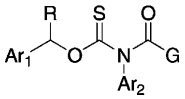


D

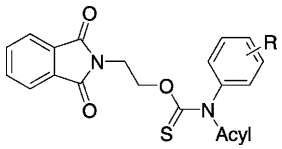


E

ATC Analogues

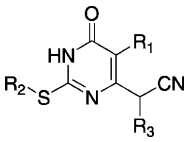


F

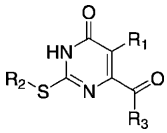


G

DABO Analogues



H



I

A: [R_1 = Me, Et, *i*-Pr, *i*-Bu; R_2 = Me, Et, Benzyl, 3′-Methylbenzyl, 3′-Fluorobenzyl, 4′-Fluorobenzyl, $CH_2CH_2OCH_3$, CH_2CH_2OH, CH_2CH_2OAc, $PhCH_2CH_2$, c-Pr-CH_2, c-Hexl-CH_2]. **B**: [R_1 = Et, *i*-Pr; R_2 = Me, Et, Benzyl]. **C**: [R_1 = Me, Et, Benzyl; R_2 = NH_2, NO_2]. **D**: [R_1 = H, Cl; R_2 = NH_2, NHMe, NHEt, *n*-PrNH]. **E**: [R_1 = H, Cl, Br; R_2 = H, Br; R_3 = N_3, NH_2, NHMe, NHEt, *n*-PrNH, *i*-PrNH]. **F**: [R = H, CH_3; Ar_1 = phenyl, benzyl, phenoxymethyl; Ar_2 = C_6H_5, 4-F-C_6H_5, 4-NO_2-C_6H_5; G-CO = benzoyl, phenoxyacetyl, *trans*-cinnamoyl, 2-furoyl, 2-thenoyl, 4-chlorobenzoyl, 4-chloro-3-nitrobenzoyl, 2,4-dichlorobenzoyl, 3,5-dichlorobenzoyl]. **G**: [R = 3-Br, 3-NO_2, 4-Cl, 4-I, 4-NO_2; Acyl = 2-furoyl, 2-thenoyl]. **H**: [R_1 = H, Me, Et, *i*-Pr; R_2 = Et, *i*-Pr, propynyl, benzoylmethyl, cyclopentyl, 4-methoxybenzyl, 4-nitrobenzyl, 4-chlorobenzoylmethyl; R_3 = Ph, 1-naphthyl, 2,6-Cl_2-Ph]. **I**: [R_1 = H, Me, Et, *i*-Pr; R_2 = Et , *i*-Pr, allyl, cyclopentyl; R_3 = Ph, 1-naphthyl].

MTT method.[14] To ensure proper error rates, the raw IC_{50} and CC_{50} values were log-transformed ($-\log IC_{50}$ and $-\log CC_{50}$) instead. These values for the full set of NNRTIs used for training, validation and test are plotted in Figure 1.

The identification, names, chemical structures, as well as the respective values of IC_{50} and CC_{50} of the full set of 127 compounds used in this work (88 for training, 21 for validation, and 18 for test) can be assessed in the supporting information (see Supporting Informations, Tables SI-1 to SI-5).

2.2 Desirability-Based Multiobjective Optimization and Ranking

A desirability-based methodology was employed here to simultaneously optimize and rank a set of candidates according to their inhibitory efficacy over the HIV-1 RT and, the toxic effects over MT4 blood cells.[8b,d]

First, it is necessary to develop a prediction model for each response. The predicted values of each response are employed to set-up a global prediction model which is fitted to a linear function using the whole subset of inde-

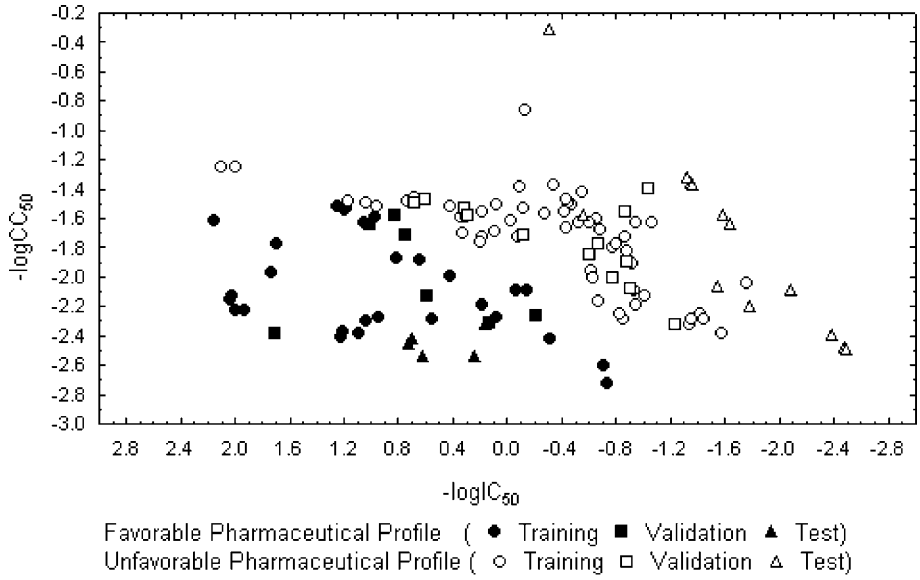


Figure 1. Plot of $-\log IC_{50}$ vs. $-\log CC_{50}$ for the full set of NNRTIs used on the training, validation and test sets.

pendent variables employed in modeling the k original responses.

Next, the predicted responses ($\hat{Y}_i$) are scaled to their respective desirability (d_i) values by means of the Derringer desirability functions.[15] Desirability functions are well-known multicriteria decision-making methods, based on the definition of a desirability function for each response in order to transform values of the responses to the same scale. Each attribute is independently transformed into a desirability value by an arbitrary function. The original value is range scaled between 0 and 1 by:

$$d_i = \frac{\hat{Y}_i - L_i}{U_i - L_i} \quad 0 \le d_1 \le 1 \tag{1}$$

where L_i and U_i are the selected minimum and maximum values, respectively.

Specifically in this work, two desirability functions (one for each response) were fitted. The toxicity over MT4 blood cells ought to be minimized. This property is expressed here through the CC_{50} value. According to the meaning, this value should be maximized in such a way that the compound with the highest CC_{50} value should be the most desirable ($d_i = 1$), but using as input $-\log CC_{50}$, these values most in turn be minimized. For estimating d_i, the target (T_i) lower value L_i was set to $-\log CC_{50} = -2.723$ (i.e., $CC_{50} =$ 529 μM) coinciding with the less toxic compound used for training, and the upper value U_i was set to -0.865 ($CC_{50} =$ 7.32 μM; i.e., the most toxic compound). The desirability function applied to $-\log CC_{50}$ was:

$$d_i = \begin{cases} 1 & \text{if } Y_i \le T_i = L_i \\ \left[\frac{\hat{Y}_i - U_i}{T_i - U_i}\right] & \text{if } U_i < Y_i < T_i \\ 0 & \text{if } Y_i \ge U_i \end{cases} \tag{2}$$

where T_i is interpreted as a small enough $-\log CC_{50}$ value, which can be L_i.

In contrast, the HIV-1 RT inhibitory activity must be maximized. Accordingly, the IC_{50} values should be minimized, but using as input $-\log IC_{50}$, these values must in turn be maximized. In this case, $U_i = T_i = -\log IC_{50} = 2.155$ (i.e., $IC_{50} =$ 0.007 μM) coinciding with the most potent compound used for training, and L_i was set to -1.575 ($IC_{50} = 37.58$ μM, i.e., the less potent compound). The specific desirability function applied was:

$$d_i = \begin{cases} 0 & \text{if } \hat{Y}_i \le L_i \\ \left[\frac{\hat{Y}_i - L_i}{T_i - L_i}\right] & \text{if } L_i < \hat{Y}_i < T_i \\ 1 & \text{if } \hat{Y}_i \ge T_i = U_i \end{cases} \tag{3}$$

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

Once the kind of function for each response is defined, the global desirability D of each i-th candidate can be evaluated as follows:

$$D = (d_1 \times d_2 \times \ldots \times d_k)^{1/k} \tag{4}$$

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.

Finally, the overall desirability D is optimized by using the simplex method.[16] The final result is finding 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.

In this work, the optimization of the overall desirability was carried on by the *Use general function optimization* option[15] of the general regression module of STATISTICA.[17] Furthermore, the spline method[18] 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 towards T_i on the two responses.

The results reached by the MOOP process (levels of descriptors for the optimal candidate) are employed as a template for a ranking algorithm based on quantitative parameters estimated from the description of the cases in order to rank candidates with unknown pharmaceutical profiles.

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

$$\Delta_i = \sum_{X=1}^{m} \Delta_{i,X} \cdot w_X \tag{5}$$

where $\Delta_{i,X}$ is the Euclidean distance between the case i and the optimal case considering the parameter(s) X and, w_X represents the weigh or influence of the variable X over the global desirability D of the case i.

The Δ_i values are normalized by means of the application of the Derringer desirability functions[15] in order to bring it to the same scale of D_i. Like this, it is possible to minimize the difference between the values of Δ_i and D_i for every case.

The weights were obtained through a nonlinear curve-fitting using the large-scale optimization algorithm[19] implemented in the "lsqcurvefit" function of MATLAB program, Version 7.2.[20]

Once minimized the differences between D_i and the normalized values of Δ_i, we achieve a highest possible degree of concordance between the description (normalized values of Δ_i) and the solution of the cases (D_i). Thus, it is possible to rank according to Δ_i new (pharmaceutically unknown) candidates only based on structural information. In this way, it will be possible to filter and identify the most promising candidates which logically will 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 candidate determined by the desirability-based MOOP process) and to discard the rest of the candidates ordered last. The ranking quality index (Ψ) was used to test the reliability of the ranking reached. Ψ encodes the degree of dissimilarity between the real (D-based) and model-based (Δ_i-based) ranking. Ψ takes values in the range of zero (identical real and model-based rankings) to one (totally dissimilar rankings). Details on the validation of the ranking algorithm employed as well as the definition and determination of Ψ can be found in reference.[8d]

2.3 Enrichment Analysis

The main goal in a VS effort is to select a subset from a large pool of compounds (typically a compound database or a virtual library) and try to maximize the number of known actives in this subset. That is, to select the most "enriched" subset as possible. So, in this experiment we are searching for the VS approach able to maximize the number of NNRTI candidates with a pharmaceutical profile equal or superior to 50% ($D_{IC50\text{-}CC50} \geq 0.5$) in a predefined fraction (χ) of the library ($\chi = 0.1 =$ top 10%; first 12 compounds). That is, to include in the top 10% fraction of the ordered library as much candidates as possible exhibiting a favorable compromise between HIV-1 RT inhibitory efficacy and MT4 blood cells toxicity. The experiment is applied to the full set of 122 NNRTIs (83/21/18 from training/validation/test set) containing 41 compounds with a pharmaceutical profile equal or superior to 50%.

The sequential VS is conducted in this work by ranking independently the library of compounds according to the two objectives considered, HIV-1 RT inhibitory efficacy (IC_{50}) and MT4 blood cells toxicity (CC_{50}). The predicted values of IC_{50} and CC_{50} derived from the initial QSAR PMs were the ranking criteria employed. After ranking, a fraction of the library is first filtered according to a predefined threshold value of inhibitory efficacy (inhibitory efficacy profile $\geq 50\%$; $d_{IC50} \geq 0.5$; $-\log IC_{50} \geq 0.196$; $IC_{50} \leq 0.64$ μM). Next, those candidates not fulfilling a predefined threshold value of safety (safety profile $\geq 50\%$; $d_{CC50} \geq 0.5$; $-\log CC_{50} \leq -1.794$; $CC_{50} \geq 62.23$ μM) are eliminated in order to keep those with adequate inhibitory efficacy and safety profiles. In this approach; as well as in the multiobjective one, the true positive fraction (χ_+) can be equal or smaller than the filtered fraction χ (i.e., $0 \leq \chi_+ \leq \chi$).

The parallel VS, as the name implies, is based on running in parallel the independent analysis of the two objectives involved on the pharmaceutical profile of the candidate (IC_{50} and CC_{50}). The conditions in this case are identical to those defined for the sequential approach, but applied in a parallel fashion. In this case, those candidates included in each top 10% filtered fraction, and fulfilling the predefined threshold value for both criteria, are selected. In this case, if the retrieved compounds in both filtered fractions are the same, the retrieved fraction $= \chi = 0.1 = 12$ compounds, otherwise the retrieved fraction $\leq 2\chi$. Consequently, $0 \leq \chi_+ \leq 2\chi$, depending of the efficacy and safety profiles of the candidates filtered in each top 10% filtered fraction.

The multiobjective VS approach proposed in this work considers the pharmaceutical profile of the candidate, rather than separately consider each property related with it. As detailed previously, the overall desirability of the candidate is the criterion employed here to measure their pharmaceutical profile. The library of NNRTIs is ranked according to a structural similarity criterion (Δ_i), top ranking those candidates structurally closer to the previously determined optimal candidate. Like in the sequential and parallel VS approaches, the top 10% of the ordered library is filtered, searching for those candidates with $D_{IC50\text{-}CC50}$ values ≥ 0.5.

Several enrichment metrics have been proposed in the literature to measure the enrichment ability of a VS protocol.[13a,b] In this work, we use some of the most extended.

Based on the analysis of the receiver operating characteristic (ROC) curve it is possible to derive the area under the ROC curve (*ROC Metric*), as well as the ratio of true positive (TP) cases and false positive (FP) cases found at the operating point of the ROC curve ($TP/FP_{\text{ROC-OP}}$).

The ROC curve method describes the sensitivity or TP rate for any possible change of the number of selected cases as a function of (1-Specificity) or FP rate.[13b] So, it is possible to identify the point in the curve with the best possible ratio between TP and FP cases (i.e., $TP/FP_{\text{ROC-OP}}$).[21] That is, the fraction of the library that must be filtered in order to maximize the number of TP cases, minimizing as much as possible the FP cost. On the other hand, the *ROC Metric* can be interpreted as the probability that a positive case will be ranked earlier than a negative one within a rank-ordered list.[13a]

From the accumulation curve we can deduce enrichment from the area under the curve (*AUAC*), from the yield of actives (*Ya*) at certain filtered fractions (5%, 10%, 20% and 50%), as well as from the fraction of the database that has to be screened in order to retrieve a certain percentage (100%) of the TP cases (screening percentage, $\chi_{100\%}$).

The accumulation curve is based on the empirical cumulative distribution function (CDF) where on the abscissa is the relative rank or data fraction, χ, and on the ordinate is the cumulative fractional count of actives retrieved up to χ when the compounds are examined from best to worst according to a scoring or ranking method. So, *AUAC* can be interpreted as the probability that a positive case, selected from the empirical CDF defined by the rank-ordered list, will be ranked before a case randomly selected from a uniform distribution.[13a]

The yield of actives (*Ya*) is one of the most popular descriptors for evaluating VS methods. Defined as the ratio between the number of TP cases and the number of selected cases (n), it quantifies the probability that one of the n selected cases is active. In other words, it represents the hit rate that would be achieved if all cases selected by the VS protocol would be tested for activity. However, it contains no information about the increase of the ratio of TP cases to decoys (FP cases) within a VS case selection compared to a random selection.[13b]

$$Ya = \frac{TP}{n} = \frac{\chi_+}{\chi} \tag{6}$$

On the other hand, the enrichment factor (*EF*) takes into account the improvement of the hit rate by a VS protocol compared to a random selection. This metric has the advantage of answering the question: how enriched in TP cases, the set of k cases that I select for screening will be, compared to the situation where I would just pick the k cases randomly?

$$EF = \frac{TP/n}{N_+/N} \tag{7}$$

where TP and n have been defined previously, and N and N_+ are the total number of cases, and the number of positive cases in the library, respectively. The maximum value that EF can take is $1/\chi$ if $\chi \geq N_+/N$, N/N_+ if $\chi < N_+/N$, and the minimum value is zero.[13a]

2.4 Computational and Statistical Details

Reasonable optimized geometries for all compounds were obtained by resorting to the MM2 molecular mechanics force field [22] implemented in the MOPAC 6.0 program.[23] The optimized structures were then uploaded to the DRAGON software package[24] to compute a total of 1664 molecular descriptors. As part of the necessary variable reduction, descriptors with constant or near constant values and those which were highly pair-correlated ($|R| > 0.95$) were excluded. The variable selection approach used in this work to establish the quantitative structure-acivity relationships (QSAR) models was the Genetic Algorithm (GA)[25] by means of the BuildQSAR software package.[26] Table 2 depicts the DRAGON molecular descriptors selected by the GA method, which were finally applied to model the HIV-1 RT inhibitory activity and toxicity over MT4 blood cells of the library of compounds used in this study.

As 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 using the STATISTICA software package.[17] The goodness of fit for each Predictive Model (PM) was assessed by examining the determination coefficient (R^2_{FIT}), the standard deviation (s_{FIT}), Fisher's statistics (F), and the ratio between the number of compounds (N) and the number of adjustable parameters in the model (p), known as the ρ statistics. The stability of the models was addressed by means of a leave-one-out cross-validation technique over the training set of NNRTIs (R^2_{LOOCV} and s_{LOOCV}).[21,27] The predictive ability was measured by examining the determination coefficient on validation (R^2_{VAL}) and test (R^2_{TEST}) sets, respectively.

Table 2. DRAGON molecular descriptors selected by the GA method.

Symbol	Definition	Class	Type	Property
N-075	R–N–R/R–N–X	Atom-centred fragments	1D	IC_{50}
MAXDP	Maximal electrotopological positive variation	Topological descriptors	2D	IC_{50}
X1sol	Solvation connectivity index chi-1	Conectivity indices	2D	IC_{50}
SIC0	Structural information content (neighborhood symmetry of 0-order)	Information indices	2D	IC_{50}
GATS1p	Geary autocorrelation – lag 1 weighted by atomic polarizabilities	2D Autocorrelations	2D	IC_{50}
Espm15r	Spectral moment 15 from edge adj. matrix weighted by resonance integrals	Edge adjacency indices	2D	IC_{50}
Eig1v	Leading eigenvalue from van der Waals weighted distance matrix	Eigenvalue-based indices	2D	IC_{50}
Ks	K global shape index weighted by atomic electrotopological states	WHIM descriptors	3D	IC_{50}
R8u+	R maximal autocorrelation of lag 8 unweighted	GETAWAY descriptors	3D	IC_{50}
R8m	R autocorrelation of lag 8 weighted by atomic masses	GETAWAY descriptors	3D	IC_{50}
nROH	Number of hydroxyl groups	Functional group counts	1D	CC_{50}
C-003	CHR3	Atom-centred fragments	1D	CC_{50}
MATS3m	Moran autocorrelation – lag 3 weighted by atomic masses	2D Autocorrelations	2D	CC_{50}
MATS5e	Moran autocorrelation – lag 5 weighted by atomic Sanderson electronegativities	2D Autocorrelations	2D	CC_{50}
RDF070p	Radial Distribution Function – 7.0 weighted by atomic polarizabilities	RDF descriptors	3D	CC_{50}
Mor18e	3D-MoRSE - signal 18 weighted by atomic Sanderson electronegativities	3D-Morse descriptors	3D	CC_{50}
H7e	H autocorrelation of lag 7 weighted by atomic Sanderson electronegativities	GETAWAY descriptors	3D	CC_{50}
R8p	R autocorrelation of lag 8 weighted by atomic polarizabilities	GETAWAY descriptors	3D	CC_{50}

The overall desirability determination coefficient for training ($R^2_{FIT,D}$) and leave-one-out cross validation ($R^2_{LOOCV,D}$) were used to test uncertainty in predicting the overall desirability function and the reliability of the simultaneous optimization of the k responses over the independent variables domain, respectively.[8b,d,28] $R^2_{FIT,D}$ and $R^2_{LOOCV,D}$ are defined as follows:

$$R^2_{FIT,D} = 1 - \frac{SSE}{SSTO} = 1 - \frac{\sum (D_{Y_i} - D_{\hat{Y}_i})^2}{\sum (D_{Y_i} - \bar{D}_{Y_i})^2} \tag{8}$$

where D_{Yi} and $D_{\hat{Y}i}$ 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.

$$R^2_{LOOCV,D} = 1 - \frac{SSE_{LOO-CV}}{SSTO} = 1 - \frac{\sum (D_{Y_i} - D_{\hat{Y}_i}(LOO-CV))^2}{\sum (D_{Y_i} - \bar{D}_{Y_i})^2} \tag{9}$$

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

The overall desirability determination coefficient was also determined on the validation ($R^2_{VAL,D}$) and test ($R^2_{TEST,D}$) sets.

3 Results and Discussion

3.1 NNRTIs Multiobjective Virtual Screening via Desirability-Based Multiobjective Optimization and Ranking

Following the strategy outlined previously, we began by seeking the best MLR models relating each property to the DRAGON molecular descriptors. Both properties, $-\log IC_{50}$ and $-\log CC_{50}$, were mapped as a linear function of respective subsets of ten and eight variables previously selected by GA. The resulting best-fit models are given below (equations 10 and 11) together with the respective statistical regression parameters.

$$\begin{aligned} -\log IC_{50} = & -37.474(\pm 5.535) \\ & -1.693(\pm 0.240)MAXDP \\ & -0.911(\pm 0.168)X1sol \\ & -18.385(\pm 4.756)SIC0 \\ & -3.748(\pm 0.999)GATS1p \\ & +2.809(\pm 0.384)ESpm15r \\ & +0.035(\pm 0.005)Eig1v \\ & -3.637(\pm 0.874)Ks \\ & +84.691(\pm 18.171)R8u+ \\ & +2.318(\pm 0.869)R8m \\ & -0.440(\pm 0.229)N-075 \end{aligned} \tag{10}$$

$N = 88$; $R^2_{FIT} = 0.72$; $s_{FIT} = 0.58$; $F = 20.20$;

$p < 0.01$; $\rho = 8.00$; $R^2_{LOOCV} = 0.66$; $s_{LOOCV} = 0.65$

$$
\begin{aligned}
-\log CC_{50} = & -2.336(\pm 0.175) \\
& -1.149(\pm 0.399)MATS3m \\
& -0.698(\pm 0.212).MATS5e \\
& +0.021(\pm 0.009)RDF070p \\
& +0.268(\pm 0.076)Mor18e \\
& -0.556(\pm 0.208)H7e \\
& +1.731(0.598).R8p \\
& -0.485(\pm 0.148)nROH \\
& +0.123(\pm 0.057)C-003
\end{aligned} \tag{11}
$$

$N = 88$; $R^2_{FIT} = 0.52$; $s_{FIT} = 0.27$; $F = 10.59$; $p < 0.01$; $\rho = 9.78$; $R^2_{LOOCV} = 0.38$; $s_{LOOCV} = 0.31$

Although the $-\log IC_{50}$ model exhibits a satisfactory goodness of fit in the initial training set of 88 NNRTIs, this is not the case for the $-\log CC_{50}$ model, even when its variables are significantly related with the property. This is usually due to the presence of outliers in the training set.

In order to detect those training compounds that influence model parameters to a marked extent, we plotted the leverage value for each compound versus their respective standardized residual value. This type of plots, if applied to test instead of training compounds, can also be used for checking the applicability domain (AD) of the model, a theoretical region in chemical space, defined by the model descriptors and modeled response.[29]

A prediction should be considered unreliable for compounds with a high leverage value (i.e., with $h > h^*$, being the critical value $h^* = 3p/N$). 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 remaining data, and hence, such predictions must be considered with great care.[29]

Here, it is very important to highlight that only predicted data for chemicals belonging to the chemical domain of the training set should be proposed for further screening of new HIV-1 RT inhibitors.

The applicability domain of the PMs determined by plotting the leverage values (h) versus standardized residuals (Std. Res.) of the 88 training compounds is shown in Figure 2. From this plot, the AD is established inside a squared area within ± 2 standard deviations and a leverage threshold h^* of 0.307 and 0.375 for the $-\log CC_{50}$ and $-\log IC_{50}$ models, respectively.

According to the analysis, five compounds exhibited an outlier behavior. Specifically, two outliers (compounds **9** and **19**) were found for the $-\log CC_{50}$ model, one outlier (compound **44)** for the $-\log IC_{50}$ model, and two common outliers (compounds **46** and **73**).

In order to keep a common training set for both models, the five outlier compounds were removed from the initial training set. The new models obtained (Equations 12 and 13) after refitting are shown below.

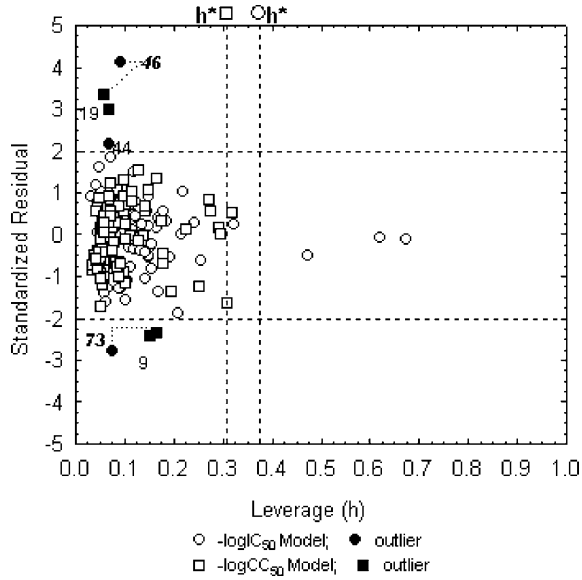


Figure 2. Applicability domain of the MLR PMs.

$$
\begin{aligned}
-\log IC_{50} = & -36.893(\pm 4.425) \\
& -1.650(\pm 0.192)MAXDP \\
& -0.904(\pm 0.138)X1sol \\
& -20.207(\pm 3.859)SIC0 \\
& -3.600(\pm 0.821)GATS1p \\
& +2.772(\pm 0.311)ESpm15r \\
& +0.035(\pm 0.004)Eig1v \\
& -3.110(\pm 0.696)Ks \\
& +78.791(\pm 14.967)R8u+ \\
& +2.832(\pm 0.692)R8m \\
& -0.416(\pm 0.182)N-075
\end{aligned}
$$

$N = 83$; $R^2_{FIT} = 0.82$; $s_{FIT} = 0.46$; $F = 32.12$; $p < 0.01$; $\rho = 7.55$; $R^2_{LOOCV} = 0.75$; $s_{LOOCV} = 0.53$; $R^2_{VAL} = 0.74$; $R^2_{TEST} = 0.72$ (12)

$$
\begin{aligned}
-\log CC_{50} = & -2.460(\pm 0.137) \\
& -1.804(\pm 0.325)MATS3m \\
& -0.669(\pm 0.163).MATS5e \\
& +0.032(\pm 0.007)RDF070p \\
& +0.215(\pm 0.061)Mor18e \\
& -0.743(\pm 0.162)H7e \\
& +1.435(0.470).R8p \\
& -0.495(\pm 0.114)nROH \\
& +0.154(\pm 0.047)C-003
\end{aligned}
$$

$$
N = 83;\ R^2_{\mathrm{FIT}} = 0.70;\ s_{\mathrm{FIT}} = 0.21;\ F = 22.04;\ \mathrm{p} < 0.01;\ \rho = 9.22;\ R^2_{\mathrm{LOOCV}} = 0.61;\ s_{\mathrm{LOOCV}} = 0.24;\ R^2_{VAL} = 0.57;\ R^2_{\mathrm{TEST}} = 0.50 \tag{13}
$$

As can be noticed, the goodness of fit of both models is significantly improved, especially taking into account that the values of R^2_{FIT} and R^2_{LOOCV} of the $-\log CC_{50}$ model are now 0.70 and 0.61, respectively.

As detailed previously, the evaluation of the predictive ability of the respective models was conducted by two independent sets of NNRTIs never used for training. The first one – validation set – comprises 21 NNRTIs randomly selected from an initial pool of 109 compounds. The second evaluation set – test set – corresponds to a subset of 18 compounds, within the AD of both models, taken from the set of 53 NNRTIs discarded due to reported inaccurate values for one or both properties.

Specifically, the values of R^2_{VAL} and R^2_{TEST} for the $-\log CC_{50}$ model were 0.57 and 0.50, respectively; whereas for the $-\log IC_{50}$ model they were 0.74 and 0.72, respectively. These values can be improved if, after checking the respective ADs, outlier compounds are removed. Actually, the predictive ability of the $-\log CC_{50}$ model in validation and test sets is higher if we do not consider outlier compounds ($R^2_{\mathrm{VAL}} = 0.65$, $R^2_{\mathrm{TEST}} = 0.81$). The predictive ability of the $-\log IC_{50}$ model in this case is also improved ($R^2_{\mathrm{TEST}} = 0.86$). The outliers can be identified by checking the ADs of both models for the validation and test set compounds detailed in Figures SI-1 and SI-2 of the supporting information.

Summarizing, the models are good both in terms of their statistical significance and predictive ability. No violations of the basic MLR assumptions were found that could compromise the reliability of the resulting predictions (see details in Table SI-6 of the supporting information).

The overall desirability function exhibits good statistical quality as indicated by the $R^2_{\mathrm{D,FIT}}$ (0.73). Moreover, a $R^2_{\mathrm{D,LOOCV}}$ value of 0.65 provides an adequate level of reliability regarding the method for predicting $D_{\mathrm{IC50\text{-}CC50}}$. This conclusion is reinforced by the high values of R^2_{D} obtained for validation and test set compounds ($R^2_{\mathrm{D,VAL}} = 0.86$, $R^2_{\mathrm{D,TEST}} = 0.69$). Table 3 contains the expected and predicted desirability values attributable to each response plus the individual and overall desirability values for the training, validation and test sets. The IC_{50} and CC_{50} values for the full data are also provided in this table.

At the same time, the performance evaluation of the overall desirability function in a classification task instead of regression can be informative too about its reliability for further tasks of MOOP and ranking. That is, to evaluate their performance in the identification of NNRTI candidates with favorable pharmaceutical profiles ($D_{IC50\text{-}CC50} \geq 0.5$). From Table 4 we can note that in all the subsets, the accuracy, sensitivity and specificity values are always higher than 80%. The excellent classification performance achieved by using an overall desirability function of predictions derived from our two MLR models supports the consistency of these models as evaluation functions of the MOOP process. This ensures the reliability of the optimal theoretical NNRTI candidate obtained, and consequently the quality of the subsequent ranking process using it as a template.

So, based on the satisfactory accuracy and predictive ability of the developed PMs we can proceed with an adequate level of confidence to the simultaneous optimization of the HIV-1 RT inhibitory activity and the toxicity over MT4 blood cells of the library of compounds. 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 5. In particular, the theoretical levels of the predictive variables required to reach a desirable ($D_{IC50\text{-}CC50} = 1.000$) NNRTI candidate with the best possible compromise between HIV-1 RT inhibitory efficacy ($IC_{50} = 0.001\ \mu\mathrm{M}$) and toxicity over MT4 blood cells ($CC_{50} = 563.638\ \mu\mathrm{M}$) are shown. As can be noticed, although we found the levels of the descriptors that simultaneously produce the most desirable combination of properties, none of these could be substantially improved. This is a logic result since the specific binding mechanism of this family of RT inhibitors "*a priori* promise" a favorable pharmaceutical profile (compromise between inhibitory efficacy and toxicity).[3–4] This is another reason why, in order to overcome the virus-NNRTIs resistance, is more feasible to look for new candidates with acceptable trade-offs between inhibitory efficacy and safety, rather than individually optimize one or another property.

The levels of the predictive variables required to reach a desirable NNRTI candidate are used as a pattern to rank the library used for training. The optimal set of weights w_i leading to the maximal concordance between descriptions (Δ_i) and solutions (D_i) of compounds used for training is shown in Table 6. The computed values of D_i, Δ_i and the normalized values of Δ_i ($^D\Delta_i$) for the library of compounds used for ranking are detailed in Table SI-7 of the supporting information material.

Table 3. Observed and predicted values of the optimized properties and their respective individual and overall desirability values for the training, validation and test set compounds used on the desirability-based MOOP process.

ID	IC_{50} (μM)	$-\log IC_{50}$	*Pred.* $-\log IC_{50}$	d_{IC50}	*Pred.* d_{IC50}	CC_{50} (μM)	$-\log CC_{50}$	*Pred.* $-\log CC_{50}$	d_{CC50}	*Pred.* d_{CC50}	$D_{IC50\text{-}CC50}$	Pred. $D_{IC50\text{-}CC50}$
1	5.06	−0.704	−0.176	0.270	0.405	405	−2.607	−2.208	0.938	0.723	0.503	0.541
2	5.370	−0.730	−0.608	0.264	0.295	529	−2.723	−2.839	1.000	1.000	0.514	0.543
3	0.02	1.699	1.431	0.884	0.815	60	−1.778	−1.985	0.491	0.603	0.659	0.701
4	0.186	0.730	0.476	0.636	0.571	30.33	−1.482	−1.507	0.332	0.346	0.460	0.444
5	0.107	0.971	0.942	0.698	0.691	39.07	−1.592	−1.693	0.391	0.446	0.522	0.555
6	2.7	−0.431	0.721	0.340	0.634	29.46	−1.469	−1.508	0.325	0.346	0.333	0.468
7	0.093	1.032	0.547	0.713	0.590	196.9	−2.294	−2.053	0.769	0.640	0.741	0.614
8	0.062	1.208	1.030	0.758	0.713	237	−2.375	−2.210	0.812	0.724	0.785	0.718
9	0.26	0.585	–	–	–	250	−2.398	–	–	–	–	–
10	0.09	1.046	0.805	0.717	0.655	42.23	−1.626	−1.824	0.409	0.516	0.542	0.582
11	0.156	0.807	0.974	0.656	0.699	74.4	−1.872	−2.074	0.542	0.651	0.596	0.674
12	1.065	−0.027	0.386	0.443	0.549	41.54	−1.618	−1.937	0.406	0.577	0.424	0.563
13	0.284	0.547	0.966	0.590	0.697	193.9	−2.288	−2.118	0.766	0.674	0.672	0.685
14	0.009	2.032	1.686	0.969	0.880	143.2	−2.156	−2.150	0.695	0.692	0.820	0.780
15	0.009	2.027	2.292	0.967	1.035	133	−2.124	−2.317	0.677	0.781	0.810	0.899
16	0.114	0.943	0.450	0.691	0.565	187.93	−2.274	−2.269	0.758	0.755	0.724	0.653
17	0.018	1.735	1.566	0.893	0.850	92.83	−1.968	−2.074	0.593	0.651	0.728	0.744
18	0.012	1.928	1.056	0.942	0.719	170.6	−2.232	−2.281	0.736	0.762	0.832	0.740
19	0.028	1.553	–	–	–	32.15	−1.507	–	–	–	–	–
20	0.081	1.093	1.084	0.729	0.727	243	−2.386	−2.269	0.818	0.756	0.772	0.741
21	0.06	1.222	1.937	0.762	0.944	256.32	−2.409	−2.281	0.831	0.762	0.796	0.848
22	58	−1.763	−1.039	0.000	0.185	111	−2.045	−1.955	0.635	0.587	0.000	0.329
23	4.5	−0.653	−0.314	0.283	0.370	40.4	−1.606	−1.610	0.399	0.401	0.336	0.385
24	4.2	−0.623	−0.862	0.291	0.230	91	−1.959	−1.722	0.589	0.461	0.414	0.326
25	4	−0.602	−1.001	0.296	0.195	43	−1.633	−1.613	0.414	0.403	0.350	0.280
26	4.3	−0.633	−0.663	0.288	0.281	102	−2.009	−1.835	0.615	0.522	0.421	0.383
27	7.7	−0.886	−0.734	0.224	0.263	66.6	−1.823	−1.817	0.516	0.512	0.340	0.367
28	10.3	−1.013	−1.065	0.192	0.178	133	−2.124	−1.928	0.677	0.572	0.360	0.319
29	11.6	−1.064	−0.975	0.178	0.201	43	−1.633	−1.820	0.414	0.514	0.272	0.322
30	8.8	−0.944	−1.083	0.209	0.174	43	−1.633	−1.618	0.414	0.406	0.294	0.265
31	8.4	−0.924	−0.739	0.214	0.262	82	−1.914	−1.982	0.564	0.601	0.348	0.396
32	8.6	−0.934	−1.225	0.212	0.137	125.2	−2.098	−2.024	0.663	0.624	0.375	0.293
33	1.4	−0.146	−0.211	0.413	0.396	122.4	−2.088	−2.200	0.658	0.718	0.521	0.533
34	6	−0.778	−0.761	0.251	0.256	63	−1.799	−1.906	0.503	0.560	0.356	0.379
35	1.2	−0.079	0.553	0.430	0.591	53	−1.724	−2.013	0.463	0.618	0.446	0.604
36	0.38	0.420	0.933	0.557	0.688	100	−2.000	−1.824	0.611	0.516	0.583	0.596
37	0.007	2.155	1.455	1.000	0.821	41	−1.613	−1.721	0.403	0.461	0.634	0.615
38	0.01	2.000	2.257	0.960	1.026	18	−1.255	−1.415	0.210	0.296	0.449	0.551
39	0.008	2.097	1.972	0.985	0.953	18	−1.255	−1.547	0.210	0.367	0.455	0.591
40	0.01	2.000	1.937	0.960	0.944	168	−2.225	−2.032	0.732	0.628	0.839	0.770
41	0.66	0.180	0.242	0.496	0.512	53.19	−1.726	−1.870	0.463	0.541	0.479	0.526
42	0.8	0.097	0.151	0.475	0.489	49.39	−1.694	−1.530	0.446	0.358	0.460	0.418
43	2.7	−0.431	0.347	0.340	0.539	46.09	−1.664	−1.683	0.430	0.440	0.382	0.487
44	0.09	1.046	–	–	–	163.86	−2.214	–	–	–	–	–
45	4.71	−0.673	−0.245	0.278	0.387	148.32	−2.171	−1.974	0.703	0.597	0.442	0.481
46	0.002	2.699	–	–	–	10.81	−1.034	–	–	–	–	–
47	6.34	−0.802	−0.409	0.245	0.346	58.95	−1.770	−1.641	0.487	0.418	0.346	0.380
48	7.28	−0.862	−0.890	0.230	0.223	53.81	−1.731	−1.882	0.466	0.547	0.327	0.349
49	1.17	−0.068	−0.075	0.433	0.431	122.35	−2.088	−1.794	0.658	0.500	0.534	0.464
50	0.18	0.745	−0.357	0.640	0.359	51.88	−1.715	−1.528	0.458	0.357	0.541	0.358
51	4.85	−0.686	−0.184	0.275	0.403	47.99	−1.681	−1.620	0.439	0.406	0.348	0.405
52	1.25	−0.097	−0.450	0.425	0.335	24.52	−1.390	−1.242	0.282	0.203	0.347	0.261
53	0.64	0.194	−0.148	0.500	0.412	58.02	−1.764	−2.080	0.484	0.654	0.492	0.519
54	1.33	−0.124	−0.201	0.418	0.399	34.19	−1.534	−1.590	0.360	0.391	0.388	0.395
55	2.22	−0.346	−0.687	0.362	0.275	23.46	−1.370	−1.537	0.272	0.362	0.314	0.315
56	1.34	−0.127	0.039	0.418	0.460	7.32	−0.865	−1.269	0.000	0.218	0.000	0.316
57	3.53	−0.548	−0.691	0.310	0.274	26.79	−1.428	−1.293	0.303	0.230	0.307	0.251
58	37.58	−1.575	−1.796	0.048	0.000	241.61	−2.383	−2.522	0.817	0.892	0.198	0.000
59	2.05	−0.312	−0.285	0.370	0.377	263.19	−2.420	−2.323	0.837	0.784	0.557	0.544
60	27.84	−1.445	−1.285	0.081	0.122	196.54	−2.293	−2.111	0.769	0.671	0.250	0.286

Table 3. (Continued)

ID	IC_{50} (μM)	$-\log IC_{50}$	*Pred.*$-\log IC_{50}$	d_{IC50}	*Pred.*d_{IC50}	CC_{50} (μM)	$-\log CC_{50}$	*Pred.*$-\log CC_{50}$	d_{CC50}	*Pred.*d_{CC50}	$D_{IC50\text{-}CC50}$	Pred.$D_{IC50\text{-}CC50}$
61	6.79	−0.832	−1.029	0.238	0.187	178.92	−2.253	−2.119	0.747	0.675	0.421	0.356
62	0.46	0.337	0.167	0.536	0.493	39.6	−1.598	−1.642	0.394	0.418	0.460	0.454
63	0.5	0.301	−0.216	0.527	0.395	38.78	−1.589	−1.662	0.390	0.429	0.453	0.412
64	0.21	0.678	0.255	0.623	0.515	28.73	−1.458	−1.750	0.319	0.476	0.446	0.495
65	0.057	1.244	1.052	0.768	0.719	33.2	−1.521	−1.655	0.353	0.425	0.521	0.553
66	0.65	0.187	−0.085	0.498	0.428	35.58	−1.551	−1.831	0.369	0.520	0.429	0.472
67	0.38	0.420	0.670	0.557	0.621	32.75	−1.515	−1.332	0.350	0.252	0.442	0.395
68	0.063	1.201	0.359	0.756	0.542	35.3	−1.548	−1.632	0.368	0.413	0.527	0.473
69	1.86	−0.270	0.099	0.381	0.475	37.11	−1.569	−1.627	0.379	0.410	0.380	0.441
70	0.83	0.081	0.024	0.471	0.456	32.57	−1.513	−1.713	0.349	0.457	0.405	0.456
71	2.96	−0.471	0.020	0.330	0.455	32.57	−1.513	−1.359	0.349	0.266	0.339	0.348
72	0.092	1.036	0.535	0.714	0.586	30.96	−1.491	−1.652	0.337	0.424	0.491	0.499
73	12.53	−1.098	–	–	–	171.92	−2.235	–	–	–	–	–
74	0.11	0.959	0.657	0.695	0.618	32.82	−1.516	−1.696	0.351	0.447	0.493	0.526
75	0.067	1.174	0.666	0.750	0.620	30.23	−1.480	−1.610	0.331	0.401	0.498	0.499
76	2.77	−0.442	0.394	0.337	0.551	31.4	−1.497	−1.069	0.340	0.110	0.339	0.246
77	0.23	0.638	−0.049	0.613	0.438	77.07	−1.887	−1.934	0.550	0.576	0.581	0.502
78	22.02	−1.343	−1.246	0.107	0.132	210.09	−2.322	−1.844	0.784	0.527	0.290	0.264
79	3.37	−0.528	−0.812	0.315	0.243	42.19	−1.625	−1.758	0.409	0.481	0.359	0.342
80	0.65	0.187	−0.193	0.498	0.401	154.5	−2.189	−2.011	0.712	0.617	0.596	0.497
81	2.65	−0.423	−0.304	0.342	0.373	36.3	−1.560	−1.759	0.374	0.481	0.358	0.423
82	0.83	0.081	0.082	0.471	0.471	187.34	−2.273	−1.953	0.757	0.585	0.597	0.525
83	0.47	0.328	−0.001	0.534	0.450	50.13	−1.700	−1.880	0.449	0.546	0.490	0.496
84	0.065	1.187	1.063	0.753	0.721	34.19	−1.534	−1.918	0.360	0.567	0.521	0.639
85	25.99	−1.415	−0.846	0.089	0.234	180.48	−2.256	−2.005	0.749	0.613	0.258	0.379
86	7.06	−0.849	−0.158	0.233	0.410	195.23	−2.291	−2.022	0.767	0.623	0.423	0.505
87	8.78	−0.943	0.013	0.209	0.453	155.47	−2.192	−1.773	0.714	0.489	0.387	0.471
88	22.66	−1.355	−1.153	0.104	0.156	192.9	−2.285	−2.248	0.764	0.744	0.282	0.341
					Validation Set ($R^2_{D,VAL(IC50\text{-}CC50)}$ = 0.856)							
89	0.733	0.135	0.607	0.484	0.605	207.7	−2.317	−2.071	0.782	0.649	0.615	0.627
90	1.63	−0.212	0.658	0.396	0.618	181.6	−2.259	−2.024	0.750	0.624	0.545	0.621
91	0.256	0.592	0.526	0.601	0.584	134.44	−2.129	−2.079	0.680	0.653	0.639	0.618
92	0.0195	1.710	1.414	0.886	0.811	244	−2.387	−2.230	0.819	0.735	0.852	0.772
93	6	−0.778	−0.024	0.251	0.444	103	−2.013	−1.940	0.618	0.579	0.394	0.507
94	8	−0.903	−0.859	0.220	0.231	122	−2.086	−1.955	0.657	0.586	0.380	0.368
95	7.6	−0.881	−0.402	0.225	0.348	80	−1.903	−1.701	0.559	0.450	0.355	0.396
96	4	−0.602	−0.335	0.296	0.365	70.7	−1.849	−1.887	0.530	0.550	0.396	0.448
97	1.3	−0.114	−0.683	0.421	0.276	51.5	−1.712	−1.827	0.456	0.518	0.438	0.378
98	16.94	−1.229	−1.056	0.136	0.180	210.37	−2.323	−1.963	0.785	0.591	0.327	0.327
99	4.69	−0.671	−0.935	0.279	0.211	60.37	−1.781	−1.963	0.493	0.591	0.371	0.353
100	10.87	−1.036	−0.899	0.186	0.221	24.69	−1.393	−1.727	0.284	0.464	0.230	0.320
101	0.18	0.745	0.446	0.640	0.564	52.5	−1.720	−1.755	0.460	0.479	0.543	0.520
102	0.15	0.824	0.652	0.660	0.617	38.4	−1.584	−1.797	0.387	0.502	0.506	0.556
103	0.48	0.319	0.245	0.531	0.513	34.28	−1.535	−1.694	0.361	0.446	0.438	0.478
104	0.51	0.292	0.172	0.525	0.494	38.4	−1.584	−1.550	0.387	0.369	0.451	0.427
105	0.25	0.602	0.420	0.604	0.557	29.6	−1.471	−1.673	0.326	0.435	0.444	0.492
106	7.37	−0.867	−0.027	0.229	0.443	36.12	−1.558	−1.120	0.373	0.137	0.292	0.247
107	0.21	0.678	0.535	0.623	0.586	31.47	−1.498	−1.652	0.341	0.424	0.461	0.499
108	4.69	−0.671	−0.933	0.279	0.212	59.58	−1.775	−1.873	0.490	0.542	0.370	0.339
109	0.099	1.004	0.649	0.706	0.616	43.89	−1.642	−1.729	0.418	0.465	0.544	0.535
					Test Set ($R^2_{D,TEST(IC50\text{-}CC50)}$ = 0.694)							
110	3.65	−0.562	0.102	0.307	0.476	> 37.88	−1.578	−2.031	0.384	0.627	0.343	0.547
111	0.693	0.159	0.718	0.491	0.633	≥ 209	−2.320	−1.819	0.783	0.513	0.620	0.570
112	0.203	0.693	0.921	0.627	0.685	> 263.2	−2.420	−2.139	0.837	0.685	0.724	0.685
113	0.187	0.728	0.382	0.636	0.547	> 290	−2.462	−2.242	0.860	0.741	0.739	0.637
114	> 38.3	−1.583	−1.066	0.046	0.178	38.3	−1.583	−2.080	0.387	0.654	0.133	0.341
115	> 44	−1.643	−1.960	0.031	0.000	44	−1.643	−1.649	0.419	0.422	0.113	0.000
116	> 22.80	−1.358	−0.111	0.103	0.422	22.8	−1.358	−1.535	0.265	0.361	0.166	0.390
117	0.24	0.620	0.039	0.608	0.460	> 344.35	−2.537	−1.893	0.900	0.553	0.740	0.504
118	0.58	0.237	−0.310	0.510	0.371	> 346.26	−2.539	−1.670	0.901	0.433	0.678	0.401
119	≥ 35.73	−1.553	−1.376	0.054	0.099	116.66	−2.067	−1.755	0.647	0.479	0.186	0.218

Table 3. (Continued)

ID	IC_{50} (μM)	$-\log IC_{50}$	*Pred.*$-\log IC_{50}$	d_{IC50}	*Pred.*d_{IC50}	CC_{50} (μM)	$-\log CC_{50}$	*Pred.*$-\log CC_{50}$	d_{CC50}	*Pred.*d_{CC50}	$D_{IC50\text{-}CC50}$	Pred.$D_{IC50\text{-}CC50}$
120	>2.09	−0.320	−0.506	0.368	0.321	2.09	−0.320	−1.049	0.000	0.099	0.000	0.178
121	>23.53	−1.372	−0.989	0.100	0.198	23.53	−1.372	−1.281	0.273	0.224	0.165	0.211
122	>21.03	−1.323	−1.291	0.112	0.121	21.03	−1.323	−1.651	0.247	0.423	0.167	0.226
123	>304.41	−2.483	−1.310	0.000	0.116	>304.41	−2.483	−2.143	0.871	0.688	0.000	0.282
124	>246.65	−2.392	−2.248	0.000	0.000	246.65	−2.392	−2.393	0.822	0.822	0.000	0.000
125	>123.56	−2.092	−2.337	0.000	0.000	123.56	−2.092	−2.158	0.660	0.696	0.000	0.000
126	>312.50	−2.495	−1.512	0.000	0.064	>312.50	−2.495	−1.872	0.877	0.542	0.000	0.187
127	≥60.75	−1.784	−1.924	0.000	0.000	157.43	−2.197	−2.211	0.717	0.725	0.000	0.000

Table 4. Classification performance of the overall desirability function on training, validation and test sets.

TRINING SET								EVALUATION SETS							
Fit					***LOOCV***			***Validation***					***Test***		
Classification Matrix								*Classification Matrix*							
	Obs.				Obs.				Obs.				Obs.		
		+	−			+	−			+	−			+	−
Pred.	+	**25**	8	Pred.	+	**25**	11	Pred.	+	**7**	1	Pred.	+	**4**	1
	−	4	**46**		−	4	**43**		−	0	**13**		−	1	**12**
%				**Statistic**	%			%				**Statistic**	%		
85.54				Accuracy	81.93			95.24				Accuracy	88.89		
82.21				Sensitivity	82.21			100.00				Sensitivity	80.00		
85.19				Specificity	79.63			92.86				Specificity	92.31		
14.81				FP Rate	20.37			7.14				FP Rate	7.69		
17.79				FN Rate	17.79			0.00				FN Rate	20.00		

Table 5. Results of the desirability-based MOOP process.

Predictors Optimum Level						
MAXDP = 4.013	*Ks* = 0.800			*RDF070p* = 15.444		
X1sol = 16.882	*R8u+* = 0.038			*Mor18e* = −2.794		
SIC0 = 0.368	*R8m* = 1.022			*H8e* = 0.107		
GATS1p = 1.180	*N-075* = 3.000			*R8p* = 0.118		
ESpm15r = 22.385	*MATS3m* = 0.035			*nROH* = 0.000		
Eig1v = 260.585	*MATS5e* = 0.321			*C-003* = 0.000		
Pharmaceutical Profile	HIV-1 RT Inhibition Profile			MT4 Blood Cells Toxicity Profile		
$D_{IC50\text{-}CC50}$	$-\log IC_{50}$	IC_{50}	d_{IC50}	$-\log CC_{50}$	CC_{50}	d_{CC50}
1.000	2.860	0.001 μM	1.000	−2.751	563.638′ μM	1.000

Table 6. Optimal set of weights for ranking.

Variable	*MAXDP*	*X1sol*	*SIC0*	*GATS1p*	*ESpm15r*	*Eig1v*	*Ks*	*R8u+*	*R8m*
w_i	0.1135	−0.5515	4.9162	0.1876	1.1991	0.0197	−2.8436	56.1139	−0.0093
Variable	*N−075*	*MATS3m*	*MATS5e*	*RDF070p*	*Mor18e*	*H8e*	*R8p*	*nROH*	*C−003*
w_i	0.6871	3.6896	−1.8542	0.0362	0.2277	−0.8664	−3.1562	−0.8067	−0.0728

Based on Δ_i, it is possible to arrive at a ranking of the training set of NNRTIs with a corrected ranking quality index (Ψ^*) of 0.357 representing a percentage of ranking quality ($R_{\%}$) of 64.34. On the other hand, better ranking quality indices were obtained for the validation ($R_{\%}$ = 70.91) and test sets ($R_{\%}$ = 72.84). In addition, if the full set of 122 NNRTI compounds is considered (including all subset of compounds together), we obtain a percentage of ranking quality over 60% ($R_{\%}$ = 62.14). These ranks, compared with their respective perfect ranking, are shown in Figure 3.

Remarkably, the ranking attained ($R_{\%}$) in all subsets are similar to the predictability values exhibited on the PMs (R^2) as well as on the MOOP process (R^2_D). Specifically, in the training set, for the $-\log IC_{50}$ model, a $R_{\%}$ = 64 is supported

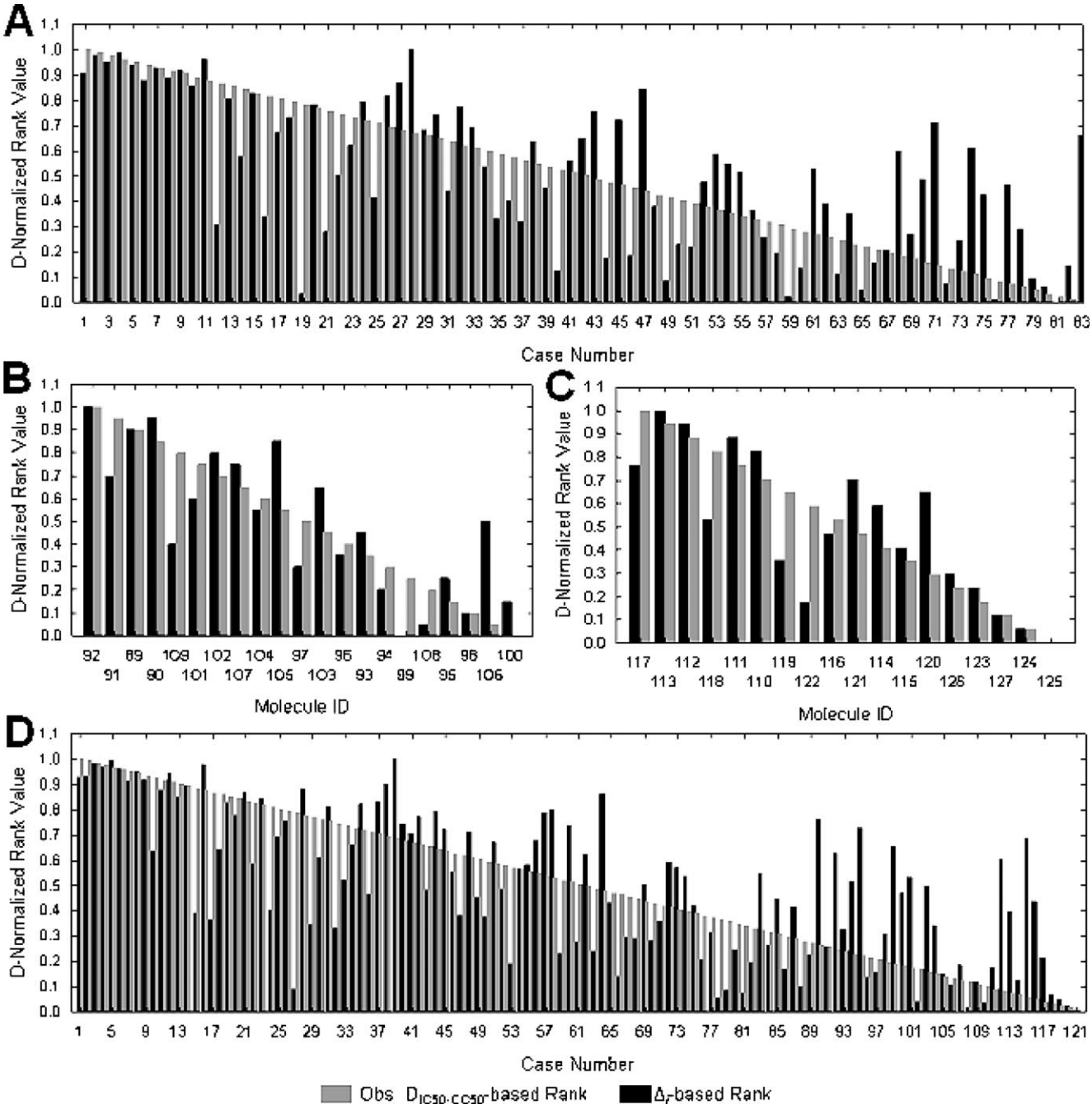


Figure 3. Δ_r-based ranking attained for the (A) training, (B) validation, (C) test, and (D) full set of NNRTIs compounds.

by a R^2_{LOOCV} value of 0.75 and, for the $-\log CC_{50}$ model, a R^2_{LOOCV} value of 0.61 by a $R^2_{D,LOOCV}$ value of 0.65. The same occurs with the validation and test sets (validation set: $R_{\%}=73$, $R^2_{-\log IC50}=0.74$, $R^2_{-\log CC50}=0.57$, $R^2_D=0.86$; test set: $R_{\%}=71$, $R^2_{-\log IC50}=0.72$, $R^2_{-\log CC50}=0.50$, $R^2_D=0.69$). This fact indicates that the quality of both process (desirability-based MOOP and ranking) are dependent on the quality of the initial set of PMs suggesting that the ranking algorithm reflects the quality of the PMs and the MOOP process on which it is based.

However, 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 ones for synthesis and biological assessment.

With this regard, we decided to test the ability of this multiobjective VS strategy to prioritize NNRTI candidates with favorable pharmaceutical profiles ($D_{IC50\text{-}CC50} \geq 0.5$) disperse in a data set of NNRTI decoys. NNRTI decoys are

physically similar but chemically distinct from NNRTIs, so that they are unlikely to be binders of the HIV RT. Specifically, we used as positive cases the 12 HIV RT known ligands with favorable pharmaceutical profiles included on the validation and test sets, and 36 decoys (negative cases) for each known ligand (432 decoys) were randomly selected from the database of HIV RT decoys included on the directory of useful decoys (DUD).[30]

We only considered those decoys included on the AD of our prediction models at a ratio of 36 decoys per ligand, as recommended by Huang et al.[30] The final set of 444 compounds is ranked according to their structural similarity (Δ_i) with the previously determined optimal candidate, and the enrichment ability of this strategy is finally tested according to the enrichment metrics previously detailed and now depicted in Table 7.

The respective values of *AUAC* and *ROC Metric* obtained suggest that the method is able to rank a NNRTI candidate with a favorable pharmaceutical profile earlier than a NNRTI decoy with a probability around 0.8. At the same time, $TP/FP_{ROC\text{-}OP}$ informs that, to obtain the best performance is necessary to filter 23.2% of the library, in turn leading to find 83.3% of the TP cases at a cost of only 21.5% of FP cases, which represents a $EF_{MAX}=3.592$. Furthermore, all the positive cases can be found at the first 32% of the library. On the other hand, a third of the compounds retrieved, after filtering the top 10% of the library, were NNRTI candidates with a favorable pharmaceutical profile ($Ya_{10\%}=0.33$), which represents an $EF_{10\%}=3.364$, being 10.09 the maximum possible value of *EF* for this data fraction.

The respective ROC, accumulation, and enrichment curves can be checked in the Figure SI-3 of the supporting information. The ranked list of 12 NNRTIs with favorable pharmaceutical profile and 432 NNRTI decoys based on Δ_i can be consulted in Table SI-8 of the supporting information material.

So, considering the previous results, one may well expect that larger (real or virtual) libraries of molecules (always inside the applicability domain of the PMs), like combinatorial libraries, could be correctly ranked; prioritizing in this way those candidates (top ranked) with more favorable compromise between inhibitory efficacy and safety.

Table 7. Enrichment metrics for Δ_i-based ranking of the data set collected form DUD.

Enrichment Metrics	MOOP Rank
ROC Curve Information	
ROC Metric	0.798
$TP/FP_{ROC\text{-}OP}$	0.833/0.215
Accumulation Curve Information	
AUAC	0.828
$\chi_{100\%}$	0.320
$Ya_{10\%}$	0.333
Enrichment Curve Information	
$EF_{10\%}$	3.364
EF_{Max}	3.592

3.2 Multiobjective versus Sequential and Parallel NNRTIs Virtual Screening

Filtering the most promising candidates having the best compromise between inhibitory efficacy and toxicity confers to the process of discovery and development of new NNRTI drugs an elevated degree of rationality which is difficult to reach via traditional QSARs which optimize sequentially each property. The sequential optimization of the properties comprising the final pharmaceutical profile of a drug candidate implies to overlook at some stage properties equally decisive to reach a successful drug or, at least, to find only by chance a candidate with acceptable profiles of all properties simultaneously. That is, a potent candidate once identified via QSAR has a high probability of being discarded later as a drug because of an unacceptable toxicological profile with the useless expenses of time and resources in synthesis and pharmacological assays.[31]

Equally difficult is the choice of using a panel of models (i.e., a parallel screening based on QSAR models to respectively map the inhibitory efficacy and toxicity) since it is not very probable to find a candidate with all the properties simultaneously optimized and if this happens the results are more by chance than fruit of a rational drug development strategy.

For instance, the suitability of a multiobjective VS approach can be checked if we compare the enrichment achieved in the screening of NNRTI candidates with a favorable pharmaceutical profile from the full set of 122 NNRTI compounds, just considering the inhibitory efficacy profile (the predicted values of $-\log IC_{50}$: $Pred.-\log IC_{50}$) in opposition to use the pharmaceutical profile information deduced from Δ_i.

In general, the overall enrichment performance of the Δ_i-based rank is comparable (just slightly superior) to the $Pred.-\log IC_{50}$-based rank. Inspecting the respective ROC and accumulation curves depicted in Figure 4, we can note that for both cases the probability to rank a positive case earlier than a negative case is always around 0.7 (see *ROC Metric* and *AUAC* values in Table 8). In addition, to retrieve 100% of the positive cases through the Δ_i-based rank it is necessary to screen almost 87% of the library contrasted with only a 54% via the *Pred-based* rank (see the $\chi_{100\%}$values). According to this information, there is no reason to privilege one or another ranking criterion, and consequently, neither a reason to substitute the current approach (i.e., prioritization of drug candidates based on their pharmacological efficacy). However, analyzing the enrichment achieved by applying each ranking criterion at specific fractions, instead of using metrics based on the whole data set, the previous conclusion is not supported.

Actually, the enrichment achieved by the Δ_i-based rank in the initial fraction (up to the top 10% of the dataset) is superior to the obtained through the $Pred.-logIC_{50}$-based rank (see the respective values of $Ya_{5\%}$, $Ya_{10\%}$, $EF_{5\%}$, and $EF_{10\%}$ in Table 8). In a lesser degree, the same behavior is

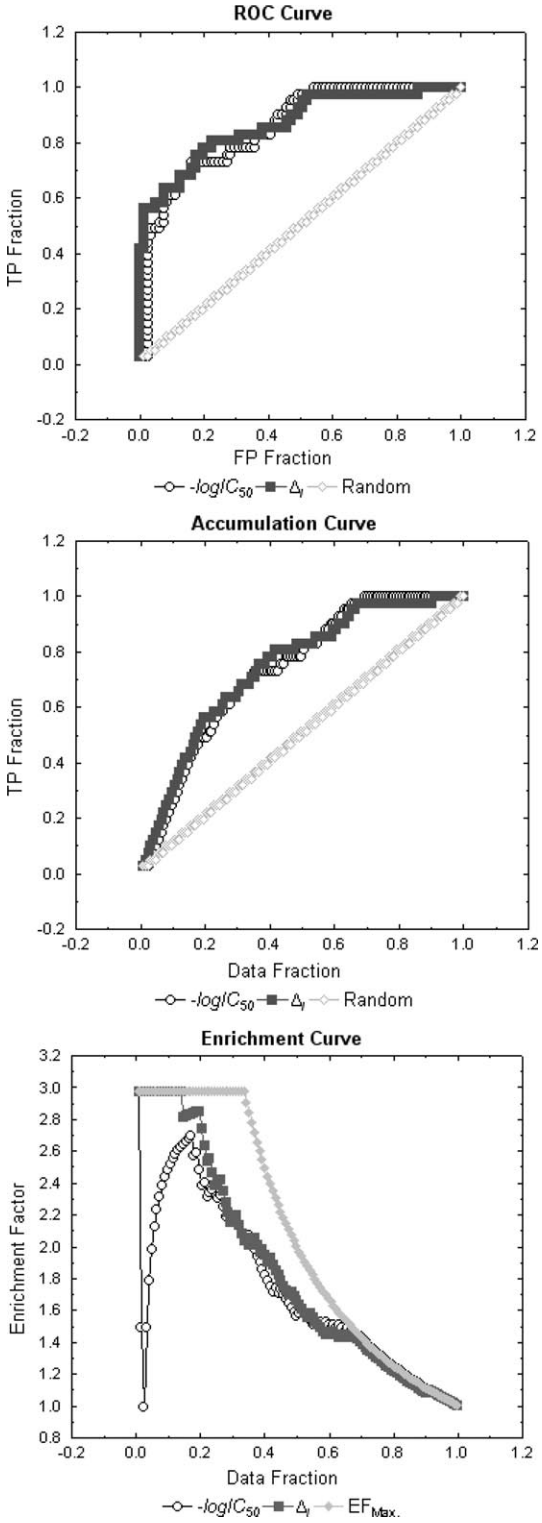


Figure 4. ROC, accumulation, and enrichment curves for the $-\log IC_{50}$- and Δ_i-based ranks of the full set of 122 NNRTI compounds.

observed in later fractions of the data set (after the top 20%) as indicated by the respective values of $Ya_{20\%}$, $Ya_{50\%}$, $EF_{20\%}$, and $EF_{50\%}$. Another metric supporting the use of Δ_i over the $Pred.-logIC_{50}$ values as ranking criterion in a VS effort is $TP/FP_{ROC\text{-}OP}$. The operating point for both ranks is found after screening approximately the same fraction of the dataset (top 17% and 19% for the $Pred.-logIC_{50}$- and Δ_i-based ranks, respectively). Nevertheless, the *TP/FP* ratio achieved by the Δ_i-based rank is significantly better (Δ_i rank: $TP/FP_{ROC\text{-}OP}=0.56/0.01$; $Pred.-logIC_{50}$ rank: $TP/FP_{ROC\text{-}OP}=0.46/0.02$).

The comparison of the enrichment curve of each approach with the ideal enrichment curve for the present data set allows confirming the previous statement. Note in Figure 4 that the enrichment curve obtained for the Δ_i-based rank resembles the ideal curve better than the $Pred.-\log IC_{50}$-based rank, especially on initial and final fractions.

Anyhow, VS endeavors also consider safety criteria in subsequent steps. So, if the screening is conducted in a sequential manner, starting with the selection of candidates fulfilling a previously established threshold for the inhibitory efficacy ($Pred.-logIC_{50}\geq 0.196$; $Pred.IC_{50}\leq 0.64$ μM; $Pred.d_{IC50}\geq 0.5$) and further eliminating those candidates with an unfavorable safety profile ($Pred.-logCC_{50}\leq -1.794$; $Pred.CC_{50}\geq 62.23$ μM; $Pred.d_{CC50}\geq 0.5$), the area of selected candidates is reduced. As a consequence, 41% of the candidates (17 out of 41) with favorable pharmaceutical profiles ($D_{IC50\text{-}CC50}\geq 0.5$) are mistakenly discarded (see Figure 5A). However, by considering the compromise between inhibitory efficacy and safety of the candidates through a multiobjective virtual screening ($Pred.D_{IC50\text{-}CC50}\geq 0.5$) is possible to retrieve up to 88% of the candidates with acceptable pharmaceutical profiles included on the library (see Figure 5B).

This reveals the importance of considering multiple properties simultaneously since the sequential application of property filters could have led to the elimination of the candidate, despite it having a good balance between most of the properties.[32] The importance of achieving a balance across a range of criteria is also recognized by other groups.[33]

However, that can be settle on in a more detailed way by simulating a VS attempt over the same data set through three different VS approaches, and conducting a retrospective analysis of the performance of each approach by comparing the respective degree of enrichment achieved at the top 10% of the data set. As referred to above, the multiobjective VS approach proposed in this work is compared with two of the approaches – QSAR-based sequential and parallel VS – currently employed on drug discovery.

The sequential selection guides retrieving 75% of the pharmaceutically acceptable compounds included on the top 10% fraction of the data set, which represents an $EF_{10\%}=2.232$. Similar but inferior results were achieved through a parallel screening ($Ya_{10\%}=0.6$; $EF_{10\%}=1.785$). These results although very good are outperformed when the selection of compounds was made based on a multiobjective criterion (the structural similarity to an optimal can-

Table 8. Enrichment metrics for predicted inhibitory (*Pred.−logIC$_{50}$*) and Δ_i-based ranking of the full set of 122 NNRTI compounds.

Pred.−logIC$_{50}$ Rank	Enrichment Metrics	Δ_i Rank
ROC Curve Information		
0.654	ROC Metric	0.668
0.46/0.02	TP/FP$_{ROC\text{-}OP}$	0.56/0.01
Accumulation Curve Information		
0.730	AUAC	0.740
0.543	$\chi_{100\%}$	0.864
0.667	$Ya_{5\%}$	1.000
0.833	$Ya_{10\%}$	1.000
0.833	$Ya_{20\%}$	0.958
0.780	$Ya_{50\%}$	0.829
Enrichment Curve Information		
1.984	$EF_{5\%}$	2.976
2.480	$EF_{10\%}$	2.976
2.480	$EF_{20\%}$	2.852
1.561	$EF_{50\%}$	1.659
2.692	EF_{Max}	2.976

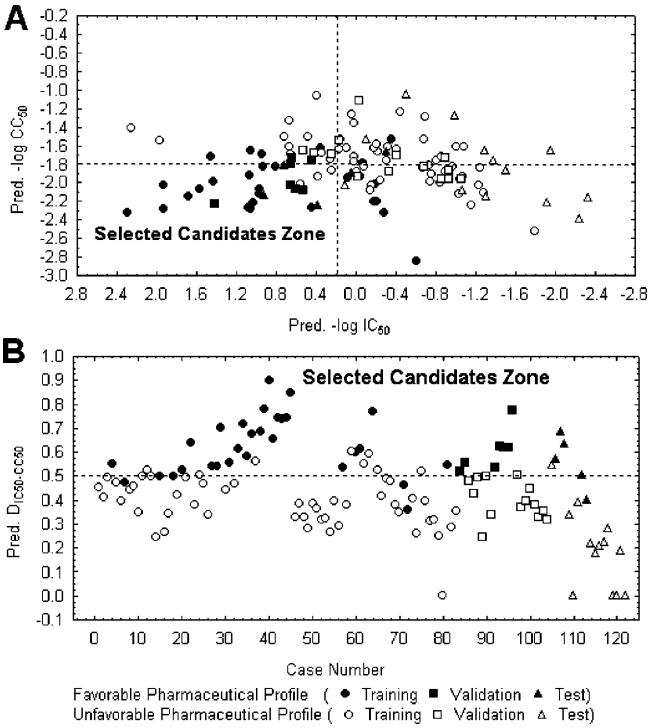


Figure 5. Graphical representation of the results for (A) a sequential screening [based on the inhibitory efficacy (*Pred.−logIC$_{50}$*) and safety (*Pred.−logCC$_{50}$*) profiles], and (B) a multiobjective screening [based on the pharmaceutical profile (*Pred.D$_{IC50\text{-}CC50}$*)], of the full set of 122 NNRTI compounds.

didate, Δ_i). In the latter case, it was possible to retrieve 100% included on the same fraction of the data, reaching the maximum possible *EF* value for this fraction ($EF_{10\%}$= 2.976). More significant is the fact that compounds, initially selected, were rejected by the sequential or the parallel VS approach, even when they actually exhibited a pharmaceutically acceptable profile (false negative compounds, *FN*). Specifically, one out of twelve, and three out of twenty compounds were mistakenly discarded through the sequential and the parallel approach, respectively. All these results are detailed in Tables 9–11.

4 Conclusions

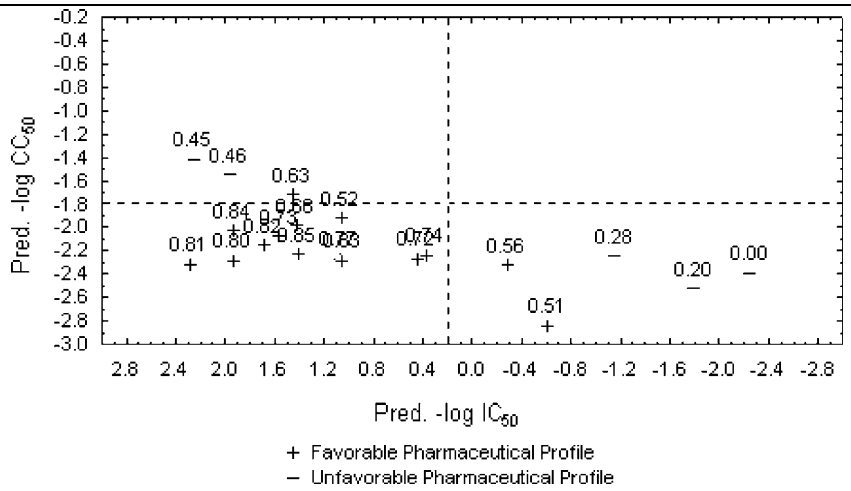
The results obtained in this work allow highlighting the benefits of exploiting a combined strategy of desirability-based multiobjective optimization and ranking as valuable tools in drug discovery and development process. The data herein obtained allow to determine the theoretical levels of a set of molecular descriptors leading to a pharmaceutically desirable HIV-1 NNRTI candidate and use it as a pattern to rank libraries of new compounds according to the degree of structural similarity. The developed MOOP strategy can be efficiently employed as a VS tool for the identification and prioritization of new NNRTI hits with acceptable trade-offs of the inhibitory efficacy towards the HIV-1 RT and the toxic effects towards MT4 blood cells. The comparative study between the sequential, parallel and multiobjective VS approaches of the selected library of compounds revealed that the multiobjective approach can be superior to the other approaches. Moreover, it can rule out the exclusion of pharmaceutically acceptable candidates.

The data obtained so far provide evidences that support the beneficial application of the multiobjective VS strategy in the identification of NNRTIs hits with appropriate trade-offs between potency and safety. The adjustment of the multiple criteria in hit-to-lead identification and lead optimization is considered to increase the likelihood of the candidate to evolve into a successful antiretroviral drug.

Table 9. Ordered list of NNRTI candidates, obtained through parallel filtering (according to the predicted values of $-\log IC_{50}$ and $-\log CC_{50}$) of the top 10% of the full set of 122 NNRTIs.[a]

Parallel virtual screening

(HIV−1 RT Inhibitory Efficacy Profile (IC_{50}) and MT4 Blood Cells Safety Profile (CC_{50}) ≥ 50%)



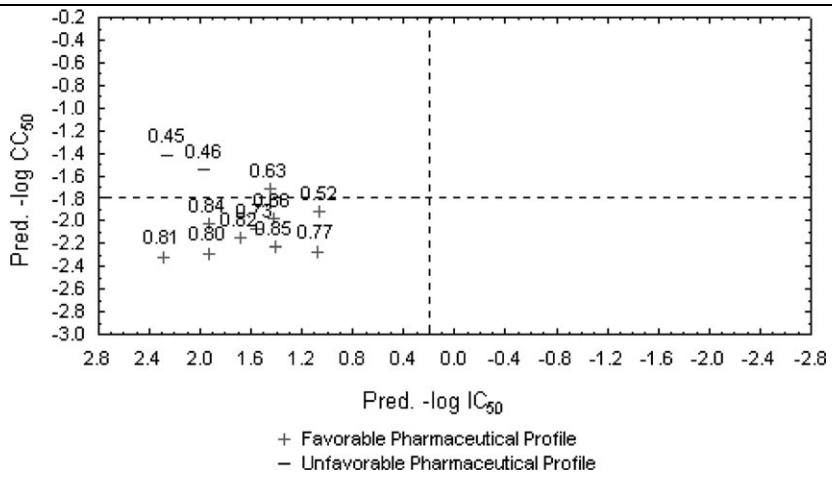
ID	NNRTI Analogue	*Pred.* $-logIC_{50}$	*Pred.* d_{IC50}	*Pred.* IC_{50} Class	*Pred.* $-logCC_{50}$	*Pred.* d_{CC50}	*Pred.* CC_{50} Class	$D_{IC50\text{-}CC50}$	Control Class ($D_{IC50\text{-}CC50}$)
15	DATA	2.292	1.035	+	−2.317	0.781	+	0.810	+
38	ATC	2.257	1.026	+	−1.415	0.296	−	0.449	−
39	ATC	1.972	0.953	+	−1.547	0.367	−	0.455	−
21	DATA	1.937	0.944	+	−2.281	0.762	+	0.796	+
40	ATC	1.937	0.944	+	−2.032	0.628	+	0.839	+
14	DATA	1.686	0.880	+	−2.150	0.692	+	0.820	+
17	DATA	1.566	0.850	+	−2.074	0.651	+	0.728	+
37	ATC	1.455	0.821	+	−1.721	0.461	−	0.634	+
3	ATC	1.431	0.815	+	−1.985	0.603	+	0.659	+
92	DATA	1.414	0.811	+	−2.230	0.735	+	0.852	+
20	DATA	1.084	0.727	+	−2.269	0.756	+	0.772	+
84	HEPT	1.063	0.721	+	−1.918	0.567	+	0.521	+
2	HEPT	−0.608	0.295	−	−2.839	1.000	+	0.514	+
58	DABO	−1.796	0.000	−	−2.522	0.892	+	0.198	−
124	DABO	−2.248	0.000	−	−2.393	0.822	+	0.000	−
59	DABO	−0.285	0.377	−	−2.323	0.784	+	0.557	+
18	DATA	1.056	0.719	+	−2.281	0.762	+	0.832	+
16	DATA	0.450	0.565	+	−2.269	0.755	+	0.724	+
88	HEPT	−1.153	0.156	−	−2.248	0.744	+	0.282	−
113	DATA	0.382	0.547	+	−2.242	0.741	+	0.739	+
	Enrichment Metrics			$Ya_{10\%}=0.600$			$EF_{10\%}=1.785$		

[a] The corresponding overall desirability ($D_{IC50\text{-}CC50}$) values are placed over each compound represented in the graph.

Table 10. Ordered list of NNRTI candidates, obtained through sequential filtering (according to the predicted values of $-\log IC_{50}$ and $-\log CC_{50}$) of the top 10% of the full set of 122 NNRTIs. The corresponding overall desirability ($D_{IC50\text{-}CC50}$) values are placed over each compound represented in the graph.

Sequential virtual screening

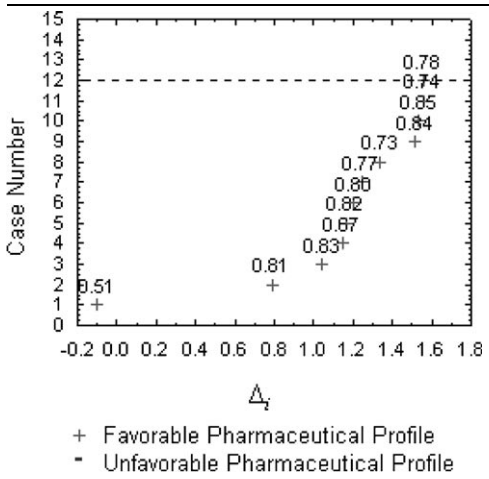
(HIV−1 RT Inhibitory Efficacy Profile (IC_{50}) and MT4 Blood Cells Safety Profile (CC_{50}) ≥ 50%)



ID	NNRTI Analogue	*Pred.* $-logIC_{50}$	*Pred.* d_{IC50}	*Pred.* IC_{50} Class	*Pred.* $-logCC_{50}$	*Pred.* d_{CC50}	*Pred.* CC_{50} Class	$D_{IC50\text{-}CC50}$	Control class ($D_{IC50\text{-}CC50}$)
15	DATA	2.292	1.035	+	−2.317	0.781	+	0.810	+
38	ATC	2.257	1.026	+	−1.415	0.296	−	0.449	−
39	ATC	1.972	0.953	+	−1.547	0.367	−	0.455	−
21	DATA	1.937	0.944	+	−2.281	0.762	+	0.796	+
40	ATC	1.937	0.944	+	−2.032	0.628	+	0.839	+
14	DATA	1.686	0.880	+	−2.150	0.692	+	0.820	+
17	DATA	1.566	0.850	+	−2.074	0.651	+	0.728	+
37	ATC	1.455	0.821	+	−1.721	0.461	−	0.634	+
3	ATC	1.431	0.815	+	−1.985	0.603	+	0.659	+
92	DATA	1.414	0.811	+	−2.230	0.735	+	0.852	+
20	DATA	1.084	0.727	+	−2.269	0.756	+	0.772	+
84	HEPT	1.063	0.721	+	−1.9+8	0.567	+	0.521	+
	Enrichment Metrics			$Ya_{10\%}=0.750$			$EF_{10\%}=2.232$		

Table 11. Ordered list of NNRTI candidates, obtained through multiobjective filtering (according to Δ_i) of the top 10% of the full set of 122 NNRTIs. The corresponding overall desirability ($D_{IC50\text{-}CC50}$) values are placed over each compound represented in the graph.

MULTIOBJECTIVE VIRTUAL SCREENING (*Pharmaceutical Profile ($D_{IC50\text{-}CC50}$) ≥ 50%)*



ID	NNRTI Analogue	Δ_i	*Pred.* D_{IC50}	$D_{IC50\text{-}CC50}$	Control Class ($D_{IC50\text{-}CC50}$)
2	HEPT	0.097	0.543	0.514	+
15	DATA	0.790	0.899	0.810	+
18	DATA	1.045	0.740	0.832	+
13	DATA	1.148	0.685	0.672	+
14	DATA	1.163	0.780	0.820	+
21	DATA	1.202	0.848	0.796	+
20	DATA	1.246	0.741	0.772	+
17	DATA	1.336	0.744	0.728	+
40	ATC	1.517	0.770	0.839	+
92	DATA	1.535	0.772	0.852	+
7	DATA	1.555	0.614	0.741	+
8	DATA	1.557	0.718	0,785	+
Entrichment Metrics		$Ya_{10\%}=1.000$		$EF_{10\%}=2.976$	

Acknowledgements

The authors acknowledge the Portuguese *Fundação para a Ciência e a Tecnologia* (FCT) for financial support (SFRH/BD/30698/2006 and SFRH/BSAB/930/2009).

References

[1] H. Jonckheere, J. Anne, E. De Clercq, *Med. Res. Rev.* **2000**, *20*(2), 129–154.
[2] R. W. Buckheit, *Expert Opin. Investig. Drugs* **2001**, *10*(8), 1423–1442.
[3] a) G. Hajos, S. Riedi, J. Molnar, D. Szabo, *Drugs Future* **2000**, *25*, 47–62; b) J. Ren, J. Milton, K. L. Weaver, S. A. Short, D. I. Stuart, D. K. Stammers, *Structure* **2000**, *8*(10), 1089–1094.
[4] E. de Clercq, *Med. Res. Rev.* **1996**, *16*, 125–157.
[5] E. De Clercq, *Int. J. Antimicrob. Agents* **2009**, *33*(4), 307–320.
[6] a) E. De Clercq, *Biochem. Pharmacol.* **1994**, *47*(2), 155–169; b) R. G. Nanni, J. Ding, A. Jacobo-Molina, S. H. Hughes, E. Arnold, *Perspect. Drug Discov. Design* **1993**, *1*(1), 129–150; c) C. Tantillo, J. Ding, A. Jacobo-Molina, R. G. Nanni, P. L. Boyer, S. H. Hughes, R. Pauwels, K. Andries, P. A. Janssen, E. Arnold, *J. Mol. Biol.* **1994**, *243*(3), 369–387.
[7] a) C. J. Manly, S. Louise-May, J. D. Hammer, *Drug Discov. Today* **2001**, *6*(21), 1101–1110; b) K. H. Bleicher, H. J. Bohm, K. Muller, A. I. Alanine, *Nat. Rev. Drug Discov.* **2003**, *2*(5), 369–378.
[8] a) O. Nicolotti, I. Giangreco, T. F. Miscioscia, A. Carotti, *J. Chem. Inf. Model.* **2009**, *49*(10), 2290–2302; b) M. Cruz-Monteagudo, F. Borges, M. N. Cordeiro, *J. Comput. Chem.* **2008**, *29*(14), 2445–2459; c) O. Nicolotti, V. J. Gillet, P. J. Fleming, D. V. Green, *J. Med. Chem.* **2002**, *45*(23), 5069–5080; d) M. Cruz-Monteagudo, F. Borges, M. N. D. S. Cordeiro, J. L. Cagide Fajin, C. Morell, R. Molina Ruiz, Y. Cañizares-Carmenate, E. Rosa Dominguez, *J. Comb. Chem.* **2008**, *10*(6), 897–913; e) C. A. Nicolaou, J. Apostolakis, C. S. Pattichis, *J. Chem. Inf. Model.* **2009**; f) F. Yamashita, H. Hara, T. Ito, M. Hashida, *J. Chem. Inf. Model.* **2008**, 48(2), 364–369; g) A. Machado, E. Tejera, M. Cruz-Monteagudo, I. Rebelo, *Eur. J. Med. Chem.* **2009**, *44*(12), 5045–5054.
[9] H. J. Federsel, *Drug Discov. Today* **2006**, *11*(21–22), 966–974.
[10] a) J. Xu, A. Hagler, *Molecules* **2002**, *7*, 566–700; b) D. Butina, M. D. Segall, K. Frankcombe, *Drug Discov. Today* **2002**, *7*(11 Suppl), S83–S88.
[11] J. A. DiMasi, R. W. Hansen, H. G. Grabowski, *J. Health Econ.* **2003**, *22*(2), 151–185.
[12] a) A. Ranise, A. Spallarossa, S. Schenone, O. Bruno, F. Bondavalli, L. Vargiu, T. Marceddu, M. Mura, P. La Colla, A. Pani, *J. Med. Chem.* **2003**, *46*(5), 768–781; b) G. F. Sun, X. X. Chen, F. E. Chen, Y. P. Wang, E. De Clercq, J. Balzarini, C. Pannecouque, *Chem. Pharm. Bull. (Tokyo)* **2005**, *53*(8), 886–892; c) L. Ji, F. E. Chen, E. De Clercq, J. Balzarini, C. Pannecouque, *J. Med. Chem.* **2007**, *50*(8), 1778–1786; d) Y. Z. Xiong, F. E. Chen, J. Balzarini, E. De Clercq, C. Pannecouque, *Eur. J. Med. Chem.* **2008**, *43*(6), 1230–1236.
[13] a) J. F. Truchon, C. I. Bayly, *J. Chem. Inf. Model.* **2007**, *47*(2), 488–508; b) J. Kirchmair, P. Markt, S. Distinto, G. Wolber, T. Langer, *J. Comput. Aided Mol. Des.* **2008**, *22*(3–4), 213–228; c) U. Fechner, G. Schneider, *Chembiochem* **2004**, *5*(4), 538–540.
[14] R. Pauwels, J. Balzarini, M. Baba, R. Snoeck, D. Schols, P. Herdewijn, J. Desmyter, E. De Clercq, *J. Virol. Methods* **1988**, *20*, 309–321.
[15] G. Derringer, R. Suich, *J. Quality Technol.* **1980**, *12*(4), 214–219.
[16] a) J. A. Nelder, R. Mead, *Computer J.* **1965**, *7*, 308–313; b) R. Fletcher, C. M. Reeves, *Computer Journal* **1964**, *7*, 149–154; c) R. Hooke, T. A. Jeeves, *J. Assoc. Comp. Machin.* **1961**, *8*, 212–229.
[17] Statsoft_Inc., *STATISTICA*, **2001**.
[18] a) C. De Boor, *A Practical Guide to Splines*, Springer-Verlag, New York, **1978**; b) C. F. Gerald, P. O. Wheatley, *Applied Numerical Analysis*, Addison Wesley, Reading, MA, **1989**.
[19] a) T. F. Coleman, Y. Li, *SIAM J. Optim.* **1996**, *6*, 418–445; b) T. F. Coleman, Y. Li, *Math. Program.* **1994**, *67*(2), 189–224.
[20] *MATLAB*, The MathWorks, Inc., **2006**.
[21] I. H. Witten, E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*, Morgan Kaufmann, San Francisco, **2005**.
[22] a) U. Burkert, N. L. Allinger, *Molecular Mechanics*, ACS, Washington, DC, USA, **1982**; b) T. Clark, *Computational Chemistry*, Wiley, New York, USA, **1985**.
[23] J. Frank, *MOPAC*, Seiler Research Laboratory, US Air Force Academy, Colorado Springs, CO, **1993**.
[24] R. Todeschini, V. Consonni, M. Pavan, *DRAGON Software*, Milano Chemometrics, Milano, **2002**.
[25] a) R. Leardi, R. Boggia, M. Terrile, *J. Chemom.* **1992**, *6*, 267–281; b) A. Yasri, D. Hartsough, *J. Chem. Inf. Comput. Sci.* **2001**, *41*(5), 1218–1227; c) Y. Zhu, Y. Yu, X. Chen, *Zhonghua Yu Fang Yi Xue Za Zhi* **1999**, *33*(1), 21–25; d) K. Hasegawa, T. Kimura, K. Funatsu, *J. Chem. Inf. Comput. Sci.* **1999**, *39*(1), 112–120.
[26] a) D. Barbosa de Oliveira, A. C. Gaudio, BuildQSAR, Physics Department-CCE, University of Espírito Santo, Vitória ES, Brasil, **2000**; b) D. Barbosa de Oliveira, A. C. Gaudio, *Quant. Struct.-Act. Relat.* **2000**, *19*, 599–601.
[27] a) G. Cruciani, M. Baroni, S. Clementi, G. Costantino, D. Riganelli, B. Skagerberg, *J. Chemom.* **1992**, *6*(6), 335–346; b) H. Van Waterbeemd, *Chemometric Methods in Molecular Design*, Vol. 2 (Eds: R. Mannhold, P. Krogsgaard-Larsen, H. Timmerman), Wiley-VCH, New York, **1995**.
[28] B. Jancic-Stojanovic, A. Malenovic, D. Ivanovic, T. Rakic, M. Medenica, *J. Chromatogr. A* **2009**, *1216*(8), 1263–1269.
[29] a) A. C. Atkinson, *Plots, Transformations, and Regression*, Clarendon Press, Oxford, UK, **1985**; b) L. Eriksson, J. Jaworska, A. P. Worth, M. T. Cronin, R. M. McDowell, P. Gramatica, *Environ. Health Perspect.* **2003**, *111*(10), 1361–1375.
[30] N. Huang, B. K. Shoichet, J. J. Irwin, *J. Med. Chem.* **2006**, *49*(23), 6789–6801.
[31] J. Drews, *Drug Discov. Today* **1998**, *3*, 491–494.
[32] V. J. Gillet, *Curr. Opin. Chem. Biol.* **2008**, *12*(3), 372–378.
[33] M. D. Segall, A. P. Beresford, J. M. Gola, D. Hawksley, M. H. Tarbit, *Expert Opin. Drug Metab. Toxicol.* **2006**, *2*(2), 325–337.

Received: October 12, 2009
Accepted: February 19, 2010
Published online: April 14, 2010

SUPPORTING INFORMATION

PRIORITIZING HITS WITH APPROPRIATE TRADE-OFFS BETWEEN HIV-1 REVERSE TRANSCRIPTASE INHIBITORY EFFICACY AND MT4 BLOOD CELLS TOXICITY THROUGH DESIRABILITY-BASED MULTI-OBJECTIVE OPTIMIZATION AND RANKING

Maykel Cruz-Monteagudo, Hai PhamThe, M. Natalia D. S. Cordeiro, Fernanda Borges

CONTENTS

Table SI-1. Identification (ID), names and references relative to the data collected.

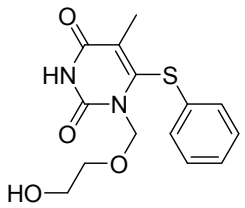
ID	Compound Name	Ref.[a]
1	1-[(2-Hydroxyethoxy)methyl]-6-(phenylthio) thymine (**HEPT**)	A, B
2	2,3-dideoxyinosine (**DDI**)	B,C
3	Trovirdine	D
4	4-[4-Methylamino-6-(1-naphthoxy)-1,3,5-triazine-2-yl]aminobenzonitrile (**8a**)	C
5	4-[4-Amino-6-(1-naphthoxy)-1,3,5-triazine-2-yl] aminobenzonitrile (**8b**)	C
6	4-[4-n-Propylamino-6-(1-naphthoxy)-1,3,5-triazine-2-yl]aminobenzonitrile (**8c**)	C
7	4-[4-Methylamino-6-(4-chloro-1-naphthoxy)-1,3,5-triazine-2-yl] amino benzonitrile (**8d**)	C
8	4-[4-Amino-6-(4-chloro-1-naphthoxy)-1,3,5-triazine-2-yl]aminobenzonitrile (**8e**)	C
9	4-[4-Methylamino-6-(2-naphthoxy)-1,3,5-triazine-2-yl]aminobenzonitrile (**9a**)	C
10	4-[4-Amino-6-(2-naphthoxy)-1,3,5-triazine-2-yl]aminobenzonitrile (**9b**)	C
11	4-[4-Ethylamino-6-(1,6-dibromo-2-naphthoxy)-1,3,5-triazine-2-yl] amino benzonitrile (**9c**)	C
12	4-[4-i-Propylamino-6-(6-bromo-2-naphthoxy)-1,3,5-triazine-2-yl] amino benzonitrile (**9f**)	C
13	4-[4-Methylamino-6-(6-bromo-2-naphthoxy)-1,3,5-triazine-2-yl] amino benzonitrile (**9g**)	C
14	4-[4-Methylamino-6-(1-bromo-2-naphthoxy)-1,3,5-triazine-2-yl] amino benzonitrile (**9h**)	C
15	4-[4-Amino-6-(1-bromo-2-naphthoxy)-1,3,5-triazine-2-yl] amino benzonitrile (**9i**)	C
16	4-[4-Ethylamino-6-(1-chloro-2-naphthoxy)-1,3,5-triazine-2-yl] amino benzonitrile (**9k**)	C
17	4-[4-Methylamino-6-(1,6-dibromo-2-naphthoxy)-1,3,5-triazine-2-yl] amino benzonitrile (**9m**)	C
18	4-[4-Methylamino-6-(1-chloro-2- naphthoxy)-1,3,5-triazine-2-yl] amino benzonitrile (**9o**)	C
19	4-[4-Amino-6-(1-chloro-2-naphthoxy)-1,3,5-triazine-2-yl]aminobenzonitrile (**9p**)	C
20	4-[4-Amino-6-(6-bromo-2-naphthoxy)-1,3,5-triazine-2-yl]aminobenzonitrile (**9q**)	C
21	4-[4-Azido-6-(1-chloro-2-naphthoxy)-1,3,5-triazine-2-yl]aminobenzonitrile (**9r**)	C
22	O-(Benzyl) 2-furoyl (phenyl) thiocarbamate (**13q**)	D
23	O-(2-Phenethyl) (E)-cinnamoyl (phenyl) thiocarbamate (**15b)**	D
24	O-(2-Phenethyl) benzoyl (phenyl) thiocarbamate (**15c**)	D
25	O-(2-Phenethyl) 4-chlorobenzoyl (phenyl)thiocarbamate (**15g**)	D
26	O-(2-Phenethyl) 2-furoyl (phenyl) thiocarbamate (**15q**)	D
27	O-(2-Phenoxyethyl) (E)-Cinnamoyl (phenyl)thiocarbamate (**17b**)	D
28	O-(2-Phenoxyethyl) 4-chlorobenzoyl (phenyl)thiocarbamate (**17g**)	D
29	O-(2-Phenoxyethyl) 2,4-dichlorobenzoyl(phenyl)thiocarbamate (**17k**)	D
30	O-(2-Phenoxyethyl) 3,5-dichlorobenzoyl (phenyl)thiocarbamate (**17m**)	D
31	O-(2-Phenoxyethyl) 2-furoyl (phenyl) thiocarbamate (**17q**)	D
32	O-(2-Phenoxyethyl) phenyl(thien-2-yl carbonyl)thiocarbamate (**17r**)	D
33	(±) O-(1-Methyl-2-phenoxyethyl) phenoxyacetyl(phenyl)thiocarbamate (**19a**)	D
34	(±) O-(1-Methyl-2-phenoxyethyl) 4-nitrophenyl(thien-2-yl carbonyl) thiocarbamate (**22r**)	D
35	O-[2-(1,3-Dioxo-1,3-dihydro-2H-isoindol-2-yl)ethyl]3-bromophenyl (thien-2-ylcarbonyl)thiocarbamate (**35r**)	D
36	O-[2-(1,3-Dioxo-1,3-dihydro-2H-isoindol-2-yl)ethyl] 3-nitrophenyl(thien-2-ylcarbonyl)thiocarbamate (**36r**)	D
37	O-[2-(1,3-Dioxo-1,3-dihydro-2H-isoindol-2-yl)ethyl]4-chlorophenyl(2-furoyl)thiocarbamate (**41q**)	D
38	O-[2-(1,3-Dioxo-1,3-dihydro-2H-isoindol-2-yl)ethyl]4-iodophenyl(thien-2-ylcarbonyl)thiocarbamate (**43r**)	D
39	O-[2-(1,3-Dioxo-1,3-dihydro-2H-isoindol-2-yl)ethyl]2-furoyl(4-nitrophenyl)thiocarbamate (**45q**)	D
40	O-[2-(1,3-Dioxo-1,3-dihydro-2H-isoindol-2-yl)ethyl] 4-nitrophenyl(thien-2-ylcarbonyl)thiocarbamate (**45r**)	D
41	6-[α-Cyano-(1-naphthylmethyl)]-3,4-dihydro-2-isopropylthiopyrimidin-4(3H)-one (**3a**)	A
42	6-[α-Cyano-(1-naphthylmethyl)]-2-cyclopentylthio-3,4-dihydro-5-methyl pyrimidin-4(3H)-one (**3c**)	A
43	6-[α-Cyano-(1-naphthylmethyl)]-3,4-dihydro-5-methyl-2-propynyl thiopyrimidin-4(3H)-one (**3d**)	A
44	2-Benzoylmethylthio-6-[α-cyano-(1-naphthylmethyl)]-3,4-dihydro-5-methyl pyrimidin-4(3H)-one (**3e**)	A
45	2-(4-Chlorobenzoylmethylthio)-6-[α-cyano-(1-naphthylmethyl)]-3,4-dihydro-5-methylpyrimidin-4(3H)-one (**3f**)	A
46	6-(α-Cyanobenzyl)-3,4-dihydro-2-isopropylthio-5-methylpyrimidin-4(3H)-one (**3g**)	A
47	6-(α-Cyanobenzyl)-2-cyclopentylthio-3,4-dihydro-5-methylpyrimidin-4(3H)-one (**3h**)	A
48	2-Benzoylmethylthio-6-(α-cyanobenzyl)-3,4-dihydro-5-methylpyrimidin-4(3H)-one (**3i**)	A
49	6-[α-Cyano-(1-naphthylmethyl)]-3,4-dihydro-5-ethyl-2-ethylthiopyrimidin-4(3H)-one (**3j**)	A
50	6-[α-Cyano-(1-naphthylmethyl)]-2-cyclopentylthio-3,4-dihydro-5-ethyl pyrimidin-4(3H)-one (**3l**)	A
51	6-[α-Cyano-(1-naphthylmethyl)]-3,4-dihydro-5-ethyl-2-propynylthiopyrimidin-4(3H)-one (**3n**)	A
52	6-[α-Cyano-(1-naphthylmethyl)]-3,4-dihydro-5-ethyl-2-(4-nitrobenzylthio) pyrimidin-4(3H)-one (**3o**)	A
53	6-(α-Cyanobenzyl)-3,4-dihydro-5-ethyl-2-isopropylthiopyrimidin-4(3H)-one (**3q**)	A
54	6-(α-Cyanobenzyl)-3,4-dihydro-5-ethyl-2-(4-nitrobenzylthio)-pyrimidin-4(3H)-one (**3r**)	A
55	6-(α-Cyano-2,6-dichlorobenzyl)-3,4-dihydro-5-ethyl-2-(4-nitrobenzylthio) pyrimidin-4(3H)-one (**3s**)	A
56	6-[α-Cyano-(1-naphthylmethyl)]-2-cyclopentylthio-3,4-dihydro-5-isopropyl pyrimidin-4(3H)-one (**3t**)	A
57	6-(α-Cyanobenzyl)-3,4-dihydro-5-isopropyl-2-(4-methoxybenzylthio) pyrimidin-4(3H)-one (**3v**)	A
58	3,4-Dihydro-2-ethylthio-6-(1-naphthoyl) pyrimidin-4(3H)-one (**4a**)	A
59	6-Benzoyl-3,4-dihydro-2-isopropylthio-5-methylpyrimidin-4(3H)-one (**4e**)	A
60	3,4-Dihydro-5-ethyl-2-ethylthio-6-(1-naphthoyl)pyrimidin-4(3H)-one (**4g**)	A
61	3,4-Dihydro-5-ethyl-2-isopropylthio-6-(1-naphthoyl)pyrimidin-4(3H)-one (**4h**)	A
62	1-Ethoxymethyl-5-methyl-6-(1-naphthylthio)uracil (**7a**)	B
63	1-Methoxymethyl-5-ethyl-6-(1-naphthylthio)uracil (**7c**)	B
64	1-[3-Methyl(benzyloxy)methyl]-5-ethyl-6-(1-naphthylthio)uracil (**7f**)	B
65	1-Ethoxymethyl-5-isopropyl-6-(1-naphthylthio)uracil (**7h**)	B
66	1-[(2-Methoxyethyloxy)methyl]-5-isopropyl-6-(2-naphthylthio)uracil (**7i**)	B
67	1-[(Cyclopropylmethoxy)methyl]-5-isopropyl-6-(1-naphthylthio)uracil (**7j**)	B
68	1-[(Benzyloxy)methyl]-5-isopropyl-6-(1-naphthylthio)uracil (**7k**)	B
69	1-Ethoxymethyl-5-propyl-6-(1-naphthylthio)uracil (**7m**)	B
70	1-[(Benzyloxy)methyl]-5-propyl-6-(1-naphthylthio)uracil (**7n**)	B

Table SI-1. (*Continued…*)

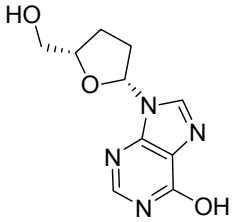
ID	Compound Name	Ref.[a]
71	1-[(Benzyloxy)methyl]-5-isobutyl-6-(1-naphthylthio)uracil (**7p**)	B
72	1-[(4-Fluorobenzyloxy)methyl]-5-ethyl-6-(1-naphthylthio)uracil (**7r**)	B
73	1-[(Cyclohexylmethoxy)methyl]-5-isopropyl-6-(1-naphthylthio)uracil (**7s**)	B
74	1-[(3-Fluorobenzyloxy)methyl]-5-isopropyl-6-(1-naphthylthio)uracil (**7t**)	B
75	1-[(4-Fluorobenzyloxy)methyl]-5-isopropyl-6-(1-naphthylthio)uracil (**7u**)	B
76	1-[(2-Phenylethoxy)methyl]-5-isopropyl-6-(1-naphthylthio)uracil (**7v**)	B
77	1-(Hydroxyethoxymethyl)-6-(α-naphthalenethio)-5-isopropyluracil (**7w**)	B
78	1-Methoxymethyl-5-ethyl-6-(2-naphthylthio)uracil (**8a**)	B
79	1-Ethoxymethyl-5-ethyl-6-(2-naphthylthio)uracil (**8b**)	B
80	1-[(Benzyloxy)methyl]-5-ethyl-6-(2-naphthylthio)uracil (**8c**)	B
81	1-Ethoxymethyl-5-isopropyl-6-(2-naphthylthio)uracil (**8e**)	B
82	1-[(Benzyloxy)methyl]-5-isopropyl-6-(2-naphthylthio)uracil (**8f**)	B
83	1-Methoxymethyl-5-ethyl-6-(1-nitro-2-naphthylthio)uracil (**11a**)	B
84	1-[(Benzyloxy)methyl]-5-ethyl-6-(1-nitro-2-naphthylthio)uracil (**11c**)	B
85	1-Methoxymethyl-5-ethyl-6-(1-amino-2-naphthylthio)uracil (**12a**)	B
86	1-Ethoxymethyl-5-ethyl-6-(1-amino-2-naphthylthio)uracil (**12b**)	B
87	1-[(Benzyloxy)methyl]-5-ethyl-6-(1-acetamino-2-naphthylthio)uracil (**13**)	B
88	1-Methoxymethyl-5-ethyl-6-(1-chloro-2-naphthylthio)uracil (**15**)	B
89	4-[4-Ethylamino-6-(4-chloro-1-naphthoxy)-1,3,5-triazine-2-yl] amino benzonitrile (**8f**)	C
90	4-[4-n-Propylamino-6-(4-chloro-1-naphthoxy)-1,3,5-triazine-2-yl] amino benzonitrile (**8g**)	C
91	4-[4-n-Propylamino-6-(1,6-dibromo-2-naphthoxy)-1,3,5-triazine-2-yl] amino benzonitrile (**9d**)	C
92	4-[4-Amino-6-(1,6-dibromo-2-naphthoxy)-1,3,5-triazine-2-yl] amino benzonitrile (**9n**)	C
93	O-(2-Phenoxyethyl) 2-phenoxyacetyl (phenyl)thiocarbamate (**17a**)	D
94	O-(2-Phenoxyethyl) benzoyl (phenyl) thiocarbamate (**17c**)	D
95	O-(2-Phenoxyethyl) 4-chloro-3-nitrobenzoyl(phenyl)thiocarbamate (**17n**)	D
96	O-(2-Phenoxyethyl) benzoyl(4-fluorophenyl)thiocarbamate (**18c**)	D
97	(±) O-(1-Methyl-2-phenoxyethyl) 2-furoyl(phenyl)thiocarbamate (**19q**)	D
98	2-Allylthio-3,4-dihydro-5-ethyl-6-(1-naphthoyl)pyrimidin-4(3H)-one (**4i**)	A
99	6-Benzoyl-2-cyclopentylthio-3,4-dihydro-5-ethylpyrimidin-4(3H)-one (**4l**)	A
100	2-Cyclopentylthio-3,4-dihydro-5-isopropyl-6-(1-naphthoyl)pyrimidin-4(3H)-one (**4m**)	A
101	1-[(Benzyloxy)methyl]-5-methyl-6-(1-naphthylthio)uracil (**7b**)	B
102	1-Ethoxymethyl-5-ethyl-6-(1-naphthylthio)uracil (**7d**)	B
103	1-[(Benzyloxy)methyl]-5-ethyl-6-(1-naphthylthio)uracil (**7e**)	B
104	1-Methoxymethyl-5-isopropyl-6-(1-naphthylthio)uracil (**7g**)	B
105	1-[(3-Methyl-phenylmethyloxy) methyl]-5-isopropyl-6-(1-naphthyl-thio)uracil (**7l**)	B
106	1-Ethoxymethyl-5-isobutyl-6-(1-naphthylthio)uracil (**7o**)	B
107	1-[(3-Fluorobenzyloxy)methyl]-5-ethyl-6-(1-naphthylthio)uracil (**7q**)	B
108	1-Methoxymethyl-5-isopropyl-6-(2-naphthylthio)uracil (**8d**)	B
109	1-Ethoxymethyl-5-ethyl-6-(1-nitro-2-naphthylthio)uracil (**11b**)	B
110	1-[(Benzyloxy)methyl]-5-ethyl-6-(1-amino-2-naphthylthio)uracil **(12c)**	B
111	4-[4-i-Propylamino-6-(1,6-dibromo-2-naphthoxy)-1,3,5-triazine-2-yl]aminobenzonitrile **(9e)**	C
112	4-[4-n-Propylamino-6-(1-bromo-2-naphthoxy)-1,3,5-triazine-2-yl]aminobenzonitrile **(9j)**	C
113	4-[4-n-Propylamino-6-(1-chloro-2-naphthoxy)-1,3,5-triazine-2-yl]aminobenzonitrile **(9l)**	C
114	O-(2-Furylmethyl) benzoyl(phenyl)thiocarbamate **(14c)**	D
115	O-(2-Phenylsulfanylethyl) benzoyl(phenyl)thiocarbamate **(23c)**	D
116	6-[α-Cyano-(1-naphthylmethyl)]-2-cyclopentylthio-3,4-dihydropyrimidin-4(3H)-one **(3b)**	A
117	6-[α-Cyano-(1-naphthylmethyl)]-3,4-dihydro-5-ethyl-2-isopropylthiopyrimidin-4(3H)-one **(3k)**	A
118	2-Allylthio-6-[α-cyano-(1-naphthylmethyl)]-3,4-dihydro-5-ethylpyrimidin-4(3H)-one **(3m)**	A
119	2-(4-Chlorobenzoylmethylthio)-6-[α-cyano-(1-naphthylmethyl)]-3,4-dihydro-5-ethylpyrimidin-4(3H)-one **(3p)**	A
120	6-[α-Cyano-(1-naphthylmethyl)]-3,4-dihydro-5-isopropyl-2-(4-methoxybenzylthio)pyrimidin-4(3H)-one **(3u)**	A
121	6-(α-Cyano-2,6-dichlorobenzyl)-3,4-dihydro-5-isopropyl-2-(4-methoxybenzylthio)pyrimidin-4(3H)-one **(3w)**	A
122	2-Cyclopentylthio-3,4-dihydro-6-(1-naphthoyl)pyrimidin-4(3H)-one **(4b)**	A
123	2-Benzoylmethylthio-3,4-dihydro-6-(1-naphthoyl)pyrimidin-4(3H)-one **(4c)**	A
124	6-Benzoyl-3,4-dihydro-2-ethylthiopyrimidin-4(3H)-one **(4d)**	A
125	3,4-Dihydro-5-methyl-6-(1-naphthoyl)-2-propynylthiopyrimidin-4(3H)-one **(4f)**	A
126	2-Benzylthio-3,4-dihydro-5-ethyl-6-(1-naphthoyl)pyrimidin-4(3H)-one **(4j)**	A
127	2-Benzylthio-3,4-dihydro-5-ethyl-6-(2-naphthoyl)pyrimidin-4(3H)-one **(4k)**	A

[a]References used as source for data collection: **A**- L. Ji, F.E. Chen, E. De Clercq, J. Balzarini, and C. Pannecouque. Synthesis and anti-HIV-1 activity evaluation of 5-alkyl-2-alkylthio-6-(arylcarbonyl or alpha-cyanoarylmethyl)-3,4-dihydropyrimidin-4(3H)-ones as novel non-nucleoside HIV-1 reverse transcriptase inhibitors. J Med Chem. 50:1778-1786 (2007). **B**- G.F. Sun, X.X. Chen, F.E. Chen, Y.P. Wang, E. De Clercq, J. Balzarini, and C. Pannecouque. Nonnucleoside HIV-1 reverse-transcriptase inhibitors, part 5. Synthesis and anti-HIV-1 activity of novel 6-naphthylthio HEPT analogues. Chem Pharm Bull (Tokyo). 53:886-892 (2005). **C**- Y.Z. Xiong, F.E. Chen, J. Balzarini, E. De Clercq, and C. Pannecouque. Non-nucleoside HIV-1 reverse transcriptase inhibitors. Part 11: structural modulations of diaryltriazines with potent anti-HIV activity. Eur J Med Chem. 43:1230-1236 (2008). **D**- A. Ranise, A. Spallarossa, S. Schenone, O. Bruno, F. Bondavalli, L. Vargiu, T. Marceddu, M. Mura, P. La Colla, and A. Pani. Design, synthesis, SAR, and molecular modeling studies of acylthiocarbamates: a novel series of potent non-nucleoside HIV-1 reverse transcriptase inhibitors structurally related to phenethylthiazolylthiourea derivatives. J Med Chem. 46:768-781 (2003).

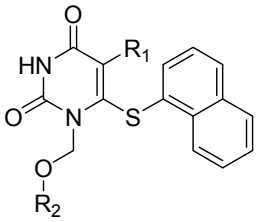
Table SI-2. Chemical structures, IC_{50} and CC_{50} values (expressed in μM) for HEPT analogues



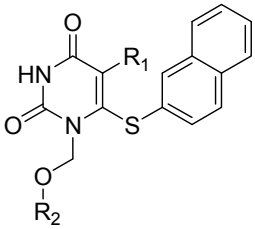
HEPT*
IC_{50}= 5.06; CC_{50}=405



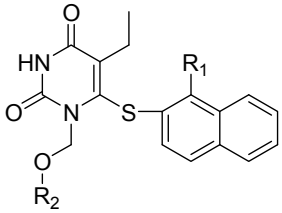
DDI*
IC_{50}= 5.37; CC_{50}=529



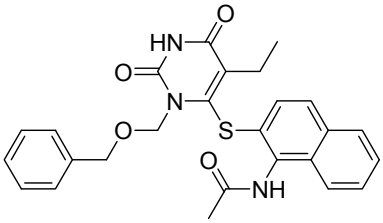
62-77; 101-107



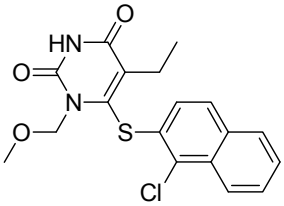
78-82; 108



83-86; 109-110



87
IC_{50}= 8.78; CC_{50}= 155.47

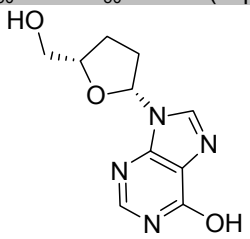


88
IC_{50}=22.66; CC_{50}=192.9

ID	R_1	R_2	IC_{50}	CC_{50}	ID	R_1	R_2	IC_{50}	CC_{50}
62	Me	Et	0.46	39.6	**80**	Et	Benzyl	0.65	154.5
63	Et	Me	0.5	38.78	**81**	*i*-Pr	Et	2.65	36.3
64	Et	3'-Methylbenzyl	0.21	28.73	**82**	*i*-Pr	Benzyl	0.83	187.34
65	*i*-Pr	Et	0.057	33.2	**83**	Me	NO_2	0.47	50.13
66	*i*-Pr	$CH_2CH_2OCH_3$	0.65	35.58	**84**	Benzyl	NO_2	0.065	34.19
67	*i*-Pr	c-Pr-CH_2	0.38	32.75	**85**	Me	NH_2	25.99	180.48
68	*i*-Pr	Benzyl	0.063	35.3	**86**	Et	NH_2	7.06	195.23
69	Pr	Et	1.86	37.11	**101**	Me	Benzyl	0.18	52.5
70	*i*-Bu	Et	0.83	32.57	**102**	Et	Et	0.15	38.4
71	Et	3'-Fluorobenzyl	2.96	32.57	**103**	Et	Benzyl	0.48	34.28
72	*i*-Bu	c-Hexyl-CH_2	0.092	30.96	**104**	*i*-Pr	Me	0.51	38.4
73	*i*-Bu	3'-Fluorobenzyl	12.53	171.92	**105**	*i*-Pr	3'-Methylbenzyl	0.25	29.6
74	*i*-Bu	4'-Fluorobenzyl	0.11	32.82	**106**	*i*-Bu	Benzyl	7.37	36.12
75	*i*-Bu	$PhCH_2CH_2$	0.067	30.23	**107**	Et	4'-Fluorobenzyl	0.21	31.47
76	*i*-Bu	CH_2CH_2OAc	2.77	31.4	**108**	*i*-Pr	Me	4.69	59.58
77	*i*-Bu	CH_2CH_2OH	0.23	77.07	**109**	Et	NO_2	0.099	43.89
78	Et	Me	22.02	210.09	**110**	Benzyl	NH_2	3.65	>37.88
79	Et	Et	3.37	42.19					

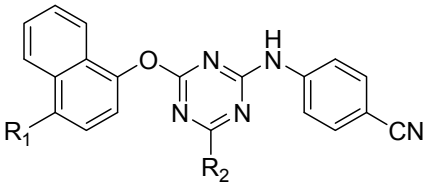
* Compound used as reference drug on anti-HIV activity and cytotoxicity assays.

Table SI-3. Chemical structures, IC_{50} and CC_{50} values (expressed in μM) for DATA analogues.

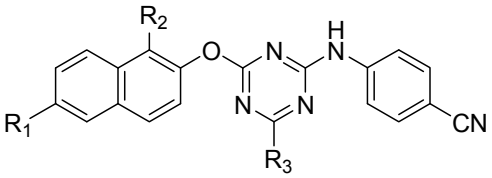


DDI*

IC_{50}= 5.37; CC_{50}=529



4-8; 89-90

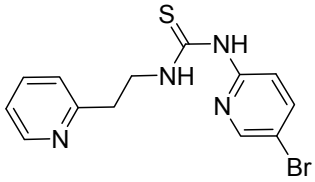


9-21, 91-92; 111-113

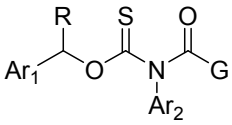
ID	R_1	R_2	IC_{50}	CC_{50}	ID	R_1	R_2	R_3	IC_{50}	CC_{50}
4	H	NHMe	0.186	30.33	**9**	H	H	NHMe	0.26	250
5	H	NH_2	0.107	39.07	**10**	H	H	NH_2	0.09	42.23
6	H	*n*-PrNH	2.7	29.46	**11**	Br	Br	NHEt	0.156	74.4
7	Cl	NHMe	0.093	196.9	**12**	Br	H	*i*-PrNH	1.065	41.54
8	Cl	NH_2	0.062	237	**13**	H	Br	NHMe	0.284	193.9
89	Cl	NHEt	0.733	207.7	**14**	Br	H	NHMe	0.0093	143.2
90	Cl	*n*-PrNH	1.63	181.6	**15**	Br	H	NH_2	0.0094	133
					16	Cl	H	NHEt	0.114	187.93
					17	Br	Br	NHMe	0.0184	92.83
					18	Cl	H	NHMe	0.0118	170.6
					19	Cl	H	NH_2	0.028	32.15
					20	H	Br	NH_2	0.0808	243
					21	Cl	H	N_3	0.06	256.32
					91	Br	Br	*n*-PrNH	0.256	134.44
					92	Br	Br	NH_2	0.0195	244
					111	Br	Br	*i*-PrNH	0.693	≥209
					112	Br	H	*n*-PrNH	0.203	>263.2
					113	Cl	H	*n*-PrNH	0.187	>290

* Compound used as reference drug on anti-HIV activity and cytotoxicity assays.

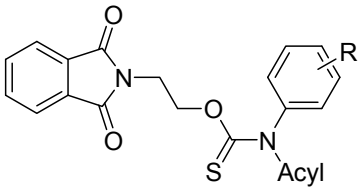
Table SI-4. Chemical structures, IC_{50} and CC_{50} values (expressed in μM) for ATC analogues.



Trovirdine*
IC_{50}= 0.02; CC_{50}= 60



22-34; 93-97; 114-115

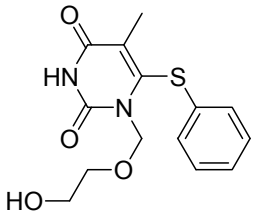


35-40

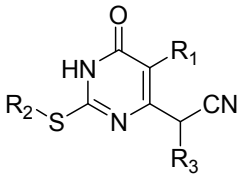
ID	Ar_1	R	Ar_2	G-CO	IC_{50}	CC_{50}
22	Phenyl	H	C_6H_5	2-furoyl	58	111
23	Benzyl	H	C_6H_5	*trans*-cinnamoyl	4.5	40.4
24	Benzyl	H	C_6H_5	Benzoyl	4.2	91
25	Benzyl	H	C_6H_5	4-chlorobenzoyl	4	43
26	Benzyl	H	C_6H_5	2-furoyl	4.3	102
27	Phenoxymethyl	H	C_6H_5	*trans*-cinnamoyl	7.7	66.6
28	Phenoxymethyl	H	C_6H_5	4-chlorobenzoyl	10.3	133
29	Phenoxymethyl	H	C_6H_5	2,4-dichlorobenzoyl	11.6	43
30	Phenoxymethyl	H	C_6H_5	3,5-dichlorobenzoyl	8.8	43
31	Phenoxymethyl	H	C_6H_5	2-furoyl	8.4	82
32	Phenoxymethyl	H	C_6H_5	2-thenoyl	8.6	125.2
33	Phenoxymethyl	CH_3	C_6H_5	Phenoxyacetyl	1.4	122.4
34	Phenoxymethyl	CH_3	4-NO_2-C_6H_5	2-thenoyl	6	63
93	Phenoxymethyl	H	C_6H_5	Phenoxyacetyl	6	103
94	Phenoxymethyl	H	C_6H_5	Benzoyl	8	122
95	Phenoxymethyl	H	C_6H_5	4-chloro-3-nitrobenzoyl	7.6	80
96	Phenoxymethyl	H	4-F-C_6H_5	Benzoyl	4	70.7
97	Phenoxymethyl	CH_3	C_6H_5	2-furoyl	1.3	51.5
114	2-furyl	H	C_6H_5	Benzoyl	>38.3	38.3
115	Phenoxythiomethyl	H	C_6H_5	Benzoyl	>44	44
	R		**Acyl**			
35	3-Br		2-thenoyl		1.2	53
36	3-NO_2		2-thenoyl		0.38	100
37	4-Cl		2-furoyl		0.007	41
38	4-I		2-thenoyl		0.01	18
39	4-NO_2		2-furoyl		0.008	18
40	4-NO_2		2-thenoyl		0.01	168

* Compound used as reference drug on anti-HIV activity and cytotoxicity assays.

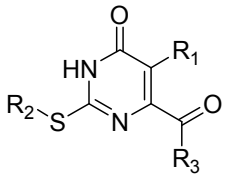
Table SI-5. Chemical structures, IC_{50} and CC_{50} values (expressed in μM) for DABO analogues.



1. HEPT*
IC_{50}= 5.06; CC_{50}=405



41-57; 116-121



58-61; 98-100; 122-127

ID	R_1	R_2	R_3	IC_{50}	CC_{50}
41	H	*i*-Pr	1-naphthyl	0.66	53.19
42	Me	Cyclopentyl	1-naphthyl	0.8	49.39
43	Me	Propynyl	1-naphthyl	2.7	46.09
44	Me	Benzoylmethyl	1-naphthyl	0.09	163.86
45	Me	4-chlorobenzoylmethyl	1-naphthyl	4.71	148.32
46	Me	*i*-Pr	Ph	0.002	10.81
47	Me	Cyclopentyl	Ph	6.34	58.95
48	Me	Benzoylmethyl	Ph	7.28	53.81
49	Me	Et	1-naphthyl	1.17	122.35
50	Et	Cyclopentyl	1-naphthyl	0.18	51.88
51	Et	Propynyl	1-naphthyl	4.85	47.99
52	Et	4-nitrobenzyl	1-naphthyl	1.25	24.52
53	Et	*i*-Pr	Ph	0.64	58.02
54	Et	4-nitrobenzyl	Ph	1.33	34.19
55	Et	4-nitrobenzyl	2,6-Cl_2-Ph	2.22	23.46
56	*i*-Pr	Cyclopentyl	1-naphthyl	1.34	7.32
57	*i*-Pr	4-methoxybenzyl	Ph	3.53	26.79
116	H	Cyclopentyl	1-naphthyl	>22.80	22.8
117	Et	*i*-Pr	1-naphthyl	0.24	>344.35
118	Et	Allyl	1-naphthyl	0.58	>346.26
119	Et	4-chlorobenzoylmethyl	1-naphthyl	≥35.73	116.66
120	*i*-Pr	4-methoxybenzyl	1-naphthyl	>2.09	2.09
121	*i*-Pr	4-methoxybenzyl	2,6-Cl_2-Ph	>23.53	23.53
58	H	Et	1-naphthyl	37.58	241.61
59	Me	*i*-Pr	Ph	2.05	263.19
60	Et	Et	1-naphthyl	27.84	196.54
61	Et	*i*-Pr	1-naphthyl	6.79	178.92
98	Et	Allyl	1-naphthyl	16.94	210.37
99	Et	Cyclopentyl	Ph	4.69	60.37
100	*i*-Pr	Cyclopentyl	1-naphthyl	10.87	24.69
122	H	Cyclopentyl	1-naphthyl	>21.03	21.03
123	H	Benzoylmethyl	1-naphthyl	>304.41	>304.41
124	H	Et	Ph	>246.65	246.65
125	Me	Propynyl	1-naphthyl	>123.56	123.56
126	Et	Benzyl	1-naphthyl	>312.50	>312.50
127	Et	Benzyl	2-naphthyl	≥60.75	157.43

* Compound used as reference drug on anti-HIV activity and cytotoxicity assays.

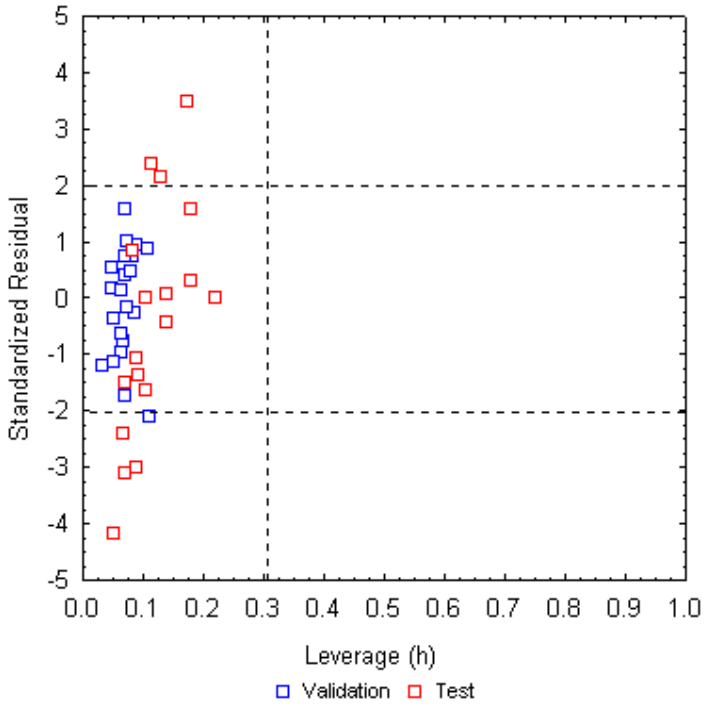


Figure SI-1. Checking the compliance of the validation and external test set compounds within the applicability domain of the $-\log CC_{50}$ model.

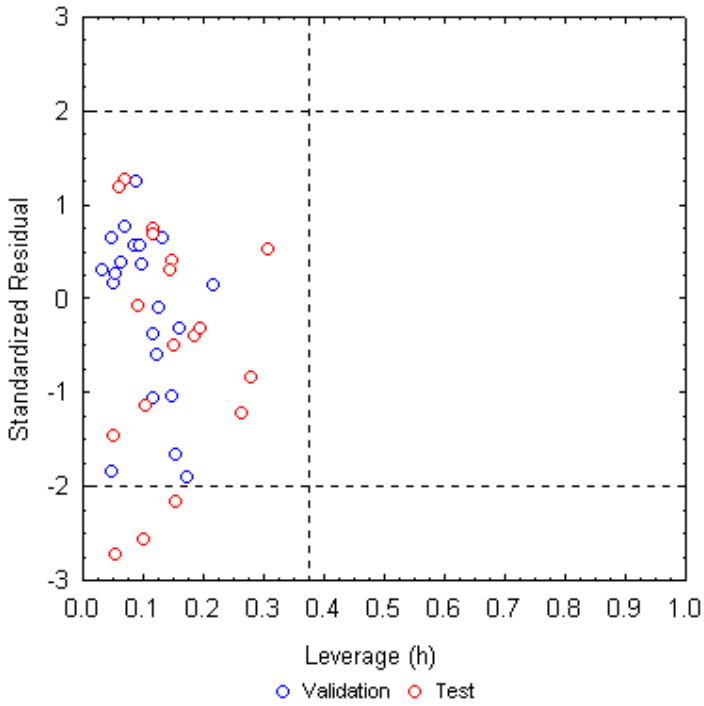


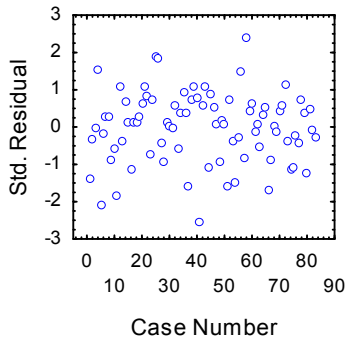
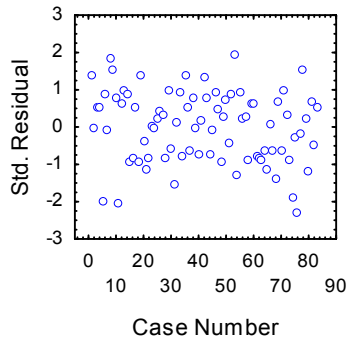
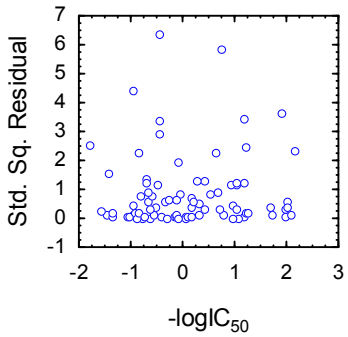
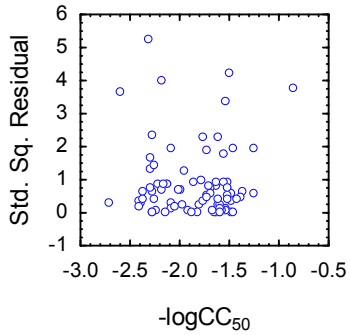
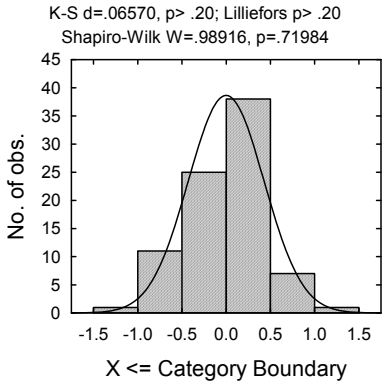
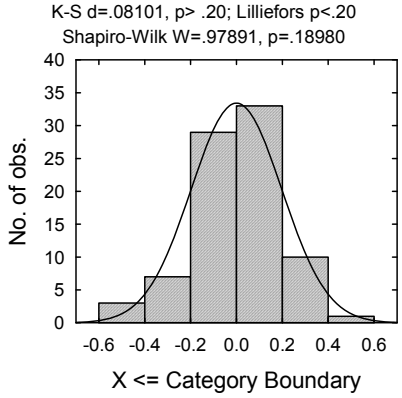
Figure SI-2. Checking the compliance of the validation and external test set compounds within the applicability domain of the $-\log IC_{50}$ model.

MLR Parametrical Assumptions

This section provides details about the checking of the pre-adopted parametric assumptions, a very important aspect in the application of linear multivariate statistical-based approaches (MLR techniques) (1). In fact, once the linear regression model has been set up, it is very important to check the parametric assumptions to assure the validity of extrapolation from the sample to the population. These include the linearity of the modeled property, normal distribution as well as the homoscedasticity and non-multicollinearity descriptors. Notice that severe violations of one or various of these assumptions can markedly compromise the reliability of the predictions resulting from our MLR models (1).

We first check the linearity hypothesis by looking at the distribution of the standardized residuals for all cases. Indeed the plots in Tables SI-6 (1st row) do not show any specific pattern, reinforcing the idea that our models do not exhibit a non-linear dependence (1). Next, we check the hypothesis of homoscedasticity (*i.e.*: homogeneity of variance of the variables), which can be confirmed by simply plotting the square of standardized residuals related to each dependent variable (1) (2nd row of plots in Tables SI-6). These plots reveal significant scatter of points, without any systematic pattern, *post-mortem* validating the pre-adopted assumption of homoscedasticity for all the PMs. The plots also provide a check for the no auto-correlation of the residuals. Moving on to the hypothesis of normally distributed residuals, one can easily confirm that the residuals follow a normal distribution by applying the Kolmogorov-Smirnov and Lilliefors statistical test (3rd row of Tables SI-6). In addition, as the term related to the error (represented by residuals) is not included in the MLR equations, the mean must be zero what actually occurs (check 4th row of Tables SI-6). The last aspect deserving special attention is the degree of multicollinearity among the variables. Highly collinear variables may be identified by examining their pair-correlations (R_{ij}). Only three pair of predictors included in the –logIC50 MLR model exhibit a value of R_{ij} higher than 0.7 indicating that the multicollinearity is not a serious problem for our models. The common interpretation of a regression coefficient as measuring the change in the expected value of the response variable, when the given predictor variable is increased by one unit while all other predictor variables are held constant, is not fully applicable when multicollinearity exists ($R \geq 0.7$). However, the predictive ability of the model is not affected at all in this situation (2).

Table SI-6. Checking the main parametric assumptions related to the MLR models used to fit the desirability functions.

	***-logIC$_{50}$* MLR Model**	***-logCC$_{50}$* MLR Model**
Linearity		
Homoscedasticity		
Normality of Residuals		
Res. Mean = 0	4.24×10^{-14}	-5.55×10^{-16}
Non Collinearity	See Correlation Matrix	See Correlation Matrix

Correlation matrix. $-\log IC_{50}$ MLR Model.										
	MAXDP	***X1sol***	***SIC0***	***GATS1p***	***ESpm15r***	***Eig1v***	***Ks***	***R8u+***	***R8m***	***N-075***
MAXDP	1.000									
X1sol	0.043	1.000								
SIC0	-0.310	-0.168	1.000							
GATS1p	-0.431	-0.285	-0.192	1.000						
ESpm15r	0.539	0.3453	-0.098	-0.226	1.000					
Eig1v	-0.010	0.858	-0.032	-0.143	0.017	1.000				
Ks	-0.713	0.061	0.402	0.289	-0.379	0.209	1.000			
R8u+	-0.510	-0.320	0.426	0.064	-0.310	-0.290	0.519	1.000		
R8m	0.413	0.602	0.023	-0.438	0.464	0.518	-0.314	-0.307	1.000	
N-075	-0.951	0.098	0.273	0.434	-0.470	0.166	0.775	0.465	-0.349	1.000

Correlation matrix. $-\log CC_{50}$ MLR Model.								
	MATS3m	***MATS5e***	***RDF070p***	***Mor18e***	***H8e***	***R8p***	***nROH***	***C-003***
MATS3m	1.000							
MATS5e	0.296	1.000						
RDF070p	0.008	0.273	1.000					
Mor18e	0.249	0.268	-0.044	1.000				
H8e	-0.576	-0.202	0.224	-0.297	1.000			
R8p	-0.415	0.113	0.630	-0.310	0.565	1.000		
nROH	-0.166	-0.025	-0.061	0.415	-0.014	-0.146	1.000	
C-003	-0.113	-0.071	0.254	0.049	0.261	0.371	0.031	1.000

Table SI-7. Δ_i, ${}^{D}\Delta_i$ and D_i values of the library of compounds used for ranking.

ID	Δ_i	${}^{D}\Delta_i$	$D_{IC50\text{-}CC50}$	ID	Δ_i	${}^{D}\Delta_i$	$D_{IC50\text{-}CC50}$	ID	Δ_i	${}^{D}\Delta_i$	$D_{IC50\text{-}CC50}$
1	2.257	0.487	0.503	**45**	3.056	0.313	0.442	**89**	1.722	0.559	0.615
2	-0.097	1.000	0.514	**46**	---	---	---	**90**	1.660	0.575	0.545
3	2.778	0.374	0.659	**47**	3.096	0.304	0.346	**91**	2.333	0.402	0.639
4	2.294	0.479	0.460	**48**	3.416	0.235	0.327	**92**	1.535	0.607	0.852
5	1.972	0.549	0.522	**49**	2.504	0.433	0.534	**93**	2.791	0.285	0.394
6	2.241	0.491	0.333	**50**	2.890	0.349	0.541	**94**	3.128	0.199	0.380
7	1.555	0.640	0.741	**51**	2.698	0.391	0.348	**95**	2.977	0.237	0.355
8	1.557	0.640	0.785	**52**	3.507	0.215	0.347	**96**	2.952	0.244	0.396
9	---	---	---	**53**	2.100	0.521	0.492	**97**	2.962	0.241	0.438
10	2.052	0.532	0.542	**54**	2.416	0.453	0.388	**98**	3.367	0.138	0.327
11	1.887	0.568	0.596	**55**	2.919	0.343	0.314	**99**	3.904	0.000	0.371
12	1.796	0.588	0.424	**56**	2.264	0.486	0.000	**100**	3.201	0.180	0.230
13	1.148	0.729	0.672	**57**	2.296	0.479	0.307	**101**	2.383	0.390	0.543
14	1.163	0.725	0.820	**58**	4.493	0.000	0.198	**102**	2.189	0.439	0.506
15	0.790	0.807	0.810	**59**	3.576	0.200	0.557	**103**	2.363	0.395	0.438
16	1.633	0.623	0.724	**60**	3.483	0.220	0.250	**104**	2.456	0.371	0.451
17	1.336	0.688	0.728	**61**	3.372	0.244	0.421	**105**	2.082	0.467	0.444
18	1.045	0.751	0.832	**62**	2.753	0.379	0.460	**106**	2.501	0.359	0.292
19	---	---	---	**63**	3.151	0.292	0.453	**107**	2.254	0.423	0.461
20	1.246	0.707	0.772	**64**	2.265	0.485	0.446	**108**	3.584	0.082	0.370
21	1.202	0.717	0.796	**65**	1.601	0.630	0.521	**109**	2.799	0.283	0.544
22	3.104	0.303	0.000	**66**	3.002	0.325	0.429	**110**	2.160	0.567	0.343
23	2.525	0.429	0.336	**67**	2.228	0.493	0.442	**111**	2.140	0.571	0.620
24	2.942	0.338	0.414	**68**	2.294	0.479	0.527	**112**	1.856	0.628	0.724
25	2.672	0.397	0.350	**69**	2.691	0.393	0.380	**113**	1.714	0.656	0.739
26	2.923	0.342	0.421	**70**	2.535	0.427	0.405	**114**	2.747	0.449	0.133
27	2.984	0.329	0.340	**71**	2.351	0.467	0.339	**115**	3.435	0.311	0.113
28	2.993	0.327	0.360	**72**	2.254	0.488	0.491	**116**	3.268	0.345	0.166
29	2.853	0.357	0.272	**73**	---	---	---	**117**	2.340	0.531	0.740
30	2.595	0.413	0.294	**74**	2.582	0.416	0.493	**118**	2.747	0.449	0.678
31	3.269	0.267	0.348	**75**	2.157	0.509	0.498	**119**	3.465	0.305	0.186
32	2.909	0.345	0.375	**76**	2.906	0.346	0.339	**120**	2.674	0.464	0.000
33	2.632	0.405	0.521	**77**	2.179	0.504	0.581	**121**	2.411	0.517	0.165
34	2.491	0.436	0.356	**78**	4.028	0.101	0.290	**122**	4.144	0.169	0.167
35	2.124	0.516	0.446	**79**	3.713	0.170	0.359	**123**	3.937	0.211	0.000
36	2.260	0.486	0.583	**80**	2.730	0.384	0.596	**124**	4.552	0.088	0.000
37	1.948	0.554	0.634	**81**	3.132	0.297	0.358	**125**	4.989	0.000	0.000
38	2.432	0.449	0.449	**82**	2.430	0.450	0.597	**126**	3.824	0.234	0.000
39	2.580	0.417	0.455	**83**	2.456	0.444	0.490	**127**	4.461	0.106	0.000
40	1.517	0.648	0.839	**84**	1.890	0.567	0.521				
41	2.750	0.380	0.479	**85**	3.370	0.245	0.258				
42	2.638	0.404	0.460	**86**	2.690	0.393	0.423				
43	2.499	0.434	0.382	**87**	2.435	0.448	0.387				
44	---	---	---	**88**	2.568	0.419	0.282				

Table SI-8. Δ_i-based ranked list of 12 NNRTIs with favorable pharmaceutical profile and 432 DUD decoys.

Rank	ID	Class[a]	Δ_i	Rank	ID	Class[a]	Δ_i	Rank	ID	Class[a]	Δ_i
1	ZINC01545897	-	0.235	56	ZINC00478707	-	1.838	111	ZINC02868137	-	2.546
2	ZINC02683840	-	0.355	57	ZINC00653432	-	1.842	112	ZINC03463934	-	2.560
3	ZINC02334601	-	0.886	**58**	**112**	**+**	**1.856**	113	ZINC00374482	-	2.565
4	ZINC01553345	-	0.929	59	ZINC00614305	-	1.866	114	ZINC03366422	-	2.580
5	ZINC00464794	-	0.951	60	ZINC01405602	-	1.879	115	ZINC00052016	-	2.585
6	ZINC02889052	-	0.987	61	ZINC00534216	-	1.884	116	ZINC00110010	-	2.594
7	ZINC00041945	-	1.075	62	ZINC00573652	-	1.885	117	ZINC01281458	-	2.603
8	ZINC00041945	-	1.084	63	ZINC00540184	-	1.887	118	ZINC01281458	-	2.603
9	ZINC02889051	-	1.113	64	ZINC01002344	-	1.897	119	ZINC00502655	-	2.611
10	ZINC01010351	-	1.129	65	ZINC02632868	-	1.897	120	ZINC02728470	-	2.615
11	ZINC01340777	-	1.212	66	ZINC01856833	-	1.899	121	ZINC00856243	-	2.638
12	ZINC00756911	-	1.280	67	ZINC00052260	-	1.919	122	ZINC00667522	-	2.646
13	ZINC00009466	-	1.350	68	ZINC00161667	-	1.933	123	ZINC00272219	-	2.648
14	ZINC02798358	-	1.373	69	ZINC02741855	-	1.976	124	ZINC02728470	-	2.660
15	ZINC01856833	-	1.383	70	ZINC00794384	-	1.988	125	ZINC00550540	-	2.672
16	ZINC00223283	-	1.395	71	ZINC01965571	-	1.989	126	ZINC00629315	-	2.693
17	ZINC00114582	-	1.415	72	ZINC03231072	-	2.089	127	ZINC01424359	-	2.707
18	ZINC00009466	-	1.452	73	ZINC01780133	-	2.089	128	ZINC00670303	-	2.712
19	ZINC00185357	-	1.473	74	ZINC00290760	-	2.128	129	ZINC00590227	-	2.722
20	**92**	**+**	**1.535**	**75**	**111**	**+**	**2.140**	130	ZINC00212739	-	2.740
21	ZINC02552260	-	1.542	76	ZINC01780134	-	2.152	131	ZINC03404509	-	2.742
22	ZINC00626314	-	1.550	77	ZINC02887392	-	2.167	**132**	**118**	**+**	**2.747**
23	ZINC00125154	-	1.550	78	ZINC00649848	-	2.180	133	ZINC00332082	-	2.756
24	ZINC01496341	-	1.570	**79**	**102**	**+**	**2.189**	134	ZINC00628111	-	2.757
25	ZINC00125154	-	1.577	80	ZINC00013204	-	2.195	135	ZINC00367881	-	2.774
26	ZINC00230762	-	1.578	81	ZINC00013204	-	2.206	136	ZINC00587144	-	2.777
27	ZINC00627273	-	1.579	82	ZINC01091132	-	2.220	137	ZINC00678020	-	2.779
28	ZINC00627273	-	1.582	83	ZINC02797879	-	2.224	138	ZINC00229279	-	2.780
29	ZINC00626314	-	1.591	84	ZINC03271480	-	2.228	139	ZINC00628111	-	2.780
30	ZINC00268377	-	1.610	85	ZINC00947566	-	2.231	140	ZINC00856243	-	2.795
31	ZINC01959294	-	1.620	86	ZINC00444291	-	2.247	141	ZINC03401018	-	2.799
32	ZINC00533154	-	1.632	87	ZINC00804958	-	2.259	**142**	**109**	**+**	**2.799**
33	ZINC02784877	-	1.643	88	ZINC02740807	-	2.262	143	ZINC00678020	-	2.801
34	ZINC02764981	-	1.653	89	ZINC02796351	-	2.276	144	ZINC00973466	-	2.802
35	**90**	**+**	**1.660**	90	ZINC02265634	-	2.279	145	ZINC00212739	-	2.803
36	ZINC00161385	-	1.684	91	ZINC02087110	-	2.284	146	ZINC00305200	-	2.820
37	ZINC00533153	-	1.689	92	ZINC02570906	-	2.297	147	ZINC00549463	-	2.842
38	ZINC00526392	-	1.697	93	ZINC03251649	-	2.317	148	ZINC00090974	-	2.847
39	ZINC00526390	-	1.700	**94**	**91**	**+**	**2.333**	149	ZINC01065454	-	2.857
40	ZINC02887396	-	1.705	95	ZINC02801176	-	2.336	150	ZINC00092980	-	2.858
41	**113**	**+**	**1.714**	**96**	**117**	**+**	**2.340**	151	ZINC00367881	-	2.862
42	ZINC00526394	-	1.719	97	ZINC01050386	-	2.345	152	ZINC00079509	-	2.865
43	ZINC02798460	-	1.720	98	ZINC00377349	-	2.359	153	ZINC01599204	-	2.867
44	**89**	**+**	**1.722**	99	ZINC00347899	-	2.376	154	ZINC00088513	-	2.874
45	ZINC01124240	-	1.741	100	ZINC00477907	-	2.376	155	ZINC03242936	-	2.880
46	ZINC00526388	-	1.751	101	ZINC00347897	-	2.377	156	ZINC00092980	-	2.886
47	ZINC03268162	-	1.756	102	ZINC00477906	-	2.378	157	ZINC00124651	-	2.886
48	ZINC01012054	-	1.787	**103**	**101**	**+**	**2.383**	158	ZINC00442591	-	2.891
49	ZINC00005434	-	1.796	104	ZINC02353769	-	2.397	159	ZINC00079509	-	2.893
50	ZINC01023177	-	1.797	105	ZINC00140556	-	2.402	160	ZINC02760792	-	2.904
51	ZINC02795155	-	1.797	106	ZINC01801666	-	2.410	161	ZINC01006309	-	2.922
52	ZINC03211805	-	1.799	107	ZINC00342537	-	2.442	162	ZINC00063382	-	2.929
53	ZINC00138080	-	1.812	108	ZINC03153754	-	2.469	163	ZINC02796205	-	2.936
54	ZINC00653432	-	1.812	109	ZINC00424537	-	2.477	164	ZINC02760339	-	2.949
55	ZINC00587907	-	1.822	110	ZINC03464407	-	2.485	165	ZINC02394836	-	2.955

[a] **+:** NNRTI candidate with favorable pharmaceutical profile; -: Decoy.

Table SI-8. (*continued…*)

Rank	ID	Class[a]	Δ_i	Rank	ID	Class[a]	Δ_i	Rank	ID	Class[a]	Δ_i
166	ZINC02394834	-	2.981	216	ZINC02620592	-	3.310	266	ZINC02131933	-	3.583
167	ZINC01080693	-	2.990	217	ZINC00442700	-	3.317	267	ZINC00253706	-	3.584
168	ZINC00476882	-	2.991	218	ZINC02930611	-	3.318	268	ZINC01453047	-	3.585
169	ZINC00367887	-	2.993	219	ZINC02723816	-	3.322	269	ZINC00955089	-	3.596
170	ZINC02431635	-	2.999	220	ZINC00418213	-	3.325	270	ZINC00362176	-	3.602
171	ZINC00057774	-	3.001	221	ZINC01823161	-	3.340	271	ZINC01240239	-	3.620
172	ZINC03117934	-	3.008	222	ZINC01051738	-	3.342	272	ZINC00446819	-	3.638
173	ZINC00441722	-	3.043	223	ZINC03439950	-	3.346	273	ZINC00082028	-	3.645
174	ZINC02717965	-	3.055	224	ZINC03378763	-	3.348	274	ZINC00619588	-	3.653
175	ZINC00217414	-	3.056	225	ZINC00359364	-	3.357	275	ZINC00515566	-	3.660
176	ZINC00124653	-	3.058	226	ZINC00060691	-	3.368	276	ZINC02217197	-	3.665
177	ZINC00502653	-	3.059	227	ZINC00355692	-	3.375	277	ZINC03337427	-	3.673
178	ZINC00918931	-	3.060	228	ZINC00425130	-	3.377	278	ZINC03337430	-	3.674
179	ZINC00367887	-	3.070	229	ZINC00613650	-	3.381	279	ZINC00437873	-	3.680
180	ZINC02717965	-	3.085	230	ZINC00060692	-	3.385	280	ZINC01284917	-	3.692
181	ZINC00100218	-	3.087	231	ZINC00880795	-	3.389	281	ZINC00609495	-	3.696
182	ZINC00175529	-	3.094	232	ZINC02313343	-	3.389	282	ZINC03397220	-	3.702
183	ZINC00100217	-	3.110	233	ZINC00043485	-	3.401	283	ZINC00181957	-	3.719
184	ZINC02554065	-	3.118	234	ZINC01240300	-	3.409	284	ZINC00830170	-	3.727
185	ZINC00968562	-	3.119	235	ZINC00930756	-	3.412	285	ZINC00257428	-	3.732
186	ZINC01066688	-	3.125	236	ZINC00132231	-	3.415	286	ZINC03455248	-	3.736
187	ZINC02996697	-	3.133	237	ZINC00168555	-	3.418	287	ZINC00483964	-	3.741
188	ZINC02751969	-	3.134	238	ZINC03370391	-	3.422	288	ZINC00146575	-	3.745
189	ZINC01208576	-	3.135	239	ZINC02133800	-	3.422	289	ZINC00206253	-	3.746
190	ZINC00126671	-	3.144	240	ZINC00536316	-	3.427	290	ZINC00512947	-	3.751
191	ZINC02637323	-	3.180	241	ZINC03293975	-	3.429	291	ZINC03041286	-	3.764
192	ZINC03347131	-	3.185	242	ZINC00206257	-	3.429	292	ZINC02795613	-	3.768
193	ZINC03301981	-	3.186	243	ZINC03453581	-	3.433	293	ZINC01254638	-	3.773
194	ZINC00295845	-	3.190	244	ZINC00425212	-	3.435	294	ZINC00362304	-	3.779
195	ZINC00397739	-	3.190	245	ZINC00793931	-	3.440	295	ZINC01396436	-	3.783
196	ZINC01363169	-	3.193	246	ZINC01017382	-	3.440	296	ZINC00383373	-	3.789
197	ZINC02620593	-	3.199	247	ZINC00002820	-	3.444	297	ZINC00212477	-	3.805
198	ZINC00203966	-	3.201	248	ZINC00611671	-	3.469	298	ZINC02762792	-	3.816
199	ZINC01807569	-	3.202	249	ZINC03271480	-	3.476	299	ZINC02795457	-	3.824
200	ZINC00295845	-	3.208	250	ZINC02787988	-	3.481	300	ZINC01148852	-	3.826
201	ZINC00126675	-	3.216	251	ZINC02800427	-	3.483	301	ZINC02800075	-	3.852
202	ZINC02795292	-	3.218	252	ZINC03439837	-	3.494	302	ZINC00427326	-	3.853
203	ZINC00319875	-	3.228	253	ZINC00206254	-	3.508	303	ZINC03439911	-	3.867
204	ZINC02796206	-	3.232	254	ZINC00181958	-	3.513	304	ZINC01134533	-	3.880
205	ZINC03086123	-	3.246	255	ZINC00036045	-	3.522	305	ZINC03453578	-	3.882
206	ZINC00261521	-	3.254	256	ZINC03439928	-	3.532	306	ZINC00725836	-	3.885
207	ZINC02533264	-	3.262	257	ZINC03453781	-	3.534	307	ZINC03041273	-	3.889
208	ZINC02861945	-	3.267	258	ZINC03217270	-	3.547	308	ZINC02402393	-	3.900
209	ZINC00188300	-	3.270	259	ZINC02637498	-	3.547	309	ZINC03317791	-	3.913
210	ZINC02795291	-	3.282	260	ZINC00080410	-	3.547	310	ZINC02199758	-	3.915
211	ZINC00411264	-	3.290	261	ZINC00002820	-	3.574	311	ZINC03283331	-	3.915
212	ZINC00213528	-	3.297	262	ZINC01121160	-	3.575	312	ZINC00450736	-	3.917
213	ZINC00233029	-	3.298	263	ZINC03148025	-	3.576	313	ZINC00973242	-	3.918
214	ZINC01051747	-	3.301	264	ZINC02787703	-	3.579	314	ZINC02199757	-	3.922
215	ZINC03463939	-	3.303	265	ZINC02675831	-	3.582	315	ZINC00188740	-	3.926

[a] **+:** NNRTI candidate with favorable pharmaceutical profile; -: Decoy.

Table SI-8. (*continued…*)

Rank	ID	Class[a]	Δ_i	Rank	ID	Class[a]	Δ_i	Rank	ID	Class[a]	Δ_i
316	ZINC00429167	-	3.926	359	ZINC01994281	-	4.169	402	ZINC01053768	-	4.474
317	ZINC00212487	-	3.931	360	ZINC00052551	-	4.173	403	ZINC02794621	-	4.477
318	ZINC01447889	-	3.932	361	ZINC03453777	-	4.177	404	ZINC00487273	-	4.478
319	ZINC00429166	-	3.936	362	ZINC01399041	-	4.179	405	ZINC00237924	-	4.485
320	ZINC01067033	-	3.950	363	ZINC00038067	-	4.187	406	ZINC00223980	-	4.517
321	ZINC02309223	-	3.952	364	ZINC01066008	-	4.189	407	ZINC00629127	-	4.555
322	ZINC00549464	-	3.952	365	ZINC02620382	-	4.213	408	ZINC01062726	-	4.566
323	ZINC00413812	-	3.957	366	ZINC00450843	-	4.213	409	ZINC00260900	-	4.578
324	ZINC00090765	-	3.962	367	ZINC01216594	-	4.219	410	ZINC03384857	-	4.583
325	ZINC00433154	-	3.969	368	ZINC00307143	-	4.222	411	ZINC00397717	-	4.590
326	ZINC02555597	-	3.988	369	ZINC01281458	-	4.237	412	ZINC03399461	-	4.591
327	ZINC00265166	-	3.998	370	ZINC00918934	-	4.239	413	ZINC00616701	-	4.593
328	ZINC00425133	-	4.004	371	ZINC03441346	-	4.249	414	ZINC03086127	-	4.612
329	ZINC00359366	-	4.009	372	ZINC02213527	-	4.259	415	ZINC02294241	-	4.618
330	ZINC03455235	-	4.024	373	ZINC00223347	-	4.267	416	ZINC00497871	-	4.620
331	ZINC00469435	-	4.026	374	ZINC01476114	-	4.268	417	ZINC01218306	-	4.628
332	ZINC00554737	-	4.037	375	ZINC03328237	-	4.287	418	ZINC03217249	-	4.631
333	ZINC03250847	-	4.040	376	ZINC02746950	-	4.297	419	ZINC03453783	-	4.634
334	ZINC03401021	-	4.044	377	ZINC01180224	-	4.313	420	ZINC01288087	-	4.637
335	ZINC03322691	-	4.052	378	ZINC00487270	-	4.314	421	ZINC02635859	-	4.646
336	ZINC00330856	-	4.054	379	ZINC00348146	-	4.324	422	ZINC01614679	-	4.657
337	ZINC00462543	-	4.066	380	ZINC02610066	-	4.328	423	ZINC00298445	-	4.662
338	ZINC00101922	-	4.068	381	ZINC00342159	-	4.329	424	ZINC00206451	-	4.664
339	ZINC03173621	-	4.069	382	ZINC00257585	-	4.332	425	ZINC02879179	-	4.698
340	ZINC01202925	-	4.077	383	ZINC00476728	-	4.341	426	ZINC01071697	-	4.723
341	ZINC00880796	-	4.078	384	ZINC00568380	-	4.341	427	ZINC03077377	-	4.735
342	ZINC02195911	-	4.080	385	ZINC00257585	-	4.350	428	ZINC00564557	-	4.746
343	ZINC00055670	-	4.084	386	ZINC00470268	-	4.368	429	ZINC00049673	-	4.752
344	ZINC01122413	-	4.090	387	ZINC00478808	-	4.386	430	ZINC00563878	-	4.755
345	ZINC00903785	-	4.094	388	ZINC01004491	-	4.387	431	ZINC00457738	-	4.777
346	ZINC00536317	-	4.097	389	ZINC00146513	-	4.387	432	ZINC00179800	-	4.792
347	ZINC02620381	-	4.101	390	ZINC00365579	-	4.392	433	ZINC01091255	-	4.792
348	ZINC01364053	-	4.112	391	ZINC00267905	-	4.395	434	ZINC00038372	-	4.810
349	ZINC00031486	-	4.115	392	ZINC01741786	-	4.399	435	ZINC00179798	-	4.930
350	ZINC01202928	-	4.120	393	ZINC00609573	-	4.410	436	ZINC03372459	-	4.989
351	ZINC00412580	-	4.121	394	ZINC00263725	-	4.414	437	ZINC02521888	-	5.016
352	ZINC00101926	-	4.122	395	ZINC00381496	-	4.427	438	ZINC01054638	-	5.052
353	ZINC02718985	-	4.124	396	ZINC02796638	-	4.427	439	ZINC02319147	-	5.071
354	ZINC00103251	-	4.141	397	ZINC01437599	-	4.428	440	ZINC01437610	-	5.072
355	ZINC01091256	-	4.142	398	ZINC02889026	-	4.453	441	ZINC02639622	-	5.172
356	ZINC03453775	-	4.152	399	ZINC01810037	-	4.459	442	ZINC03283331	-	5.334
357	ZINC01994283	-	4.162	400	ZINC02868569	-	4.467	443	ZINC00067979	-	5.358
358	ZINC02718985	-	4.164	401	ZINC01399040	-	4.471	444	ZINC01393190	-	5.542

[a] **+:** NNRTI candidate with favorable pharmaceutical profile; -: Decoy.

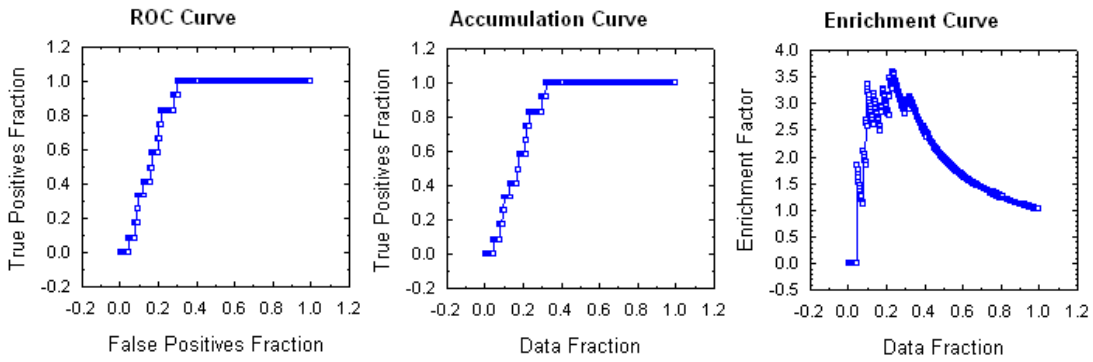


Figure SI-3. ROC, accumulation, and enrichment curves for the Δ_l-based ranking of the data set collected form DUD.

References

1. Stewart J, Gill L. Econometrics. 2nd edition ed. Allan P, editor. London: Prentice Hall; 1998.
2. Kutner MH, Nachtsheim CJ, Neter J, Li W. Multicollinearity and its effects. Applied Linear Statistical Models. Fifth ed. New York: McGraw Hill; 2005. p. 278-89.

ANNEX IV

Chem Biol Drug Des 2010; **75:** 607–618

Research Article

doi: 10.1111/j.1747-0285.2010.00971.x

Multidimensional Drug Design: Simultaneous Analysis of Binding and Relative Efficacy Profiles of N^6-substituted-4′-thioadenosines A_3 Adenosine Receptor Agonists

Maykel Cruz-Monteagudo[1,2,3,4,*], M. Natália D. S. Cordeiro[5], Marta Teijeira[6], Maykel P. González[4] and Fernanda Borges[1,*]

[1]*Department of Chemistry, Faculty of Sciences, University of Porto, 4169-007 Porto, Portugal*
[2]*Department of Organic Chemistry, Faculty of Pharmacy, University of Porto, 4150-047 Porto, Portugal*
[3]*Applied Chemistry Research Center, Faculty of Chemistry and Pharmacy, Central University of ''Las Villas'', Santa Clara 54830, Cuba*
[4]*Molecular Simulation and Drug Design Group, Chemical Bioactive Center, Central University of ''Las Villas'', Santa Clara 54830, Cuba*
[5]*REQUIMTE, Department of Chemistry, Faculty of Sciences, University of Porto, 4169-007 Porto, Portugal*
[6]*Department of Organic Chemistry, Vigo University, C.P. 36310 Vigo, Spain*
** Corresponding authors: Maykel Cruz-Monteagudo, maikelcm@uclv.edu.cu, gmailkelcm@yahoo.es; Fernanda Borges, mfernandamborges@gmail.com*

Desirability theory (DT) is a well-known multi-criteria decision-making approach. In this work, DT is employed as a prediction model (PM) interpretation tool to extract useful information on the desired trade-offs between binding and relative efficacy of N^6-substituted-4′-thioadenosines A_3 adenosine receptor (A_3AR) agonists. At the same time, it was shown the usefulness of a parallel but independent approach providing a feedback on the reliability of the combination of properties predicted as a unique desirability value. The appliance of belief theory allowed the quantification of the reliability of the predicted desirability of a compound according to two inverse and independent but complementary prediction approaches. This information is proven to be useful as a ranking criterion in a ligand-based virtual screening study. The development of a linear PM of the A_3AR agonists overall desirability allows finding significant clues based on simple molecular descriptors. The model suggests a relevant role of the type of substituent on the N^6 position of the adenine ring that in general contribute to reduce the flexibility and hydrophobicity of the lead compound. The mapping of the desirability function derived of the PM offers specific information such as the shape and optimal size of the N^6 substituent. The model herein developed allows a simultaneous analysis of both binding and relative efficacy profiles of A_3AR agonists. The information retrieved guides the theoretical design and assembling of a combinatorial library suitable for filtering new N^6-substituted-4′-thioadenosines A_3AR agonist candidates with simultaneously improved binding and relative efficacy profiles. The utility of the desirability/belief-based proposed virtual screening strategy was deduced from our training set. Based on the overall results, it is possible to assert that the combined use of desirability and belief theories in computational medicinal chemistry research can aid the discovery of A_3AR agonist candidates with favorable balance between binding and relative efficacy profiles.

Key words: A_3 adenosine receptor agonists, belief theory, Chemoinformatics, desirability theory, drug discovery, ligand-based virtual screening, simultaneous analysis

Received 22 October 2009, revised 26 February 2010 and accepted for publication 1 March 2010

Adenosine receptors (ARs) are G-protein-coupled receptors, consisting of A_1, A_2A, A_2B, and A_3 subtypes, that are activated by the endogenous agonist adenosine and blocked by natural antagonists, such as caffeine and theophylline (1). A_1 and A_3 subtypes are coupled to GI/O proteins, while A_2A and A_2B subtypes are GS protein-coupled.

There is growing evidence that ARs could be promising therapeutic targets in a wide range of pathologies (1–6). In particular, A_3AR agonists have shown to be useful to prevent ischemic damage in the brain and heart and as anti-inflammatory, anticancer, and myeloprotective agents (7–11).

Although ARs are becoming important targets in drug design and development, several problems complicate the development of new AR agonists. Kim and Jacobson (12) point out several reasons for the bottleneck in this area:

(a) The ubiquitous expression of ARs in the body would result in diverse side-effects.

(b) The low density of a given receptor subtype in a targeted tissue may reduce its desired effect in the treatment of certain diseases (8).

(c) In many cases, nucleoside derivatives have lowered maximal efficacy at the A_3AR and, consequently, behave as a partial agonist or antagonist.

(d) A major bottleneck for structure-based drug design of AR agonists or antagonists is the lack of three-dimensional (3D) structural information about G-protein-coupled receptors through standard structure determination techniques X-ray and nuclear magnetic resonance studies because of the difficulties in receptor purification and their insolubility in environments lacking phospholipids.

The problem of side-effects exposed in (a) obviously demands for selective and specific agonists to overcome it. The simultaneous study of the agonist efficacy, the binding affinity to the target AR and the binding affinity of the rest of subtypes could offer practical clues in this regard, motivating future researches in this area. In the present work, the last three problems [(b), (c), and (d)] will be tackled.

From (b) and (c), it is clear that both the binding affinity and the agonist efficacy should be simultaneously studied to develop selective A_3AR agonists. Even more, the study of the combination of both properties could be very informative and useful. However, from (d) we are aware of the little feasibility of a structure-based approach. Therefore, in cases where the receptor structure is unknown, a ligand-based approach, based only on an extensive study of structure-activity relationships (SAR), could be an informative alternative. In particular, the quantitative structure-activity relationship (QSAR) paradigm has long been of interest in the drug-design process (13, 14). Recently, an excellent review on QSAR tools to find new A_3AR agonists using 2D and 3D molecular descriptors (MDs) has been published (15).

When a medicinal chemist faces the problem of using QSAR prediction models (PM) to aid the search for new drug candidates, the desired goal is to obtain an interpretable and predictive PM. However, the fact is that the 'dominant Boolean operator' in this situation is not precisely 'AND', and more often what is desired, results to be 'OR' the dominant operator. So the interpretability of a PM is a trade-off with predictive accuracy. For example, linear regression models can be interpreted in a detailed fashion, but, generally, have lower accuracy, especially for biological activities. On the other hand, one can achieve high accuracy using a neural network model, but extracting the encoded SAR can be very difficult. In the same way, MDs with a direct physicochemical or structural meaning such as physicochemical properties or constitutional descriptors can be easily translated into structural modifications enhancing the biological profile of a molecule, whereas highly informative MDs such as the 3D ones tend to be more abstract and do not allow one easily to understand the substructures that are important for activity (16).

Thus, a PM provides to the researcher with two aspects: a set of predicted values, and information regarding the SAR(s) that are present in the dataset. Unfortunately, these two parameters are not usually provided jointly. As a consequence, it is necessary to establish priorities in an investigation weighting the importance of predictivity and interpretability, prioritize that what is determinant for the problem, and select the MDs and the modeling strategy accordingly.

At the same time, improving the profile of a molecule for the drug discovery process requires the simultaneous optimization of numerous, often competing objectives. Classic QSAR approaches usually ignore the multi-objective nature of the problem focusing on the evaluation of each single property as they became available during the drug discovery process (17). So an approach offering a simultaneous study of several biological properties determinants for a specific therapeutic activity is considered a very attractive option in computational medicinal chemistry. In this sense, desirability functions (DF) are well-known multi-criteria decision-making methods (18,19). This approach has been extensively employed in several fields (20–31). However, despite of perfectly fit with the drug development problem, reports of computational medicinal chemistry applications are at present very scarce (32,33).

Recently, a three-dimensional QSAR study (3D-QSAR) on the A_3AR agonists binding affinity and relative efficacy profiles including oxo- and thioadenosine analogs exposed the outlier nature of thioadenosine derivatives (12). In a training set of 91 compounds, five of eight outliers were 4′-thioadenosine analogs, indicating the possibility of a subtle difference in the binding mode and activation mechanisms of 4′-thioadenosine analogs in comparison with the oxo analogs. The nature of the substituents on the N^6 position of the adenine ring was found to play a significant role in the binding affinity and relative efficacy of the compounds. These interesting findings make N^6-substituted-4′-thioadenosine analogs an attractive goal in A_3AR agonists research.

Considering the medicinal and computational chemistry problems above exposed, we propose in this work the use of the desirability theory as a tool to extract useful information on the desired trade-offs between binding and relative efficacy of N^6-substituted-4′-thioadenosines A_3AR agonists. Additionally, desirability and belief theories are combined to integrate a ligand-based virtual screening (LBVS) protocol allowing the fusion of results from independent approaches to access the reliability of concurrent predictions.

Materials and Methods

Data set and computational methods

The multiple linear regression (MLR) PMs developed were based on the binding affinities (Ki_{A3}) and relative maximal efficacy (RE_{A3}) in the activation of the A_3AR reported by Jeong *et al.* (34) for a library of thirty-two N^6-substituted-4′-thioadenosines A_3AR agonists. The chemical structures and property values are depicted in the Supporting Information related to this work.

The structures of all compounds were first drawn with the aid of ChemDraw Ultra 9.0[a], and reasonable starting geometries by

resorting to the MM2 molecular mechanics force field were obtained (35,36). Molecular structures were then fully optimized with the PM3 semi-empirical Hamiltonian (37), implemented in the MOPAC 6.0 program (38). Subsequently, the optimized structures were brought into the DRAGON software package[b] for computing a total of 1664 MDs. Descriptors having constant or near constant values were excluded. Thus, from the initial set 1320 MDs remained for further variable selection and construction of the PMs (focused on predictability) involved on LBVS approach.

On the other hand, for the overall desirability PM (focused in interpretability) involved on the desirability-based interpretation approach was computed only 351 MDs (48 constitutional, 154 functional groups count, 120 atom-centered fragments and 29 molecular properties). These four families of MDs were chosen because their simple nature offers an easily structural or physicochemical interpretation of the resultant PM. To reduce noisy information that could lead to chance correlations, descriptors having constant or near constant values as well as highly pair-correlated ($|R| > 0.9$) were excluded. Consequently, from the initial set, only 32 MDs remained for further variable selection. The set of four variables finally included in the model is depicted in Table 1.

An optimization technique – the Genetic Algorithm (GA) – was applied for variable selection (39–41) by using the MOBYDIGS 1.1 software package[c]. The GA selection parameters setup was: population size = 100, maximum allowed variables in the model = 7, reproduction/mutation trade-off = 0.5 and selection bias = 50%. The determination coefficient of the leave-one-out cross-validation (Q^2_{LOO}) was employed as fitness function.

The predictive ability of the PMs was evaluated by means of internal cross-validation (CV). Specifically, the leave-one-out (LOO) technique (42) is already implicit on the GA feature selection process, being characterized by the Q^2_{LOO} and s_{LOO} statistics in eqns 12–14. Additionally, to ensure the predictive ability, the resultant PM was subjected to a bootstrap validation procedure (43) determined by 8000 resubstitutions (characterized by Q^2_{Boost} and s_{Boost} statistics in eqns 12–14). A *Y*-scrambling procedure (44) (based on 500 random permutations of the *Y*-response vector) implemented on MOBYDIGS[c] was also applied to check whether the correlations established by the respective PMs were because of chance correlations or not. See $a(R^2)$ and $a(Q^2)$ statistics in eqns 12–14, where unstable models because of chance correlations are characterized by high values and vice versa. In this way, the quality and predictive ability of the PMs can be assessed.

Table 1: Molecular descriptors (MDs) included on the overall desirability prediction model, identified through the Genetic Algorithm selection process

MDs	Definition	Family
ARR	Aromatic ratio	Constitutional descriptors
nCIR	Number of circuits	Constitutional descriptors
nCs	Number of total secondary sp3 carbon atoms	Functional groups count
ALOGP2	Squared Ghose-Crippen octanol–water partition coefficient (logP^2)	Molecular properties

We have also checked the validity of the preadopted parametric assumptions, another important aspect in the application of linear multivariate statistical-based approaches. These include the linearity of the modeled property, normal distribution of residuals as well as the homoscedasticity and non-multicollinearity of the independent variables included in the MLR model (45,46).

Finally, the applicability domain of the final PMs was identified by a leverage plot, that is to say, a plot of the standardized residuals *vs* leverages for each training compound (42,47).

Scaling properties with desirability functions

The properties Y_i were scaled to their respective desirability (d_i) values by means of the Derringer DF (19). Desirability functions are well-known multi-criteria decision-making methods, based on the definition of a DF for each property to transform their values to the same scale. Each attribute (Ki_{A3} and RE_{A3}) is independently transformed into a desirability value ($d_{(KiA3)}$ and $d_{(REA3)}$) by an arbitrary function. The original value is range scaled between 0 and 1 by:

$$d_i = \frac{\hat{Y}_i - L_i}{U_i - L_i} \qquad 0 \le d_i \le 1 \tag{1}$$

where L_i and U_i are the selected minimum and maximum values, respectively.

In this work, two specific DF (one for each property) were used.

If a property is to be maximized, its individual DF is defined as:

$$d_i = \begin{cases} 0 & \text{if } Y_i \le L_i \\ \left[\dfrac{Y_i - L_i}{T_i - L_i}\right]^s & \text{if } L_i < Y_i < T_i \\ 1 & \text{if } Y_i \ge T_i = U_i \end{cases} \tag{2}$$

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

On the other hand, if one wants to minimize a property, one might use:

$$d_i = \begin{cases} 1 & \text{if } Y_i \le T_i = L_i \\ \left[\dfrac{Y_i - U_i}{T_i - U_i}\right]^s & \text{if } U_i < Y_i < T_i \\ 0 & \text{if } Y_i \ge U_i \end{cases} \tag{3}$$

Here, T_i denotes a small enough value for the property, which can be L_i.

Specifically, RE_{A3} ought to be maximized (eqn 2) in such a way that the compound with the highest/lowest value should be the most desirable/undesirable ($d_i = 1/d_i = 0$). Specifically, L_i was set to 0%, and the upper value U_i, made equal to the target value T_i, was set to 114%. In contrast, to maximize the binding affinity to the human

A_3AR, the Ki_{A3} values most be minimized (eqn 3) where $L_i = T_i = 0.8$ nM and $U_i = 1650$ nM, coinciding with the lower and higher values of Ki_{A3} in the data set, respectively.

Anyhow, if a response is of the *target* best kind, then its individual DF is defined as:

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

The exponents s and t in eqns (2–4) determine how important is to hit the target value T_i. For $s = t = 1$, the DF increases linearly towards T_i. Large values for s and t should be selected if it is very desirable that the value of $\hat{Y}_i$ 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 $\hat{Y}_i$ above L_i, and below U_i are acceptable or if having values of $\hat{Y}_i$ considerably above L_i are not of critical importance (19).

The individual desirabilities are then combined using the geometric mean, which gives the overall desirability D_i:

$$D_i = (d_1 \times d_2 \times \ldots \times d_k)^{\frac{1}{k}} \quad (5)$$

with k denoting the number of properties.

This single value of D_i gives the overall assessment of the desirability of the combined property levels. Clearly, the range of D_i will fall in the interval [0, l] and will increase as the balance of the properties becomes more favorable.

Ranking quality

To measure the quality of the ranking obtained we employ a quantitative measure also based on the application of DF.

We will use a simple notation to represent ordering throughout this article. 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. We assume the examples are ordered incrementally from left to right. Then, the *true-order list* is O_T = 1(lowest), 2, 3, …, n (highest). 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_n, where a_i is the actual ordering position of the case that is ranked ith in O_R (see Table 2).

The ranking validation includes the following steps:

Table 2: An 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

1. Order the cases in the library according to D_i in a decreasing fashion and label each case as described earlier (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 earlier (a_1, a_2, …, a_n). This ordering corresponds to the order generated by the ranking algorithm R (O_R).

4. Normalize [through eqn (3)] the values (labels) assigned to each case on steps 1 to 3 where $L_i = T_i = 1$ y U_i = number of cases included in the library (n). In this way, we obtain 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 (${}^{OT-OR}\delta_i$):

$${}^{OT-OR}\delta_i = \left|{}^{OT}d_i - {}^{OR}d_i\right| \quad (6)$$

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

$${}^{OT-OW}\delta_i = \left|{}^{OT}d_i - {}^{OW}d_i\right| \quad (7)$$

The ideal difference is 0 for all the cases and corresponds to a perfect ranking.

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:

$$\Psi = \left|\frac{\sum_{i=1}^{n} {}^{OT-OR}\delta_i}{n}\right| \quad (8)$$

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. Like this, the closer to 0 is Ψ for a certain ranking the higher will be the quality of this ranking. In contrast, values of Ψ near to 0.5 indicate a low-ranking quality. Because the value of Ψ associated to the worst ranking is dependent of the size of the library to be ranked, this value is not exactly, but 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 the previous questions, a correction factor (F) is applied to Ψ:

$$F = \frac{2}{\Psi^{OW}} \quad (9)$$

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^*):

$$\Psi^* = \left| \frac{\sum_{i=1}^{n} {}^{OT-OR}\delta_i}{n} \right| \cdot F = \left| \frac{\sum_{i=1}^{n} {}^{OT-OR}\delta_i}{n} \right| \cdot \frac{2}{\Psi^{WR}} \qquad 0 \le \Psi \le 1 \quad (10)$$

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

$$R_{\%} = (1 - \Psi^*) \cdot 100 \qquad 0 \le R_{\%} \le 100 \qquad (11)$$

Results and Discussion

Prediction models

Once desirability scaled both Ki_{A3} and RE_{A3} responses for each compound, the corresponding overall desirability ($D_{KiA3\text{-}REA3}$) values were derived. To identify the factors governing the trade-offs between binding affinity and efficacy of this family of A_3AR agonists, the combined response $D_{KiA3\text{-}REA3}$ was mapped as a function of four simple 1D MDs with a direct structural and/or physiochemical explanation. The resulting best-fit model together with the statistical regression parameters is given below:

$$\begin{aligned} D_{KiA3-REA3} = {} & 1.557(\pm 0.292) - 0.107(\pm 0.013) \times ALOGP2 \\ & + 0.203(\pm 0.033) \times nClR - 2.783(\pm 0.595) \\ & \times ARR - 0.092(\pm 0.027) \times nCs \end{aligned} \qquad (12)$$

N = 32 R^2 = 0.781 R^2_{Adj} = 0.749 F = 24.13 s = 0.127

Q^2_{LOO} = 0.566 s_{LOO} = 0.138 Q^2_{Boost} = 0.539 s_{Boost} = 0.179
$a(R^2)$ = 0.0063 $a(Q^2)$ = −0.0039

The statistical significance and predictive ability exhibited by the model show evidence of their suitability for subsequent analyses.

No violations of the preadopted parametric assumptions were found for eqn (12).

At the same time, two QSAR PMs (for Ki_{A3} and RE_{A3}) focused on their predictive ability (identified further as prediction approach A_2) were derived to use both in combination with the previously described overall desirability PM (eqn (12), identified further as prediction approach A_1) in a LBVS strategy based on the combination of their concurrent predictions through belief theory.

The resulting best-fit models together with the statistical regression parameters are given in eqns (13 and 14):

$$\begin{aligned} Ki_{A3} = {} & -8857.67(\pm 331.482) + 10.36(\pm 1.019) \cdot D/Dr03 \\ & + 502.99(\pm 99.263) \cdot GATS3m + 5217.43(\pm 188.103) \cdot BELe3 \\ & - 453.64(\pm 45.869) \cdot Mor13u + 1110.88(\pm 57.144) \cdot Mor09v \\ & - 1258.23(\pm 101.691) \cdot Mor23v + 26703.72(\pm 3542.089) \cdot R7u + \end{aligned} \qquad (13)$$

N = 32 R^2 = 0.985 R^2_{Adj} = 0.981 F = 230.82 s = 48.796

Q^2_{LOO} = 0.977 s_{LOO} = 56.345 Q^2_{Boost} = 0.957 s_{Boost} = 61.246
$a(R^2)$ = 0.0017 $a(Q^2)$ = −0.0052

$$\begin{aligned} RE_{A3} = {} & 2559(\pm 413.56) - 3307(\pm 373.0.4) \cdot PW2 \\ & - 0.44(\pm 0.038) \cdot D/Dr06 - 143.68(\pm 28.85) \cdot ATS5v \\ & + 344.25(\pm 25.72) \cdot EEig10d + 114.72(\pm 10.54) \cdot VEA1 \\ & + 89.91(\pm 20.18) \cdot H8p - 15.68(\pm 2.32) \cdot ALOGP \end{aligned} \qquad (14)$$

N = 32 R^2 = 0.966 R^2_{Adj} = 0.956 F = 96.79 s = 5.515

Q^2_{LOO} = 0.942 s_{LOO} = 6.369 Q^2_{Boost} = 0.921 s_{Boost} = 7.182
$a(R^2)$ = 0.0017 $a(Q^2)$ = −0.0055

According to their statistics, the models are good in terms of their statistical significance and predictive ability. In opposition to eqn (12), eqns (13 and 14) were derived from a pool of variables significantly higher than the number of cases used for training. As a consequence, the risk to find chance correlations in such a vast variable space is always high. So checking the occurrence of this event is of vital importance in this case. As can be deduced from the significantly low values of $a(R^2)$ and $a(Q^2)$ obtained in the respective Y-scrambling experiments, there is no reason to ascribe to chance correlations the statistical significance and predictive ability exhibited by each PM.

With the exception of the non-multicollinearity of the independent variables included in the MLR model developed for RE_{A3}; no violations of the remaining MLR parametrical assumptions were found (48). As above-mentioned, multi-collinearity affects the common interpretation of a regression equation. However, the predictive ability of the PM is not affected in this situation (46).

See Supporting Information for details of the inspection of the parametrical assumptions as well as the establishment of the applicability domain of eqns (12–14).

Consequently, according to the statistical parameters exhibited, the goodness of fit of the PMs involved on both prediction approaches A_1 and A_2 can be considered as statistically significant. At the same time, considering their satisfactory predictive ability and the validity of the preadopted parametrical assumptions, the resultant predictions can be regarded as reliable in the domain of the N^6-substituted-4′-thioadenosines A_3AR agonists used for training and structurally coded as a linear function of the respective subsets of MDs. Therefore, all the PMs developed can be employed in a LBVS scheme with an adequate degree of reliability.

Desirability-based prediction model interpretation and theoretical design of N^6-substituted-4′-thioadenosine A_3AR agonist candidates

Based on the satisfactory accuracy, statistical significance and predictive ability of the overall desirability PM (eqn (12)) we can proceed, with an adequate level of confidence to the simultaneous analysis of the factors governing the balance between the binding affinity and relative efficacy profiles of A_3AR agonists.

Although the main variation of the subset of compounds employed is over the N^6 position of the adenine ring, the MDs employed in mapping $D_{KiA3\text{-}REA3}$ are global and not fragment based. So any inference made have to be only based on the influence of N^6 substituents over the global molecular system.

First, the information encoded in the MDs included on the model was analyzed. According to the model regression parameters, the most influencing MD is the aromatic ratio (*ARR*), followed by the Ghose-Crippen octanol–water partition coefficient (*ALOGP2*), the number of circuits (*nCIR*) and the number of total secondary sp3 carbon atoms (*nCs*). All MDs were inversely related with the overall desirability $D_{KiA3\text{-}REA3}$ of N^6-substituted-4′-thioadenosine A3AR agonists, except *nCIR*.

Specifically, *ARR* is the fraction of aromatic atoms in the hydrogen suppressed molecule graph and encodes the degree of aromaticity of the molecule. According to the model parameters, N^6 substitutions increasing the aromaticity of the molecule do not favor $D_{KiA3\text{-}REA3}$.

ALOGP2 is simply the square of the Ghose-Crippen octanol–water coefficient (*ALOGP*), which is a group contribution model for the octanol–water partition coefficient. Because these MDs encode the hydrophobic/hydrophilic character of the molecule, $D_{KiA3\text{-}REA3}$ could be favored by the presence of N^6 substituents contributing to reduce the hydrophobicity of the molecule.

The *nCIR* is a complexity descriptor, which is related to the molecular flexibility. Because *nCIR* serve as a measure of rigidity with higher numbers of circuits corresponding to reduced flexibility; cyclic and rigid or conformationally restricted N^6 substituents could increase the overall desirability of the molecular system.

Finally, the presence of secondary sp^3 carbon atoms in the molecule appears to be detrimental for $D_{KiA3\text{-}REA3}$.

According to the model, a molecule with a low aromaticity degree, without secondary sp^3 carbon atoms, and containing cyclic and rigid N^6 substituents, which contributes to reduce the hydrophobicity of the system could favor the balance of the binding affinity and relative efficacy profiles of N^6-substituted-4′-thioadenosine A_3AR agonists.

To note that these conclusions, although derived from a simple 1D model, are very similar to that obtained by 3D-CoMFA/CoMSIA approaches (12). Kim and Jacobson have concluded that a bulky group, conformationally restricted, at the N^6 position of the adenine ring will increases the A_3AR binding affinity, and that a small bulky group, at this position, might be crucial for A_3AR activation. Note the accordance of data obtained in the previous and present work: a 'conformationally restricted bulky group' is suggested by Kim and Jacobson and herein a 'cyclic and rigid substituents' on the N^6 position.

To note that although *nCIR* is not the MD more significantly related with $D_{KiA3\text{-}REA3}$, it is very informative for the property. From *nCIR,* we can infer that the bulkiness of the N_6 substituent suggested in (12) can be characterized by a cyclic rather than an alkyl substituent.

Although useful, this information is found to be incomplete because it is well known that steric factors are determinant for the design of A_3AR agonists, especially for binding affinity (12). Consequently, it is found to be important to determine the optimal size of the conformationally restricted cyclic N_6 substituent. Unfortunately, the simple inspection of the regression parameters of the PM does not offer this information. In consequence, a property/desirability profiling was carried out to identify the levels of the MDs included in the PM that simultaneously generate the most desirable combination of binding affinity and relative efficacy.

As the main goal of this analysis is to extract information on the factors governing $D_{KiA3\text{-}REA3}$ rather than optimize it, the behavior of $D_{KiA3\text{-}REA3}$ was profiled at the mean values of the four MDs rather than looking for their optimal values (see first row in Figure 1). Accordingly, it was possible to find the levels of the MDs simultaneously producing the best possible $D_{KiA3\text{-}REA3}$ in the training set employed. As can be noted in Figure 1 (second row), a A_3AR agonist candidate should exhibit a value of $D_{KiA3\text{-}REA3}$ near to 0.9 at levels of *ARR*, *nCs*, *ALOGP2,* and *nCIR* around 0.4, 2, 0, and 6; respectively.

The analysis reveal that the most favorable balance of binding affinity and agonist efficacy: the *ARR* should be not just low but near to 0.4; *ALOGP2* should be as low as possible; the number of secondary sp3 carbon atoms should be kept around two; and *nCIR* should be not just high but close to six.

Because the thioadenosine nucleus already contain three secondary sp3 carbon atoms, at least on the applicability domain of the present model, the minimum number of such atoms should be kept at three. So this type of carbons must be excluded in the substituents located at N^6 position.

At the same time, considering that the *nCIR* value of the thioadenosine nucleus is four, one can deduce that the ideal *nCIR* value of the N^6 substituent should be two. This information can be structurally translated into bicyclic N^6 type of substituents.

The inclusion in the PM of *nCIR*, instead of the number of rings in the chemical graph (*nCIC*) is also significant. Although the structural information of this pair of MDs is very similar (the number of cyclic structures in a chemical graph) their graph-theoretical information is quite different. While *nCIC* encodes the number of rings, *nCIR* includes both rings and circuits (a circuit is a larger loop around two or more rings). As an example, naphthalene contains 3 circuits and 2 rings. This is illustrated in Figure 2.

So additional information can be inferred: the bicyclic N^6 substituent should not be fused. This assumption could be related to the binding interaction of this type of fragments with the A_3AR. In fact, the presence of a certain degree of rotational freedom between the two rings of the fragment could favor its docking into the receptor cavity.

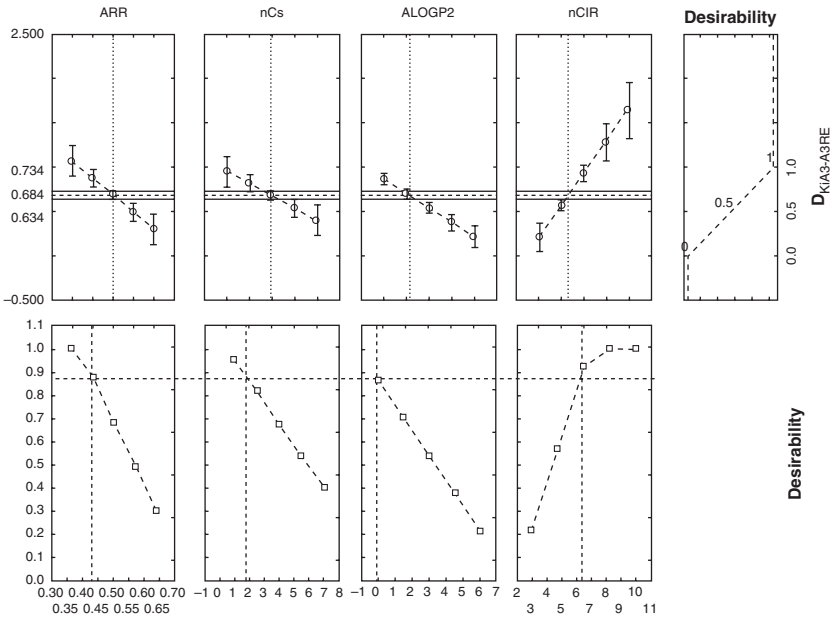


Figure 1: Property/desirability profiling of the levels of the molecular descriptors that simultaneously produce the most desirable combination of binding affinity and relative efficacy of N^6-substituted-4′-thioadenosine A_3AR agonists.

This result matches with previous experimental findings on the SAR of this family of thioadenosine derivatives (34). The SAR obtained for this family suggests that compounds with bulky N^6 substituents lost their binding to the A_3AR. Paradoxically, among compounds showing high binding affinity at the human A_3AR, two compounds substituted with a N^6-(*trans*-2-phenylcyclopropyl)amino group were found to be full agonists at the human A_3AR. In addition, it was found that compounds with α-naphthylmethyl N^6 substituents lost their binding to the A_3AR (34), which reinforce the present proposal.

From the study it was also concluded that bulky N^6 substituents only affects the binding affinity; however bulky (bicyclic) substituents such as a *trans*-2-phenylcyclopropyl group could be beneficial for agonist efficacy without lost their binding affinity. Although that experimental study do not deal with the simultaneous analysis of both properties, their experimental findings properly match with our theoretical results.

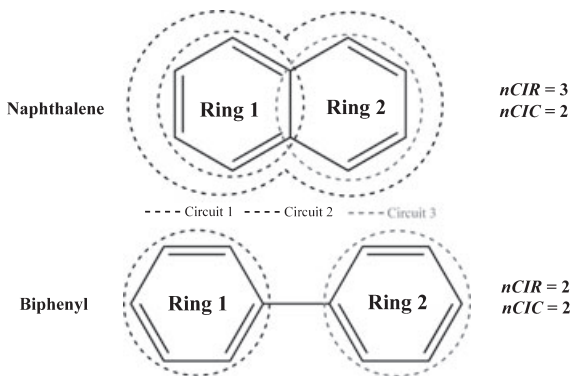


Figure 2: Graphical illustration of the definition of *nCIC* and *nCIR* for two chemical graphs.

Until now, it has been exposed the importance of bicyclic and rigid N^6 substituents contributing to reduce the hydrophobicity of the system to obtain an adequate balance between binding affinity and relative efficacy profiles of N^6-substituted-4′-thioadenosine A_3AR agonists.

At first sight, this information is pretty focused and we could expect that the task of finding promising candidates is almost performed. However, if we consider the number of attainable N^6 substituents of this type, generated from a tiny portion of the possible chemical space indicated by this information we can extrapolate the huge number of possible candidates (Table 3). To mention that this analysis has been only performed taking into account unsaturated rings and the valence of the atoms. The number of options can vary, rising or go down if we consider double bounds or chemical feasibility. Anyway, although focused, the 'haystack' is vast. So it is determinant a focused screening strategy to efficiently find some 'needle' on it.

Therefore, the previous information is employed for the theoretical design of new N^6-substituted-4′-thioadenosine analogs with adequate balances between binding affinity and agonist efficacy. Because *ARR* and *ALOGP2* cannot be easily manipulated by structural modifications, the design efforts will be mainly focused on *nCs* and *nCIR*. Thus, a combinatorial library focused on the generation of N^6-substituted-4′-thioadenosine candidates was assembled with $nCs \approx 3$ and $nCIR \approx 6$. This approach was performed with the aid of the SmiLib software (48), for the rapid assembly of combinatorial Libraries in SMILES notation. The library was directed to produce candidates with conformationally restricted bicyclic N^6 substituents while keeping at minimum the presence of secondary sp^3 carbon atoms using the 4′-thioadenosine nucleus as scaffold and a set of 25 cyclic or heterocyclic structures as linkers and building blocks. The working combinatorial scheme is shown in Table 4.

Table 3: Fraction of the chemical space determined by the N^6 substituents conformed by the possible combinations of two not fused rings linked by a single bound

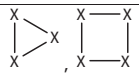
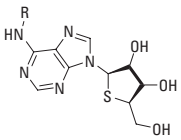
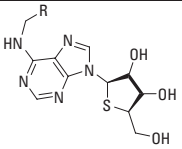
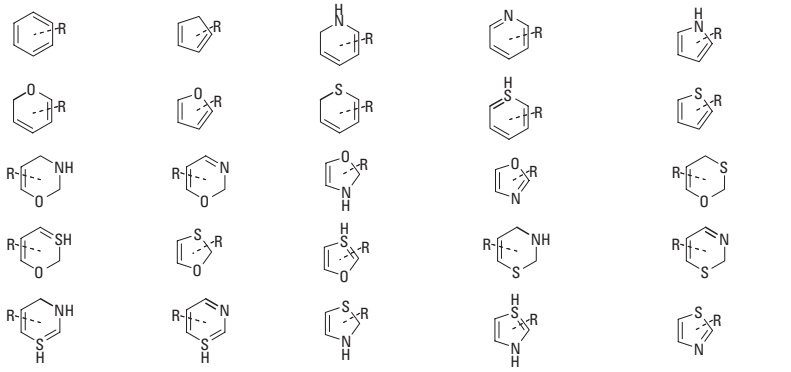
Rings (R) = 2	R-X = 35	R-X-S = 107
		
Atom Type (X) = 4		
C, O, N, S		
Substitution Places (S): Up to 4 per Ring		
(*S = No. of Ring Members if X ≠ O, Otherwise S = No. of Ring Members – No. of O atoms*)		
N^6-R-R` = 2 not fused Rings linked by a single bound = 11449		

Table 4: Scaffolds, linkers, and building blocks employed to assemble the combinatorial library

Scaffolds	
	
Linkers/Building Blocks	
	

This combinatorial strategy produced a set of more than 9000 candidates, which according to previous results can be employed in a subsequent virtual screening campaign using as ranking criterion the predicted value of $D_{KiA3\text{-}REA3}$ of each candidate. As mentioned before, only candidates included on the applicability domain of the overall desirability PM (3395 candidate molecules) should be submitted to the ranking process. Figure 3 shows the plot of the predicted $D_{KiA3\text{-}REA3}$ values of the 9782 candidate molecules versus their respective leverage values. As can be noted, predictions range from values of −0.31 to 1.70; however, candidates included on the PM applicability domain are restricted to predicted values of $D_{KiA3\text{-}REA3}$ between 0.22 and 1.44. As a result, it is possible to propose for biological screening a reduced set of candidates with a promissory balance between A_3AR binding affinity and agonist efficacy. The values of the MDs included on the overall desirability PM as well as the predicted value of $D_{KiA3\text{-}REA3}$ for a fragment of the ranked combinatorial library are shown in Table 5.

Library ranking based on the combination of desirability and belief theories

Although the idea of desirability-transforming and combining a number of related properties is in accordance with the concept of pharmaceutical profile (32,33), the usefulness of a parallel approach allowing obtaining a feedback on the reliability of the properties predicted as a unique D_i value is also desirable.

If two or more property values Y_i (previously scaled to the respective d_i values with proper DF) of a compound are combined into a unique D_i value, to map it as a MLR function of n MDs X_i (denoted as approach A_1), it is rational to expect that the resultant predicted D_i value should be similar to the inverse approach. The inverse approach consist in the independent mapping of the k properties Y_i as a MLR function of n MDs X_i, the subsequent desirability-scaling of each predicted Y_i value and the final combination of the corresponding d_i values into a unique predicted D_i value (denoted as approach A_2).

$$\begin{aligned} &Y_i \rightarrow d_i \rightarrow D_i = f(X_i) \rightarrow \Pr ed.D_i = A_1 \approx A_2 \\ &= \Pr ed.D_i \leftarrow \Pr ed.d_i \leftarrow \Pr ed.Y_i \leftarrow Y_i = f(X_i) \end{aligned} \tag{15}$$

Assuming true the previous analysis, one must anticipate that the higher is the degree of similarity between the predicted D_i values of both approaches, the higher should be their reliability, and vice versa. Clearly, the results will depend on the goodness of fit and prediction of the set of PMs involved. In addition, the

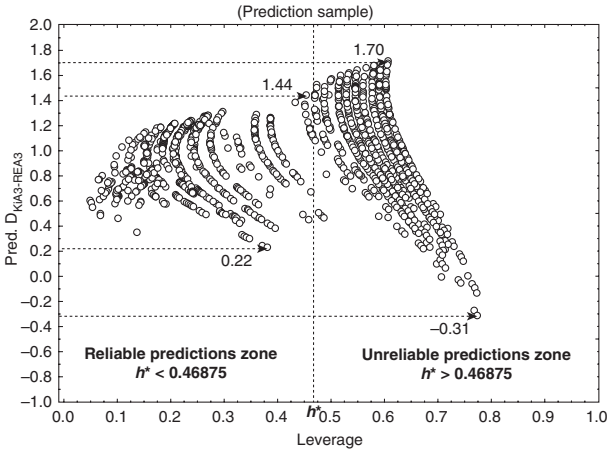


Figure 3: Predicted $D_{KiA3-REA3}$ values of the candidate molecules included on the combinatorial library plotted vs. their respective leverage values.

degree of uncertainty of PMs with different sets of MDs will be diverse.

So it is required a framework allowing the fusion of results from different approaches to access the reliability of predictions from several approaches with different degrees of uncertainty. In the present work, we select Dempster–Shafer Theory (DST) (49–51) (also known as belief theory) to achieve that goal. DST is a mathematical theory of evidence that has been developed to combine separate pieces of information that can arise from different sources (52). Dempster–Shafer Theory is based on two ideas: the idea of obtaining degrees of belief for one question from subjective probabilities for a related question, and Dempster's rule for combining such degrees of belief when they are based on independent items of evidence (52).

Table 5: Fractions of the combinatorial library ranked according to the predicted values of $D_{KiA3-REA3}$

Rank	Comb. Lib. ID*	*ARR*	*nCIR*	*nCs*	*ALOGP2*	Pred. $D_{KiA3-REA3}$
1	1.36_2	0.294	6	5	0.532	1.439
2	1.36_3	0.294	6	5	0.532	1.439
3	2.4_54	0.294	6	5	0.567	1.436
4	2.5_3	0.294	6	5	0.633	1.429
5	2.5_2	0.294	6	5	0.633	1.429
2221	1.32_55	0.455	6	3	2.161	1.000
2222	1.54_17	0.455	6	3	2.163	1.000
2223	1.17_86	0.441	6	3	2.527	1.000
2224	1.55_11	0.471	6	3	1.752	0.999
2225	1.35_40	0.441	6	3	2.541	0.998
2914	2.52_108	0.441	6	3	4.388	0.800
2915	1.34_87	0.441	6	3	4.402	0.799
2916	2.10_106	0.457	6	3	3.992	0.798
2917	1.58_90	0.357	5	3	4.7	0.798
2918	1.38_109	0.441	6	3	4.418	0.797
3343	2.35_106	0.441	6	3	7.185	0.500
3344	2.48_55	0.429	6	4	6.647	0.500
3345	2.54_53	0.441	6	3	7.198	0.499
3346	2.56_106	0.441	6	3	7.242	0.494
3347	2.48_109	0.429	6	4	6.702	0.494
3391	1.48_55	0.441	6	4	8.071	0.314
3392	1.48_109	0.441	6	4	8.132	0.307
3393	1.48_110	0.441	6	4	8.256	0.294
3394	1.48_52	0.441	6	4	8.74	0.242
3395	1.48_108	0.441	6	4	8.932	0.221

ARR, Aromatic ratio.
*Combinatorial Library identification: 1.36_2 = Scaffold1.Linker36_Building Block2.

The foundations of DST can be traced to the work of George Hooper, who published an article in the *Philosophical Transaction of the Royal Society* entitled 'A calculation of the credibility of human testimony' (50). In this article, Hooper formulated two rules relating the credibility of reports to the credibility of the reporters who make them (51).

These two rules are quite simple. The *rule for successive testimony* says that if a report has been relayed to us through a chain of n reporters, each having a degree of credibility p, then the credibility of the report is p^n. The *rule for concurrent testimony* says that if a report is concurrently attested to by n reporters, each with credibility p, then the credibility of the report is $1-(1-p)^n$; where $0 \leq p \leq 1$. Thus, the credibility of a report is weakened by transmission through a chain of reporters but strengthened by the concurrence of reporters (50,51).

If we make a simple analogy of this situation with the situation previously exposed regarding two parallel overall desirability PMs, each approached inversely, is possible to note that DST theory, specifically, the Hospers's rule for combining concurrent evidence (50,51), is fully applicable to our problem. There, it is only needed to replace 'report' with 'prediction' and 'reporter' with 'PM', and the previous paragraph will almost literally describe our problem.

Developing a *probability assignment* is the basic function in DST and is an expression of the level of confidence that can be ascribed to a particular measurement. However, in this work, we are interested on the desirability of a compound. Consequently, rather than a probability assignment for each compound, we will use the desirability values coming from both overall desirability PMs approaches (D_1 and D_2) to derive the final joint belief values (B_D):

$$B_D = 1 - (1 - D_1)(1 - D_2) \tag{16}$$

While desirability is not itself a probability, like probabilities their values also range from 0 to 1. Therefore, it can be used to derive the values of B_D for each compound. So in this way, it is possible to encode the reliability of the predicted desirability of a compound along with two inverse but complementary prediction approaches. Given this information, B_D can be used as ranking criterion in a virtual screening scheme, resulting particularly useful for LBVS.

A LBVS strategy based on B_D can be described in the sequence of steps detailed below:

1 Prediction Models setup.

Here, the predicted D_i values for each compound are derived from A_1 and A_2 as expressed in eqn (13).

2 Desirability assignment.

Because of limitations inherent to the MLR approach, the predicted desirability values not always will be included in the interval [0,1] and consequently is not possible to use it as is to derivate B_D. So in the case of the desirability values derived from the approach A_1, it is necessary to rescale using eqn (2) considering that D have to be maximized.

In the case of the approach A_2, the derivation of the respective D_i values is affected by the above-mentioned limitations of MLR, but the process is complicated by the wider range of the mapped Y_i properties. Consequently, d_i is scaled by using a two-tale (eqn (4)) using the same target T_i values employed in A_1 for each Y_i.

3 Derivation of Joint Belief B_D by the application of Hospers's Rule for Combining Concurrent Evidence.

4 B_D-based ranking.

The resultant ranking should render an ordered list, top ranking the most reliable compounds with the highest desirability values. The compounds with a higher chance to exhibit a desirable combination of the k properties modeled.

Subsequently, the B_D-based virtual screening (VS) strategy described earlier was applied to the already described training set to test their performance as ranking criterion. Considering the structural similarity between both (the combinatorial library assembled and our training set) is possible to use the latter to infer the reliability of the ranking attained for the combinatorial library. The predicted values of $D_{KiA3\text{-}REA3}$ (according to approach A_1) were also tested as ranking criterion to compare a VS strategy based on predictions coming from a single approach with a VS strategy based on the combination of concurrent predictions. The quality of the respective ranking obtained was compared according to Ψ^*, as described earlier.

Based on the analysis of our training set, the quality of the ranking attained using the predicted values of $D_{KiA3\text{-}REA3}$ is around 80%, which suggest an acceptable degree of confidence if the scheme is applied to our combinatorial library ($R_{\%}$ = 80.08%; Ψ^* = 0.1992). As can be noted in Figure 4, the use of B_D as ranking criterion ($R_{\%}$ = 82.81%; Ψ^* = 0.1719) slightly overcomes the performance of the predicted values of $D_{KiA3\text{-}REA3}$. Considering that B_D encodes in addition to the desirability of the compound, the reliability of such a prediction, it is clear their suitability at the moment to screen higher and/or structurally diverse libraries with a wider range of the mapped properties.

Conclusions

The development of a linear 1D PM of the A_3AR agonists overall desirability based on four simple MDs with a direct physicochemical or structural explanation, as well as the desirability analysis of this model, was described in this work. The results obtained provided significant clues on desired trade-offs between binding and relative efficacy of N^6-substituted-4′-thioadenosines A_3AR agonists.

The desirability-based PM interpretation strategy proposed here suggest a favorable effect over binding affinity and agonist efficacy of conformationally restricted, but not fused bicyclic N^6 substituents. The overall data provide guides to the rational design of new A_3AR agonist candidates by assembling a combinatorial library useful for the prioritization of candidates with a promissory balance between A_3AR binding affinity and agonist efficacy through a virtual screening campaign. The VS depicted protocol, based on the combined use of desirability and belief theories, exhibited a slightly superior performance compared with the single use of predicted overall desirabilities.

Finally, the combined use of desirability and belief theories in computational medicinal chemistry research was demonstrated to be a valid approach. The model was able to simultaneously consider

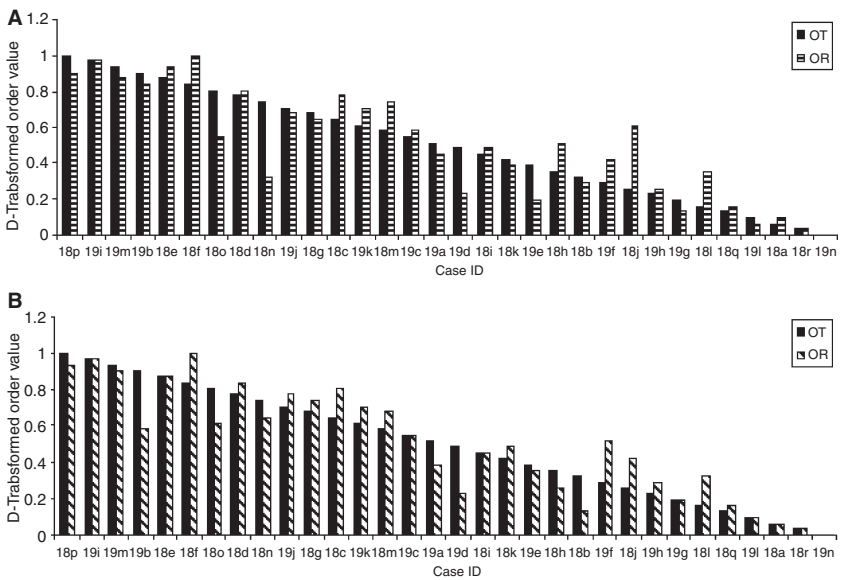


Figure 4: Ranking of the training set compounds based on B_D (top) and $D_{KiA3\text{-}REA3}$ (bottom), respectively.

several properties, in a simple an interpretable manner, and to execute a multi-target LBVS strategy.

Acknowledgment

The authors acknowledge the Portuguese *Fundação para a Ciência e a Tecnologia* (FCT) (MCM SFRH/BD/30698/2006, MNDSC SFRH/BSAB/930/2009 grants and the project PTDC/QUI/70359/2006) and Xunta de Galicia (PGIDIT07PXIB) for financial support.

References

1. Fredholm B.B., Jzerman A.P., Jacobson K.A., Klotz K.N., Linden J. (2001) International Union of Pharmacology. XXV. Nomenclature and classification of adenosine receptors. Pharmacol Rev;53:527–552.
2. Clarke B., Coupe M. (1989) Adenosine: Cellular mechanisms, pathophysiological roles and clinical applications. Int J Cardiol;23:1–10.
3. Chen Y., Corriden R., Inoue Y., Yip L., Hashiguchi N., Zinkernagel A., Nizet V, Insel P.A., Junger W.G. (2006) ATP release guides neutrophil chemotaxis via P2Y2 and A3 receptors. Science;314:1792–1795.
4. Jacobson K.A., Gao Z.G. (2006) Adenosine receptors as therapeutic targets. Nat Rev Drug Discov;5:247–264.
5. Grifantini M., Cristalli G., Franchetti P., Vittori S. (1991) Adenosine derivatives as agonists of adenosine receptors. Farmaco;46:161–169.
6. Jacobson K.A., Joshi B.V., Wang B., Klutz A., Kim Y., Ivanov A.A., Melman A., Gao Z-G. (2008) Modified nucleosides as selective modulators of adenosine receptors for therapeutic use. In: Herdewijn P., editor. Modified Nucleosides as Selective Modulators of Adenosine Receptors for Therapeutic Use. Weinheim: Wiley-VCH; p. 433.
7. Fishman P., Bar-Yehuda S. (2003) Pharmacology and therapeutic applications of A3 receptor subtype. Curr Top Med Chem;3:463–469.
8. Yan L., Burbiel J.C., Maass A., Muller C.E. (2003) Adenosine receptor agonists: from basic medicinal chemistry to clinical development. Expert Opin Emerg Drugs;8:537–576.
9. Leesar M.A., Stoddard M., Ahmed M., Broadbent J., Bolli R. (1997) Preconditioning of human myocardium with adenosine during coronary angioplasty. Circulation;95:2500–2507.
10. Conti J.B., Belardinelli L., Curtis A.B. (1995) Usefulness of adenosine in diagnosis of tachyarrhythmias. Am J Cardiol;75:952–955.
11. Madi L., Bar-Yehuda S., Barer F., Ardon E., Ochaion A., Fishman P. (2003) A3 adenosine receptor activation in melanoma cells: association between receptor fate and tumor growth inhibition. J Biol Chem;278:42121–42130.
12. Kim S.K., Jacobson K.A. (2007) Three-dimensional quantitative structure-activity relationship of nucleosides acting at the A3 adenosine receptor: analysis of binding and relative efficacy. J Chem Inf Model;47:1225–1233.
13. Brown N., Lewis R.A. (2006) Exploiting QSAR methods in lead optimization. Curr Opin Drug Discov Devel;9:419–424.
14. Hansch C. (1976) On the structure of medicinal chemistry. J Med Chem;19:1–6.
15. Gonzalez M.P., Teran C., Teijeira M., Helguera A.M. (2006) Quantitative structure activity relationships as useful tools for the design of new adenosine receptor ligands. 1. Agonist. Curr Med Chem;13:2253–2266.
16. Guha R. (2008) On the interpretation and interpretability of quantitative structure-activity relationship models. J Comput Aided Mol Des;22:857–871.
17. Nicolaou A.C., Brown N., Pattichis C.S. (2007) Molecular optimization using computational multi-objective methods. Curr Opin Drug Discov Devel;10:316–324.
18. Harrington E.C. (1965) The desirability function. Ind Qual Control;21:494–498.
19. Derringer G., Suich R. (1980) Simultaneous optimization of several response variables. J Qual Technol;12:214–219.
20. Outinen K., Haario H., Vuorela P., Nyman M., Ukkonen E., Vuorela H. (1998) Optimization of selectivity in high-performance liquid chromatography using desirability functions and mixture designs according to PRISMA. Eur J Pharm Sci;6:197–205.
21. Garcia-Gonzalez D.L., Aparicio R. (2002) Detection of vinegary defect in virgin olive oils by metal oxide sensors. J Agric Food Chem;50:1809–1814.
22. Shih M., Gennings C., Chinchilli V.M., Carter W.H. Jr (2003) Titrating and evaluating multi-drug regimens within subjects. Stat Med;22:2257–2279.
23. Kording K.P., Fukunaga I., Howard I.S., Ingram J.N., Wolpert D.M. (2004) A neuroeconomics approach to inferring utility functions in sensorimotor control. PLoS Biol;2:e330.
24. Safa F., Hadjmohammadi M.R. (2005) Simultaneous optimization of the resolution and analysis time in micellar liquid chromatography of phenyl thiohydantoin amino acids using Derringer's desirability function. J Chromatogr A;1078:42–50.
25. Pavan M., Todeschini R., Orlandi M. (2006) Data mining by total ranking methods: a case study on optimisation of the ``pulp and bleaching'' process in the paper industry. Ann Chim;96:13–27.
26. Coffey T., Gennings C., Moser V.C. (2007) The simultaneous analysis of discrete and continuous outcomes in a dose-response study: using desirability functions. Regul Toxicol Pharmacol;48:51–58.
27. Rozet E., Wascotte V., Lecouturier N., Preat V., Dewe W., Boulanger B., Hubert P. (2007) Improvement of the decision efficiency of the accuracy profile by means of a desirability function for analytical methods validation. Application to a diacetyl-monoxime colorimetric assay used for the determination of urea in transdermal iontophoretic extracts. Anal Chim Acta;591:239–247.
28. Wong W.K., Furst D.E., Clements P.J., Streisand J.B. (2007) Assessing disease progression using a composite endpoint. Stat Methods Med Res;16:31–49.
29. Cojocaru C., Khayet M., Zakrzewska-Trznadel G., Jaworska A. (2008) Modeling and multi-response optimization of pervaporation of organic aqueous solutions using desirability function approach. J Hazard Mater;167:52–63.
30. Fajar N.M., Carro A.M., Lorenzo R.A., Fernandez F., Cela R. (2008) Optimization of microwave-assisted extraction with saponification (MAES) for the determination of polybrominated flame retardants in aquaculture samples. Food Addit Contam Part A Chem Anal Control Expo Risk Assess;25:1015–1023.

31. Jancic-Stojanovic B., Malenovic A., Ivanovic D., Rakic T., Medenica M. (2009) Chemometrical evaluation of ropinirole and its impurity's chromatographic behavior. J Chromatogr A;1216:1263–1269.
32. Cruz-Monteagudo M., Borges F., Cordeiro M.N. (2008) Desirability-based multiobjective optimization for global QSAR studies: application to the design of novel NSAIDs with improved analgesic, antiinflammatory, and ulcerogenic profiles. J Comput Chem;29:2445–2459.
33. Cruz-Monteagudo M., Borges F., Cordeiro M.N.D.S., Cagide Fajin J.L., Morell C., Molina Ruiz R., Cañizares-Carmenate Y., Dominguez E.R. (2008) Desirability-Based Methods of Multiobjective Optimization and Ranking for Global QSAR Studies. Filtering Safe and Potent Drug Candidates from Combinatorial Libraries. J Comb Chem;10:897–913.
34. Jeong L.S., Lee H.W., Kim H.O., Jung J.Y., Gao Z.G., Duong H.T., Rao S., Jacobson K.A., Shin D.H., Lee J.A., Gunaga P., Lee S.K., Jin D.Z., Chon M.W., Moon H.R. (2006) Design, synthesis, and biological activity of N6-substituted-4′-thioadenosines at the human A3 adenosine receptor. Bioorg Med Chem;14:4718–4730.
35. Burkert U., Allinger N.L. (1982) Molecular Mechanics. Washington, D.C., USA: ACS.
36. Clark T. (1985) Computational Chemistry. NY, USA: Wiley.
37. Stewart J.J.P. (1989) Optimization of parameters for semiempirical methods I. Method. J Comp Chem;10:209–220.
38. Frank J. (1993) MOPAC. MOPAC. Colorado Springs, CO: Seiler Research Laboratory, US Air Force Academy.
39. Leardi R., Boggia R., Terrile M. (1992) Genetic algorithms as a strategy for feature selection. J Chemom;6:267–281.
40. Todeschini R., Consonni V., Mauri A., Pavan M. (2003) MobyDigs: Software for Regression and Classification Models by Genetic Alghorithms. In: Leardi R., editor. Nature-inspired Methods in Chemometrics: Genetic Algorithms and Artificial Neural Networks. Amsterdam: Elsevier; p. 141–167.
41. Todeschini R., Consonni V., Mauri A., Pavan M. (2004) Detecting ''bad'' regression models: multicriteria fitness functions in regression analysis. Anal Chim Acta;515:199–208.
42. Eriksson L., Jaworska J., Worth A.P., Cronin M.T., McDowell R.M., Gramatica P. (2003) Methods for reliability and uncertainty assessment and for applicability evaluations of classification- and regression-based QSARs. Environ Health Perspect;111:1361–1375.
43. Efron B. (1987) Better bootstrap confidence intervals. J Am Stat Assoc;82:171–200.
44. Lindgren F., Hansen B., Karcher W., Sjöström M., Eriksson L. (1996) Model validation by permutation tests: applications to variable selection. J Chemom;10:521–532.
45. Stewart J., Gill L. (1998) Econometrics, 2nd edn. London: Prentice Hall.
46. Kutner M.H., Nachtsheim C.J., Neter J., Li W. (2005) Multicollinearity and Its Effects. New York: McGraw Hill; p. 278–289.
47. Atkinson A.C. (1985) Plots, Transformations and Regression. Oxford: Clarendon Press.
48. Schüller A., Schneider G., Byvatov E. (2003) SMILIB: rapid Assembly of Combinatorial Libraries in SMILES Notation. QSAR Comb Sci;22:719–721.
49. Dempster A.P. (1967) Upper and lower probabilities induced by a multivalued mapping. Ann Stat;28:325–339.
50. Hooper G. (1699) A calculation of the credibility of human testimony. Philos Trans R Soc;21:359–365.
51. Shafer G. (1986) The combination of evidence. Int J Intell Syst;1:155–179.
52. Muchmore S.W., Debe D.A., Metz J.T., Brown S.P., Martin Y.C., Hajduk P.J. (2008) Application of belief theory to similarity data fusion for use in analog searching and lead hopping. J Chem Inf Model;48:941–948.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Figure S1. Correlation Matrix for Ki_{A3} Model (eqn 13).

Figure S2. Pareto chart of *t*-values for coefficients in Ki_{A3} Model (eqn 13).

Figure S3. Correlation Matrix for RE_{A3} Model (eqn 14).

Figure S4. Pareto chart of *t*-values for coefficients in RE_{A3} Model (eqn 14).

Figure S5. Applicability domain (for training set compounds) of the MLR models employed on prediction approach A_2.

Table S1. Chemical structures, MDs and property values of the library of N^6-substituted-4′-thioadenosine analogues.

Table S2. Checking the main parametric assumptions and the applicability domain of the overall desirability PM involved on the prediction approach A_1.

Table S3. Checking the main parametric assumptions related to the MLR models used in approach A_2.

Please note: Wiley-Blackwell is not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.

Notes

[a]CambridgeSoft. (2004) ChemDraw Ultra. Cambridge: CambridgeSoft.
[b]Todeschini R., Consonni V., Pavan M. (2005) DRAGON Software. Milano: Talete srl.
[c]Todeschini R., Consonni V., Pavan M. (2002) MOBY DIGS. Milan, Italy: Talete srl p. Software for Multilinear Regression Analysis and Variable Subset Selection by Genetic Algorithm.

SUPPORTING INFORMATION

MULTIDIMENSIONAL DRUG DESIGN: SIMULTANEOUS ANALYSIS OF BINDING AND RELATIVE EFFICACY PROFILES OF N^6-SUBSTITUTED-4´-THIOADENOSINES A_3 ADENOSINE RECEPTOR AGONISTS

Maykel Cruz-Monteagudo, M. Natália D.S. Cordeiro, Marta Teijeira, Maykel Pérez González, Fernanda Borges

CONTENTS

Table SI-1. Chemical structures, molecular descriptors and property values of the library of N^6-substituted-4´-thioadenosine analogues (1).

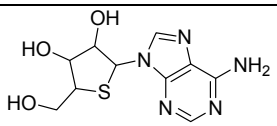
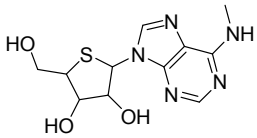
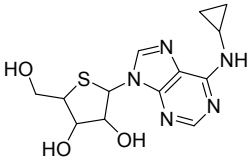
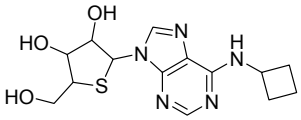
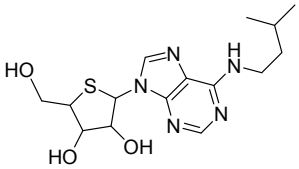
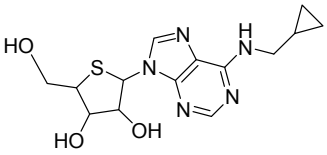
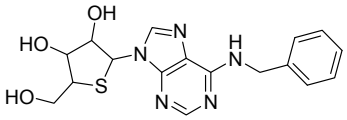
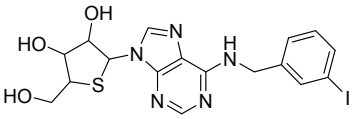
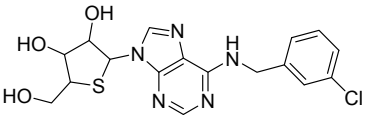
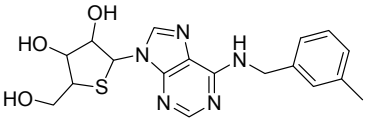
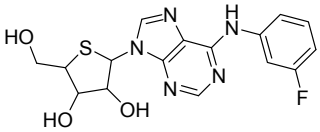
ID	STRUCTURE	PROPERTIES			MOLECULAR DESCRIPTORS			
		Ki_{A3}	RE_{A3}	$D_{KIA3-REA3}$	ARR	ALOGP2	nCIR	nCs
18a		445	0	0.000	0.476	3.482	4	3
18b		10.3	60	0.723	0.455	1.718	4	3
18c		45.2	78	0.816	0.4	0.674	5	6
18d		48	91	0.881	0.385	0.133	5	7
18e		65.3	99	0.913	0.385	0.073	4	4
18f		22.9	96	0.911	0.385	0.241	5	5
18g		155	87	0.832	0.552	0.074	5	3
18h		1.9	60	0.725	0.533	0.724	5	3
18i		6.7	62	0.736	0.533	0.878	5	3
18j		13.9	48	0.646	0.533	0.576	5	3
18k		57.6	63	0.730	0.552	0.405	5	3

Table SI-1. (*Continued...*)

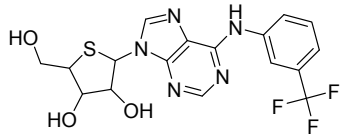
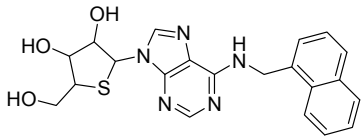
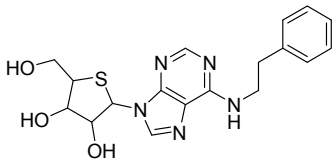
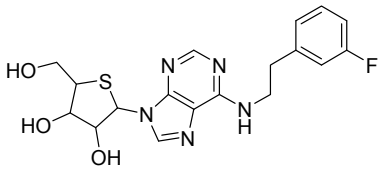
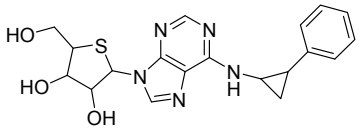
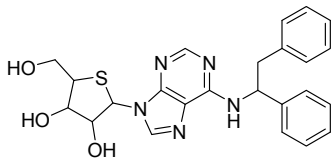
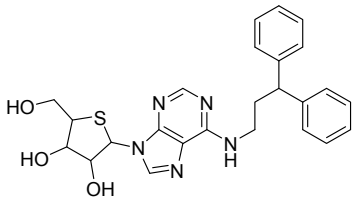
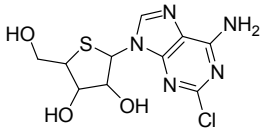
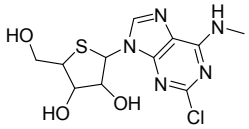
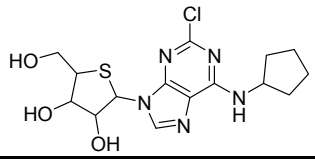
ID	STRUCTURE	PROPERTIES			MOLECULAR DESCRIPTORS			
		Ki_{A3}	RE_{A3}	$D_{KIA3-REA3}$	ARR	ALOGP2	nClR	nCs
18l		32.7	29	0.499	0.5	1.885	5	3
18m		42.2	72	0.785	0.618	1.395	7	3
18n		5.6	86	0.867	0.533	0.546	5	4
18o		11.3	89	0.881	0.516	0.285	5	4
18p		6.6	114	0.998	0.5	0.019	6	5
18q		1080	54	0.405	0.595	0.762	6	5
18r		1650	0	0.000	0.579	5.448	6	4
19a		4.9	64	0.748	0.455	1.021	4	3
19b		0.8	96	0.918	0.435	0.207	4	3
19c		94.4	68	0.750	0.357	0.897	5	8

Table SI-1. (*Continued...*)

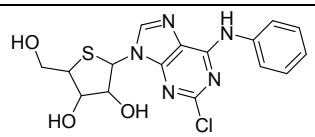
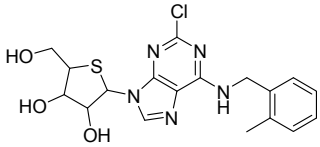
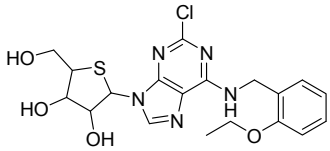
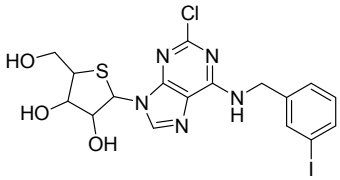
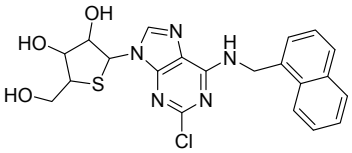
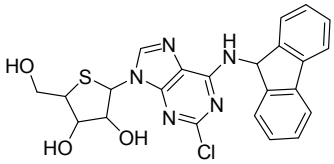
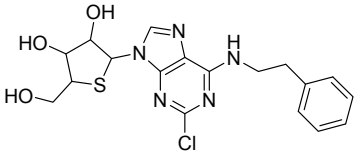
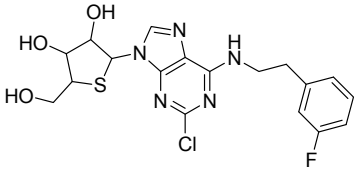
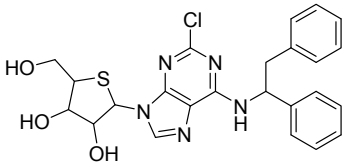
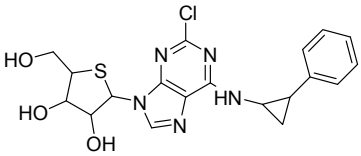
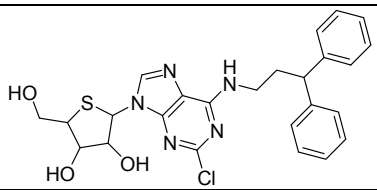
ID	STRUCTURE	PROPERTIES			MOLECULAR DESCRIPTORS			
		Ki_{A3}	RE_{A3}	$D_{KIA3-REA3}$	ARR	ALOGP2	nCIR	nCs
19d		18.2	63	0.739	0.552	1.654	5	3
19e		48.9	62	0.727	0.516	2.607	5	3
19f		17.2	60	0.722	0.485	2.134	5	3
19g		3.2	32	0.529	0.516	2.912	5	3
19h		268	45	0.575	0.6	4.148	7	3
19i		50.4	112	0.976	0.579	5.9	10	4
19j		4.4	81	0.842	0.516	0.014	5	4
19k		4.7	71	0.788	0.5	0.104	5	4
19l		1300	38	0.266	0.579	2.988	6	5
19m		1.9	102	0.946	0.485	0.518	6	5

Table SI-1. (*Continued...*)

ID	STRUCTURE	PROPERTIES			MOLECULAR DESCRIPTORS			
		Ki_{A3}	RE_{A3}	$D_{KIA3-REA3}$	ARR	ALOGP2	nClR	nCs
19n		720	0	0.000	0.564	10.174	6	4

Checking of the parametric assumptions of the MLR models (equations 12-14).

Checking of the pre-adopted parametric assumptions is a very important aspect in the application of linear multivariate statistical-based approaches (2). These include the linearity of the modeled property, normal distribution of residuals as well as the homoscedasticity and non-multicollinearity of the independent variables included in the MLR model. Once the MLR model has been set up, it is very important to check the parametric assumptions to assure the validity of extrapolation from the sample to the population. Notice that severe violations of one or various of these assumptions can markedly compromise the reliability of the predictions and inferences resulting from the MLR model (2).

We first check the linearity hypothesis by looking at the distribution of the standardized residuals for all cases. As the plots do not show any specific pattern, the idea that our PMs do not exhibit a non-linear dependence is reinforced (2).

Next, we check the hypothesis of homoscedasticity (*i.e.*: homogeneity of variance of the variables), which can be confirmed by simply plotting the square of standardized residuals related to the dependent variable (2). The data obtained reveal a significant scatter of points, without any systematic pattern, *post-mortem* validating the pre-adopted assumption of homoscedasticity for the PMs.

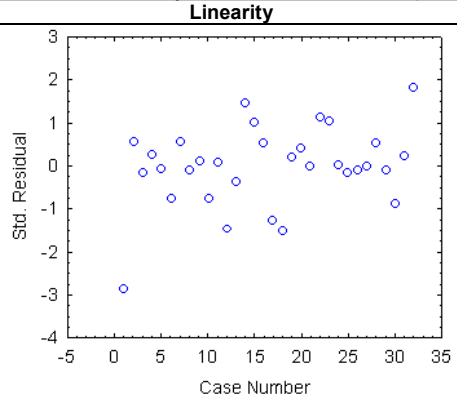
Moving on to the hypothesis of normally distributed residuals, one can easily confirm that the residuals follow a normal distribution by applying the Kolmogorov-Smirnov and Lilliefors statistical test. As the term related to the error (represented by the residuals) is not included in the MLR equation, the mean must be zero what actually occurs.

The last aspect deserving special attention is the degree of multicollinearity among the variables. Highly collinear variables may be identified by examining their pair-correlations (R_{ij}). The common interpretation of a regression coefficient as measuring the change in the expected value of the response variable, when the given predictor variable is increased by one unit while all other predictor variables are held constant, is not fully applicable when multicollinearity exists ($R \geq 0.7$). Nevertheless, the predictive ability of the model is not affecte in this situation (3). As can be noted in the correlation matrix for equation (**12**), the highest value of R_{ij} is 0.56, which confirms that the multicollinearity is not a problem in this PM and consequently the resultant inferences can be regarded as reliable. In the case of equation (**13**), the highest value of R_{ij} is 0.696, suggesting that the multicollinearity is not a serious problem in this PM and consequently the resultant predictions can be regarded as reliable. On the other hand, the multicollinearity is a problem present in equation (**14**). However, as can be noted in the pareto chart of significance of coefficients, the coefficients associated to

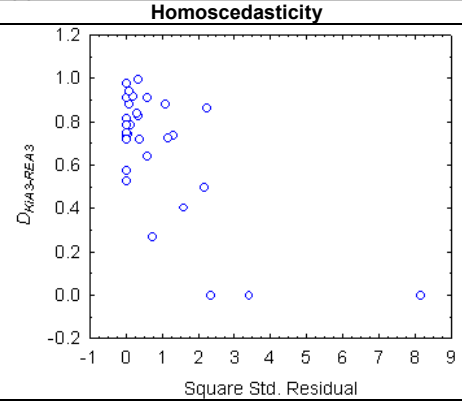
all the variables included in this model are statistically significant. Consequently, although multicollinearity is severe, actually it does not affect the statistical significance of each individual regression coefficient.

Table SI-2: Checking the main parametric assumptions and the applicability domain of the overall desirability PM involved on the prediction approach A_1.

Linearity

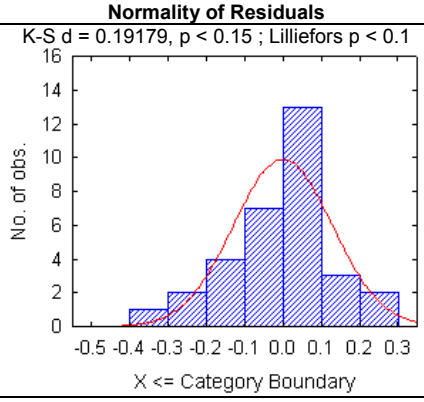


Homoscedasticity



Normality of Residuals

K-S d = 0.19179, p < 0.15 ; Lilliefors p < 0.1



Non Collinearity

	ARR	*nCs*	*ALOGP2*	*nCIR*
ARR	1.00			
nCs	-0.51	1.00		
ALOGP2	0.41	-0.15	1.00	
nCIR	0.56	0.10	0.47	1.00

Res. Mean = 0

-0.00858

Applicability Domain

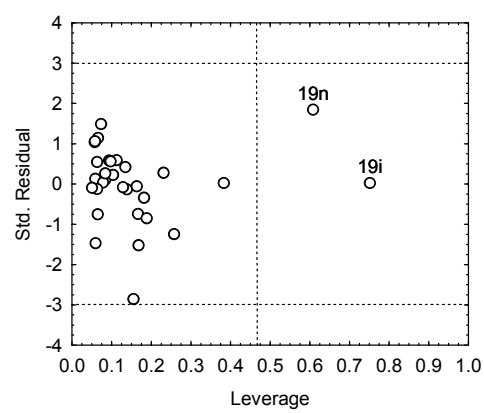
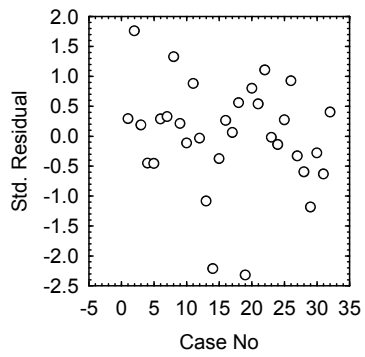
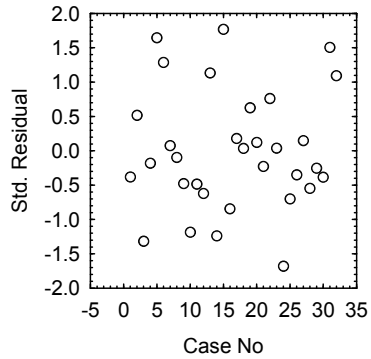
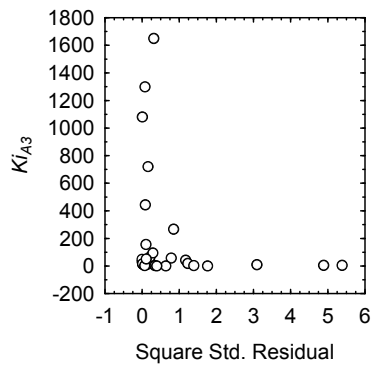
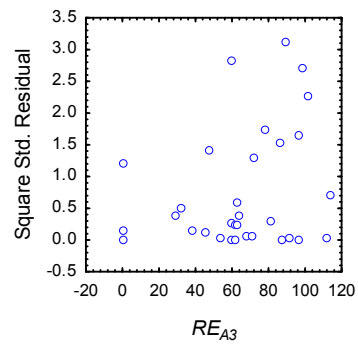
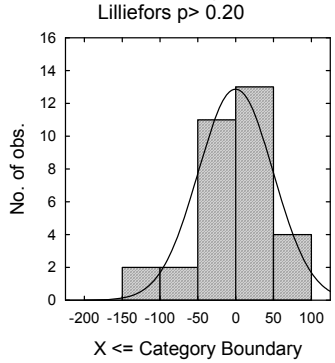
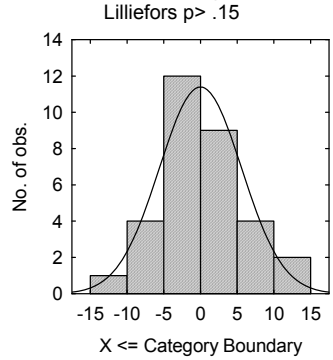


Table SI-3. Checking the main parametric assumptions related to the MLR models used in approach A_2.

	***Ki_{A3}* MLR Model (Eq. 13)**	***RE_{A3}* MLR Model (Eq. 14)**
Linearity		
Homoscedasticity		
Normality of Residuals	K-S d = 0.11479, p > 0.20 Lilliefors p> 0.20 	K-S d = 0.13735, p > 0.20 Lilliefors p> .15 
Res. Mean = 0	2.6E-12	0.000000
Non Collinearity	See Correlation Matrix	See Correlation Matrix

	D/Dr03	*GATS3m*	*BELe3*	*Mor13u*	*Mor09v*	*Mor23v*	*R7u+*
D/Dr03	1.000						
GATS3m	0.072	1.000					
BELe3	-0.201	-0.156	1.000				
Mor13u	0.437	0.001	0.224	1.000			
Mor09v	0.008	0.083	-0.587	0.189	1.000		
Mor23v	0.321	0.030	-0.696	0.039	0.621	1.000	
R7u+	0.113	0.101	-0.494	0.158	0.614	0.554	1.000

Figure SI-1. Correlation Matrix for Ki_{A3} Model (eq. 13)

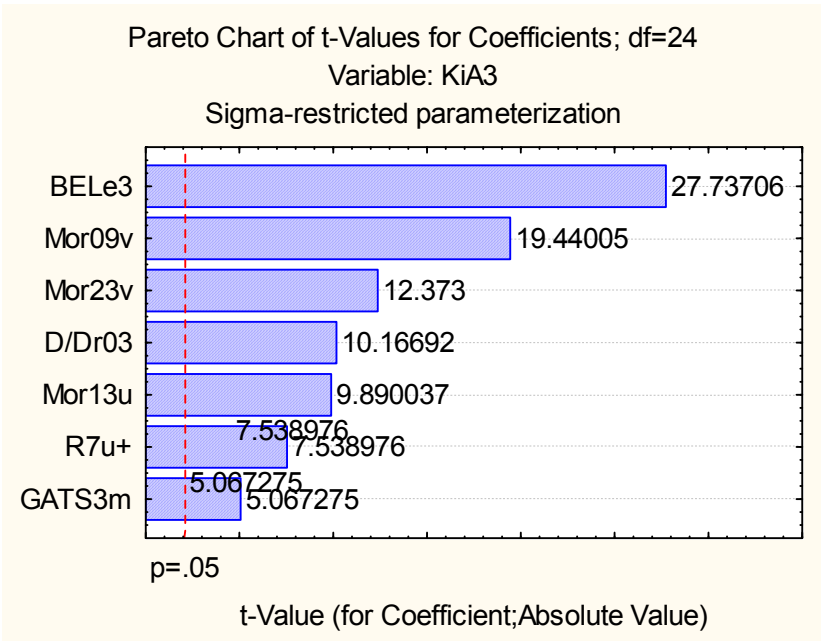


Figure SI-2. Pareto chart of t-values for coefficients in Ki_{A3} Model (eq. 13)

	PW2	D/Dr06	ATS5v	EEig10d	VEA1	H8p	ALOGP
PW2	1.000						
D/Dr06	-0.540	1.000					
ATS5v	-0.499	0.916	1.000				
EEig10d	-0.403	0.874	0.948	1.000			
VEA1	0.161	0.418	0.612	0.586	1.000		
H8p	-0.251	0.644	0.716	0.670	0.420	1.000	
ALOGP	-0.234	0.760	0.848	0.830	0.523	0.750	1.000

Figure SI-3. Correlation Matrix for RE_{A3} Model (eq. 14)

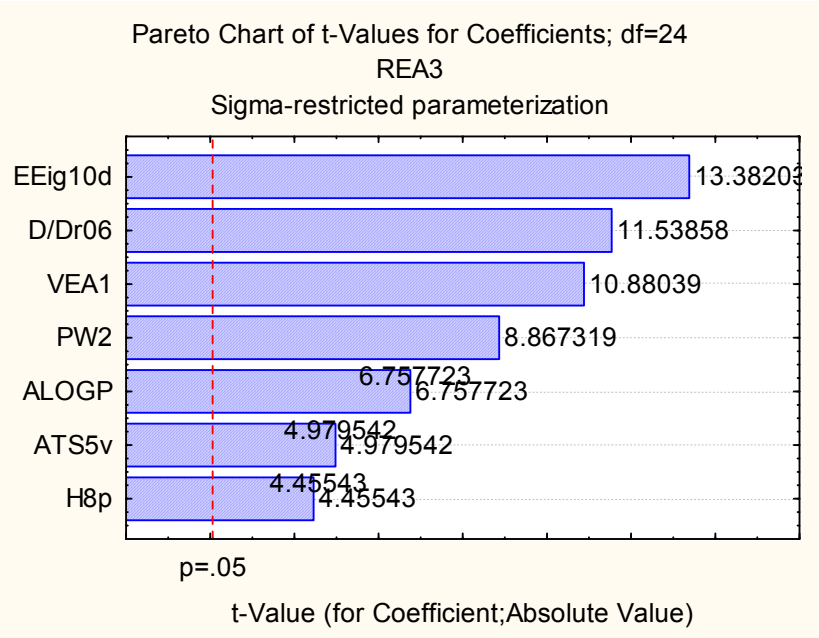


Figure SI-4. Pareto chart of t-values for coefficients in RE_{A3} Model (eq. 14)

Another aspect to consider in PMs development is the establishment of their applicability domain. The applicability domain of the PMs determined by plotting the leverage values (*h*) vs. standardized residuals (Std. Res.) of the 32 training compounds is shown in Table SI-2. From this plot, the applicability domain is established inside a squared area within ±3 standard deviations and a leverage threshold *h** of 0.468.

According to the analysis, two compounds can be regarded as structural outliers. However, no significant improvement on the goodness of fit as well as the statistical parameters was observed after their removal. So, it can be inferred that these compounds neither affect the predictive ability of the models nor the reliability of the resultant inferences, but rather enrich it with their structural information.

The applicability domain of the two PMs (Ki_{A3} and RE_{A3}) determined by plotting the leverage values (*h*) vs. standardized residuals (Std. Res.) of the 32 training compounds is shown in Figure 2. From this plot, the applicability domain is established inside a squared area within ±3 standard deviations and a leverage threshold *h** of 0.75. According to this analysis, no compounds were found to be influential for any of these PMs.

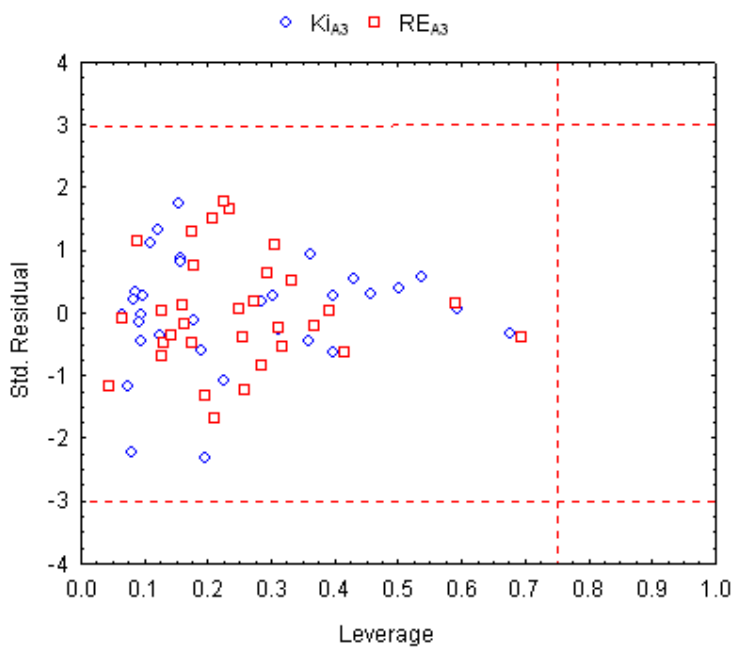


Figure SI-5: Applicability domain (for training set compounds) of the MLR models employed on prediction approach A_2.

References

1. Jeong LS, Lee HW, Kim HO, Jung JY, Gao ZG, Duong HT, et al. Design, synthesis, and biological activity of N6-substituted-4'-thioadenosines at the human A3 adenosine receptor. Bioorg Med Chem. 2006 Jul 15;14(14):4718-30.
2. Stewart J, Gill L. Econometrics. 2nd edition ed. Allan P, editor. London: Prentice Hall; 1998.
3. Kutner MH, Nachtsheim CJ, Neter J, Li W. Multicollinearity and its effects. Applied Linear Statistical Models. Fifth ed. New York: McGraw Hill; 2005. p. 278-89.

FCT Fundação para a Ciência e a Tecnologia
MINISTÉRIO DA CIÊNCIA, TECNOLOGIA E ENSINO SUPERIOR