Strategies of multi-objective optimization in drug discovery and development

Orazio Nicolotti†, Ilenia Giangreco, Antonellina Introcaso, Francesco Leonetti, Angela Stefanachi & Angelo Carotti
Dipartimento Farmaco-Chimico, University of Bari, Bari, Italy

Introduction: Drug discovery and development is a typical multi-objective problem and its successes or failures depend on the simultaneous control of numerous, often conflicting, molecular and pharmacological properties. Multi-objective optimization strategies represent a new approach to capture the occurrence of varying optimal solutions based on trade-offs among the objectives taken into account. In view of this, multi-objective optimization aims to discover a set of satisfactory compromises that may in turn be used to find the global optimal solution by optimizing numerous dependent properties simultaneously.

Areas covered: The authors review the potential of multi-objective strategies in a number of fields including: drug library design; substructure mining; the derivation of quantitative structure–activity relationship models; ranking of docking poses. The authors also discuss the potential of multi-objective strategies in controlling competing properties for absorption, distribution, metabolism and elimination/toxicity optimization.

Expert opinion: It is very clear to those who work in drug discovery and development that the success of rational drug design is largely dependent on the control of a number of, often conflicting, objectives. Therefore, multi-objective optimization methods, which have recently been introduced to the field of molecular discovery, represent the ultimate frontier in chemoinformatics. The widespread use of these multi-objective techniques has provided new opportunities in medicinal chemistry as seen through its use in a number of applications for chemoinformatics both within academia and the pharmaceutical industry.

Keywords: chemoinformatics, multi-criterion optimization, multi-objective evolutionary algorithm, multi-objective optimization

1. Introduction

Traditionally, drug discovery was considered as an inventive process mostly relying on trial-and-error and serendipity paradigms. The advent of computer-assisted techniques enhanced by far the potential of rational drug design and the chance of saving time in finding new bioactive molecules. The first unequivocal example of applying the structure-based drug design method leading to an approved drug is the carbonic anhydrase inhibitor dorzolamide[1]. Another important case study in rational drug design is imatinib, a tyrosine kinase inhibitor specifically designed to bind the bcr-abl fusion protein peculiarly expressed by Philadelphia chromosome-positive leukemia[2]. Additional examples include cimetine, the prototypical H2-receptor antagonist[3]; NSAIDs acting as selective COX-2 inhibitors[4]; enfuvirtide, a peptide HIV entry inhibitor[5]; non-benzodiazepines such as zolpidem...
Strategies of multi-objective optimization in drug discovery and development

2. Fundamentals of MOO

The linear combination of multiple objectives represents the more immediate shortcut to approach highly complex optimization problems. Basically, multiple objectives are encoded via linear combination into a single-objective function:

$$f(n) = w_1(objective - 1) + w_2(objective - 2) + \ldots + w_n(objective - n)$$

where $f(n)$ is the fitness function and $w$ are weights assigned after calibration.

The relevance of each objective is thus dependent by its own weight whose determination requires difficult and often unclear calibrations.

A popular example of linear combination of competing objectives is the Akaike information criterion (AIC), which measures the fit goodness of an estimated statistical model. Based on the law of parsimony, AIC attempts to find the model that best explains data with a minimum of free parameters. Specifically, when two or more models can equally explain data, then the one having the smallest number of parameters should be chosen. A variant of AIC was implemented to balance reliability and complexity of QSAR models by using the following AIC penalty function:

$$AIC = \log \sigma_0 + \frac{k_p}{N}$$

where $\sigma_0$ is the variance of the residuals; $N$ is the number of data points (i.e., compounds of a QSAR series); $p$ is the number of physicochemical descriptors occurring; and $k$ is a penalty factor. It goes without saying that the increment of $p$ increases the model complexity and is thus penalized.
MOO can be considered the extension of single-objective optimization (SOO) as the former encodes into the objective function a number of criteria for the optimization higher than one. As a result, unlike SOO, which aims at finding one optimum response via one objective function, MOO is the process to detect a pool of optimal equivalent solutions that meet a wider range of competing attributes relevant to a problem. In this respect, a clear example of MOO is the ‘method of marks’ voting procedure proposed by Jean-Charles de Borda [9] that represents an important step in the development of modern electoral systems, and indeed in the theory of voting more generally [10]. Such a system was initially intended for use in elections with a single winner that is designated on the basis of a rank determined by order of preference given by electors to each candidate. However, Borda count can work also as a multiple winner method when two or more seats need to be assigned. The potential of MOO methods becomes clearer in all those activities that are based on decision making that is the process of choosing the most preferred alternatives from a finite set of alternatives [11]. Such alternatives are supposed to be equivalent as they represent the best acceptable compromises among those possible.

A milestone contribution to study the MOO problem was given by Pareto [8]. In his approach, equivalent solutions, also known as non-dominated solutions, are equally good as they represent the various possible compromises among the objectives. MOO is based on the idea of Pareto optimality. In this regard, a Pareto optimal solution represents a non-dominated solution by itself as, more specifically, another solution does not exist which is better in the objectives considered. As a result, one solution dominates another one if it is either equivalent or better in all the objectives and, more strictly, it is better in at least one objective. For the ease of discussion, a multi-objective problem can be simplified to a bi-objective case. As is clearly shown in Figure 1, possible solutions are ranked according to the number of times they are dominated; thus, those non-dominated are given rank zero while those dominated are given ranks according to their position in the graph. As a result, a solution is non-dominated if the squared area, bounded by the axes and orthogonal projections from the point to the axes themselves, does not include any other point.

The entire set of non-dominated solutions represents the Pareto frontier (i.e., the curved line) obtained after simultaneously optimizing both the objectives.

As elsewhere properly reported, the amount of solutions, that is the derivation of single best solution or, alternatively, the detection of the complete Pareto frontier, is another way to classify MOO algorithms [12]. Further efforts have been also made to define three main groups of MOO algorithms [13]: the first collects those based on a priori methods, as the user is given the option to customize his or her settings before starting the optimization; the second is referred to as progressive methods, because the user can interactively guide the optimization; and the third is concerned with the a posteriori methods that give the user the chance to select the preferred solutions subset among those forming the Pareto frontier.

Pareto-based methods are tailored to take simultaneously into account a given number of properties from which a family of equivalent solutions, representing various trade-offs among the properties, is derived. One of the major benefits of the method is that user can analyze a wider range of possible solutions and thus select those matching his/her feeling and intuition.

The number of objectives (i.e., M) has the effect of increasing the time complexity of the algorithm that is of order O (MN²) where N represents the number of data point [14]. For large value of M and N, the run-time requested for searching non-dominated solutions can be thus computationally expensive and lead to non-complete convergence of the solutions or to the unevenly coverage of Pareto frontier (Figure 2). However, the implementation of niching methods can be effective to reduce the tendency to lose diversity, as well as the occurrence of premature convergence.

Furthermore, when three or more objectives are involved in MOO the bi-dimensional representation through classical Pareto frontier is unfeasible. Parallel graph coordinates are instead conveniently used to this purpose. In this representation, each line in the graph represents a non-dominated solution to the problem indicating the achieved objective values for that solution. A screenshot is given in Figure 3 reporting trade-offs among different properties characterizing a set of equivalent QSAR models.

As illustrated in Figure 3, the competing nature of the objectives is clearly shown by the crossing lines with the more accurate models consisting of larger number of terms some of which can be nonlinear.

Normally, Pareto-based approaches are joined to evolutionary algorithms that, inspired by the Darwinian theory of the survival of the fittest, are very suitable for scouting a wide search space [15]. Among evolutionary approaches, the genetic algorithms are indeed the most popular. In short, solutions, also known as individuals, are represented as strings of bits; an initial population of individuals is created, generally with random initial bits, and a fitness function is used to evaluate the quality of an individual so that the best individuals are assigned the best fitness scores. As a result, these individuals will more likely be chosen to propagate their genetic material to offspring through two basic recombination mechanisms. One of them is the mutation, which causes a random change of separate elements within a chromosome. A mutation is generally considered to be a background operator as it ensures that the probability of traversing a particular subspace is low and never zero. Mutations can generally result in pathological conditions; however, such an irregular change can also determine good results contributing to the evolution of the population through searches in new zones of the
parametric space, thus avoiding deadlock situations. The other and more powerful genetic operator is crossover, in which a portion of genetic material is taken from each parent and recombined to create new child chromosomes.

3. Applications of MOO in drug discovery and development

Several applications based on MOO have been developed and applied. Such applications can be considered as chemoinformatics tools [16] aimed at transforming data into information and information into knowledge, with the scope of accelerating decision-making process in the area of drug discovery and development. A number of examples concerned with substructure mining, QSAR, library design, docking and de novo molecular design are herein reported and discussed. In this regard, the switch from SOO to MOO enhanced the potential of all the chemoinformatics applications by enabling the exploration of multiple properties simultaneously.

3.1 MOO approaches in ligand- and structure-based design

3.1.1 Substructure mining for pharmacophore identification
The existence of recurring motifs among known drugs prompted the development of new approaches to examine the occurrence of frequently present substructures such as small chemical scaffolds, pharmacophore models and, to a large extent, privileged structures. These molecular frameworks normally exhibit both versatile binding properties [17] and good drug-like properties. As a result, they represent a solid starting point for the design of high quality leads for a wide range of biological targets through simple modifications of their functional groups. In this area, it is worth mentioning the concept of privileged structures. Initially, the idea arose from the observation that certain types of structures are biased towards certain classes of proteins. For example, spiro-based structures are preferred by GPCR receptors, while purine-based structures are preferred by kinases [18]. The rationale behind such molecular affinities is that some structures, normally hetero-cyclic rings, incorporate key chemical functional groups at the right positions and properly oriented to engage a given receptor domain. A striking example of privileged structure-based drug discovery has been the development of vardenafil (Levitra) from sildenafil (Viagra) through a me-too approach (Figure 4) [19]. Moreover, the identification of privileged structures is also a key step in the discovery of compounds biasing diverse biological targets. Numerous examples of privileged scaffolds in drugs and natural products are described in a recent paper of Welsch et al. [20].

The existence of conserved substructure and the observation that certain series of bioactive compounds [21] exhibit nearly equivalent common substructures is the proof that substructure mining is in reality a multi-objective problem.
The multi-objective nature becomes even clearer when considering that: i) 3D-common substructures can experience diverse conformational arrangements on the basis of their molecular flexibility and ii) single molecular superimposition may not explain the same mode of action of diverse molecules, as well as the different biological actions of very similar molecules as shown by multi-target ligands [22].

As often happens, distinct objectives are linearly combined through weighted sum fitness function. For instance, the program GASP (genetic algorithm superposition program), largely used for flexible molecular overlay as well as pharmacophore elucidation [23], is based on a fitness function combining the number and the similarity of molecular superimposed features, the van der Waals energy of the individual conformers and the volume integral of the superimposed conformations. A remarkable improvement is represented by the very recent program GAPE (genetic algorithm for pharmacophore elucidation) [24] which benefits from numerous enhancements resulting in a substantial increment in the reliability and applicability of the original algorithm. A striking improvement is in the perception of the pharmacophore so that each structure can contain the entire hypothesis or even partial matches. Since the publication of the original GASP program, a significant progress was the development into a multi-objective genetic algorithm (MOGA), which allows the creation of an ensemble of plausible overlays or pharmacophore hypotheses [25].

The method was initially conceived to deal with the two major issues for the pharmacophore elucidation. First, the search of the bioactive conformation that is one in the often huge number of low-energy conformers. Second, the existence of many alternative concatenations of pharmacophoric...
features or even of their subsets [26]. In this regard, it becomes clear that several plausible hypotheses could be derived from a given molecular series and the absence of structure-based data can even increase that number. In other words, it is difficult to find structures having simultaneously both high fit to pharmacophoric hypothesis and minimal conformational energy. Thus, it goes without saying that different but equivalent alternatives can exist and even in case of rigid structures equally plausible although diverse superimpositions can be determined. Unlike GASP, the method described by Cottrell et al. [26] to derive multiple pharmacophoric hypotheses involves the full conformational flexibility of the investigated structures. Besides, the simultaneously optimized objectives are the conformational energy, the volume and the feature score accounting for the energy content, the volume integral of the overlay and the goodness-of-fit to pharmacophore points. Cottrell et al. demonstrated that a single run of the multi-objective method is comparable or even better than multiple runs of GASP as the identification of a wider range of hypotheses is easily obtained. The method was further improved to explicitly account for the incorporation of partial substructure matches using as objective the conformational energy along with the quality and volume of the overlay [26].

However, a Pareto-based milestone approach, most likely the first application ever, was carried out by Handschu et al. for the flexible superposition of 3D structures [21]. The aim of that work was the identification of a set of common substructures among molecules on the basis of two objectives: the number of atoms in the substructure and the fit of the matching atoms. Intuitively, a conflict exists between these two criteria as maximizing the substructure size has the effect of determining a larger deviation in the coordinates of the superimposed atoms. The application of Pareto optimization to the superposition of vinylcyclobutane and propylcyclobutane demonstrated the contradictory aspects of the optimized objectives. In fact, a number of superimpositions belonged to the Pareto set as no other superimpositions were found having the same sizes and improved geometric fits.

Eventually, a newer approach to pharmacophore detection has been implemented in GALAHAD (genetic algorithm with linear assignment of hypermolecular alignment of datasets) which generates a set of alternative alignments by fulfilling a minimal number of assumptions about how complete pharmacophoric overlaps need to be. This algorithm adopts the fast feature multiplet technology in its operation and uses a multi-objective fitness function in order to generate multiple alignments simultaneously. GALAHAD seeks to identify a set of ligand conformations having an optimal combination of pharmacophoric similarity, steric overlap and low strain energy [27]. Very recently, Caballero reported an interesting work describing a combined approach of CoMFA (i.e., comparative molecular field analysis) and CoMSIA (i.e., comparative molecular similarity indices analysis) models with GALAHAD tridimensional pharmacophore for the rational design of novel flavonoid-based inhibitors of the aldose reductase [28].

3.1.2 QSAR and multivariate data analysis
Since the pioneering approach of Hansch et al. [29], the acronym QSAR has become a very popular word. Such a term refreshed the traditional concepts of drug design and discovery so that it represents today a new and independent topic of medicinal chemistry. Moreover, the always increasing role of QSAR is also largely documented by the number of QSAR models that have been explicitly considered in the multilateral environmental agreements established by member countries of the Organisation for Economic Co-operation and Development to assess the hazards of chemicals and, thus, to prevent severe environmental risks [30].

Nevertheless, the fundamental mission of QSAR investigations is still finding a sound relationship between the biological activity (Y) of a series of compounds and some interpretable descriptors (X) through the extraction of the most significant signals of the Y variance from the often undistinguishable fields of information and noise contained in the space of molecular physicochemical properties (X) [31].

Many approaches to QSAR have tackled this difficulty by means of customized strategies. A few examples are the forward-stepping multivariate regression analysis [32]; the Friedman’s lack of fit [33]; the FIT function conceived by Kubinyi that represents an improvement of the Fischer significance value [34]; and more recently the CVFIT index to weight the occurrence of nonlinear descriptors on the basis of a penalty scheme [35]. However, the above cited approaches were based on a single-objective function and were, thus, strategically oriented to find just one, possibly near optimal, QSAR model.

Actually, QSAR can be considered as a typical example of MOO including two or even more competitive objectives. For a better understanding, let us consider the use of the value of $r^2$ (i.e., the squared correlation between the Y response and a set of X molecular physicochemical properties). It is well known that the value of $r^2$ increases at the increment of the number of variables (X) so that more complex models tend to be more accurate.

The simple comparison between Figure 1 and Figure 5 immediately discloses the QSAR multi-objective nature. A perfect match between these two plots can be easily obtained by inverting the Y-axis. The competiveness existing between the two objectives is thus clearly evident that is the number of terms of the QSAR models and the related $r^2$ values. In fact, the latter has to be high as it represents the explained variance of the model, while the former should be low as it is a measure of the structural complexity. Similarly, it is also clear that a set of different compromises exists and, such trade-offs are those intercepted by the dotted line in Figure 5.

A milestone contribution in this field was a work describing the use of the genetic programming algorithm for model
models. The method was successful in constructing models to analyze selectivity relationships between COX-1 and -2 inhibitors [37].

Interestingly, Cruz-Monteagudo et al. introduced global QSAR studies considering simultaneously the pharmacological, pharmacokinetic and toxicological profiles of a set of molecule candidates [39]. The study relied on a MOO method based on the Derringer’s desirability function that was successful in finding a right compromise between multiple response variables. A typical example is the optimization of properties such as efficacy (i.e., potency), bioavailability (i.e., ADME properties) and toxicity (i.e., safety) to finally obtain a successful drug.

As is well known, each response variable can be predicted by submitting to least squares regression a number of independent descriptors. In certain circumstances, the same set of descriptors can forecast a pool of diverse properties by simply playing with the corresponding numerical coefficients. However, the real-life modeling of, often opposing, biological responses (i.e., maximizing potency while minimizing toxicity) requires different coefficients as well as different descriptors. In this regard, the desirability approach is a valuable strategy. It is performed in three main steps. First, the fitting models for all the multiple responses are calculated irrespective of the subset of descriptors used. Second, individual desirability functions are defined for each response. Third, the overall desirability $D$ is maximized with respect to the controllable factors.

A desirability function $d_i$ is used to assign values in the range $0 \sim 1$ (i.e., from minimum to maximum desirability) of each predicted response. Among those available, Cruz-Monteagudo used the desirability function formulated by Derringer [40] where lower, upper and target values are set for any specific response.

The global desirability is thus obtained calculating the geometric mean of the individual values of desirability as follows:

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

where $k$ represents the number of response.

Unlike other multi-objective methods, this technique is tailored to find a single best solution that depends on the descriptors used to train each predictive model; additional solutions can be found only by changing the subset of descriptors. The method was successfully applied to the simultaneous optimization of analgesic, anti-inflammatory and ulcerogenic properties of a library of fifteen 3-(3-methylphenyl)-2-substituted-amino-3H-quinazolin-4-one derivatives. Moreover, a variant of this approach was used for ranking drug candidates with unknown pharmaceutical properties from combinatorial libraries according to the degree of similarity with the previously determined optimal candidate. In particular, the method was applied to a library of 95 fluoroquinolones.

Figure 5. Variation of $r^2$ at the change of the number of physicochemical components.
reporting their Gram-negative antibacterial activity and mammalian cell cytotoxicity [41]. Another notable application of this method was the study of the selectivity of a series of arylpiperazine derivates that could interact with 5-HT1A and 5-HT2A serotonin receptor subtypes [42].

### 3.1.3 Molecular docking

Molecular docking is a method currently used to predict the preferred binding orientation of a small molecule (i.e., the ligand) to a macromolecule (i.e., the target) in order to form a stable complex. Knowledge of the binding mode may in turn be used to estimate the strength of interactions, that is, the binding affinity, through appropriate scoring functions. In principle, the search space should enumerate all the possible orientations and conformations of the protein, as well as of the ligand. In practice, the majority of docking approaches consider the protein as a rigid body and the ligand as a flexible molecule as the complete coverage of the conformational ligand–protein space is unfeasible in the time of current computer simulations. Each snapshot of the pair ligand–protein is referred to a molecular pose having its own scoring value. Among the most popular conformational search strategies so far developed, a wide application was given to systematic or stochastic torsional searches about rotatable bonds and genetic algorithms to progress new low energy conformations. The goodness of the predicted ligand–protein complex is ensured by the scoring functions [43]. The majority of scoring functions are based on the assumption that the interaction energy of a ligand to its target can be decomposed into a sum of individual contributions. In this regard, the scoring function condenses into one term (i.e., ΔGbind) any single energetic attribute; for instance, the specific protein–ligand interactions (i.e., ΔGint), the interactions of ligand and receptor with solvent (i.e., ΔGsolvent), the conformational changes in the ligand and the receptor (i.e., ΔGconf) and the local induced-fit motion during the complex formation (ΔGmotion). In other words, such a scoring function reduces the problem from multi- to single-objective via a weighted sum. However, any individual attribute can be correlated with each other and they can affect the binding affinity in several ways (i.e., positive or negative correlation). Moreover, the assumption of additivity of binding energy contribution is not rigorously valid [44].

A notable effort in the application of MOO to docking is that of Janson and Merkle [45] proposing ClustMPSO, a new hybrid particle swarm optimization algorithm for MOO. Simulating the bird flocking behavior, ClustMPSO defines a swarm of particles representing possible solutions characterized by a definite position and velocity. During the motion, the particles communicate among themselves in order to adjust their position and velocity. At any iteration, K-means algorithm divides particles into several smaller swarms, each one having its non-dominated front so that the total-dominated front is obtained from the union of the front of all the pool of smaller swarms. As far as docking is concerned, ClustMPSO minimizes two objectives: the intermolecular energy between the protein and the ligand and the intramolecular energy of the ligand. ClustMPSO demonstrated to outperform Lamarckian genetic algorithm [46] largely used to approach problems of this kind.

More recently, a MOO algorithm based on the concept of Pareto was proposed for the automated integration of structure- and ligand-based molecular design [47]. Posing and scoring themselves were transformed into objectives to automate the search of non-dominated QSAR models. Posing was assessed by measuring the atom displacements from a properly established X-ray crystal-based binding topology while scoring was evaluated by deriving regressions among docking scores and biological affinities. The goal was that of maximizing scoring while minimizing posing. It was demonstrated that several binding conformations occurred when the ligand could establish multiple contacts with different parts of the protein-binding site. Moreover, the enzyme selectivity was easily interpreted as diverse although equivalent models were made available for the user. A side benefit was also in the virtual screening campaigns of large molecular libraries of 10,240 product compounds, most of them generated through the Ugi-type three-component reaction: the obtained trade-off models outperformed docking-based runs in retrieving at high sensitivity active hits from the overwhelming number of chemically similar decoys.

The approach improved by far the results presented in a related previous work where combining posing and scoring was linearly reduced to a single-objective function [48]. The technique was challenged on a series of 3-amidinophenylalanine thrombin inhibitors that have been intensively used for testing the robustness of new medicinal computational methods [49].

### 3.1.4 De novo molecular design

De novo design approaches are aimed at discovering new molecules starting from building blocks consisting of single fragments or even single atoms. A controversial aspect is that de novo design often results in molecules whose synthetic feasibility is not always straight and easy [50].

To face this limitation, efforts have been spent to figure out how to enhance the synthetic accessibility. A few examples are the use of connection rules to join building blocks [51] or the growth of molecules from fragments obtained after the cleavage of existing compounds [52]. Despite its declared aim, the discovery of novel molecular entities is guided by their molecular similarity to other known desired molecules on the hypothesis that structurally related compounds will have also biologically related properties. A side benefit would also be that this explicit bias will encourage the scouting of a smaller and, more importantly, desirable regions of the chemical space.

This idea was profitably adopted to develop a method for predicting new chemical entities by using external QSAR [53]. However, to minimize the risk of an overestimation of the similarity to other existing molecules, the same authors improved their approach by presenting an inverse QSAR...
strategy that simultaneously optimizes molecules in the property space [54] and optimizes the residual standard deviation and the leverage to guarantee the conformity of the de novo designed molecules to the QSPR domain.

Other interesting multi-objective applications [55] were instead based on the closeness of the physicochemical property profiles (i.e., such as molecular mass, number of hydrogen-bond acceptors/donors and polar surface area) that can be directly tuned by the user.

A recent and burgeoning technique of de novo design is the fragment-based drug design (FBDD). Such an approach capitalizes on the modular binding of low-molecular mass (i.e., < 250 Da) and low-affinity fragments (i.e., $K_d$ in the range of 100 – 1000 µM) normally obtained through the decomposition of lead-like inhibitors [56]. However, the process of molecular fragmentation is highly difficult, as a serious risk of molecular oversimplification exists. As a matter of the fact, desirable fragments should maintain a low affinity of their parent leads but also should exhibit a substantial chemical diversity and ability of molecular recognition.

As a result, a fundamental rule of FBDD states that, on modification, the initial fragment lead is expected to preserve its original orientation in the protein-binding pocket and thus contributes additively to the final binding energy. A few examples of drugs developed through the FBDD approach and currently evaluated in clinical trials for diverse pathological states should be mentioned. AT7519 [57] is a cyclin-dependent kinase 2 inhibitor evaluated for its in vivo antitumor activity in nude BALB/c mice bearing early stage A2780 human ovarian carcinoma xenografts with a mean starting volume of ~ 50 mm$^3$. AT9283 [58] is instead a clinical candidate able to inhibit growth and survival of HCT116 cells and produce the polyploid cellular phenotype typically associated with Aurora B kinase inhibition. Finally, as the last example, the AT13387 compound could be cited. It is an inhibitor of the molecular chaperone heat-shock protein 90, interesting in clinical development as potential treatment for cancer [59,60].

On this assumption, a recent work of Dey and Caflisch [61] presented the algorithm GANDI (which stands for genetic algorithm-based de novo design of inhibitors) that represents a very interesting application of multiple optimization of FBDD. Basically, new molecules are automatically constructed by joining pre-docked fragments with properly sized linkers. The fitness function is, however, a linear combination of three terms accounting for the force field-based binding energy and for two measures of similarity. The first is towards known inhibitors while the second quantifies the spatial overlap to known binding modes in order to enhance the chance of scaffold hopping. The method was successfully tested on cyclin-dependent kinase 2 using a library of about 14,000 fragments and the binding mode of a known oxindole inhibitor to bias the design.

More recently, a fragment-based QSAR methodology was also proposed [62]. Using the Derringer and Suich desirability function, the optimal configuration of independent variables that could confer a balanced trade-off between selectivity and activity was optimized. The method was validated on a data set consisting of 52 hydroxysterethylamine BACE inhibitors, disclosed by GlaxoSmithKline Pharmaceuticals as potential anti-Alzheimer agents.

It is worth mentioning the recent paper of Nicolaou et al. [63] describing a new MOO de novo design algorithmic framework, that is MEGA (multi-objective evolutionary graph algorithm), for the design of structurally diverse molecules by satisfying one or more objectives that explicitly consider scores such as binding affinity, molecular similarity and a variety of properties strictly related to the chemical structure (e.g., the number of rotatable bonds, the number of hydrogen bond donors and acceptors and the molecular mass). By combining evolutionary techniques with graph theory, the algorithm enables the automated manipulation and designing of chemical structures that fulfill multiple pharmacologically important requirements concurrently. Interestingly, objectives can be used as both criteria for multiple optimization and also hard filter to circumscribe the search space. The algorithm was tested on two data sets. The first comprised a set of 53 estrogen receptor ligands, while the second included 439 compounds having a definite activity label from PubChem. A nice example of the potential of MEGA as a powerful idea generator is derivation from tamoxifen of four de novo designed structures that are representative of the Pareto frontier. As remarked by Nicolaou et al. [63], each new structure represented the more promising candidate of the list of possible novel compounds belonging to different identified structural groups.

### 3.2 Multi-objective approaches in the library design

The advent of combinatorial chemistry has permitted the synthesis of a large number of different although structurally related molecules [64]. Despite its first application in early 1960s, the invention of solid phase synthesis, for which Merrifield was awarded with the Nobel prize, such a strategy had a worldwide spread only in the 1990s [65].

Nevertheless, the impact of combinatorial chemistry was remarkable in both the pharmaceutical industry and academia. As a matter of fact, the automated preparation of large number of compounds per year implied, in parallel, a strong demand of new intelligent methods [66] of computational chemistry as the plain synthesis and the screening of large libraries did not ensure the discovery of successful compounds. In this respect, numerous strategies have thus been proposed to handle the vast number of structural possibilities and, more importantly, to select appropriate subset of virtual libraries for the actual synthesis. Initially, efforts were directed at maximizing the structural diversity on the assumption that a wide domain of structural types may ensure the coverage of large spectrum of biological responses [67]. In continuing works [68], the drug-likeness was also considered as it pertains to certain properties that would make a compound a likely
Strategies of multi-objective optimization in drug discovery and development

orally active drug. Lipinski et al. introduced a rule of thumb, called the rule of 5 [69], to filter out beforehand molecules violating $>1$ of the following criteria: not more than 5 hydrogen bond donors; not more than 10 hydrogen bond acceptors; a molecular mass not higher than 500 daltons; an octanol–water partition coefficient logP not higher than 5. Alternatively, the likelihood of a molecule exhibiting some specific bioactivity [70] can be estimated through knowledge-based approaches. The compound acquisition from likely active leads is biased towards collection of molecules having known activity and a measurable profile of some easy-to-calculate physicochemical properties such as molecular mass, number of rotatable bonds, number of hydrogen bond donors, number of hydrogen bond acceptors, number of aromatic rings, shape descriptors, polar surface area, and clogP. Moreover, the success in library design is also determined by other additional factors such as the ADMET properties, the molecular selectivity and last, but not the least, the cost of its synthesis.

It is thus clearly evident that the successful library design requires the optimization of a series of objectives. A first attempt reduced the multi-objective nature of library design to a weighted sum approach guiding the selection of compounds with both pronounced diversity and drug-likeness [71].

Another notable approach was instead based on a weighted sum function optimized via simulated annealing. Optimal solutions were biased towards reactant diversity, product novelty, similarity to known leads and pharmacokinetics properties [72]. Further works implemented the inclusion of additional criteria to constrain the design in regions of the chemical space that are considered pharmacologically appealing [73,74].

An alternative attractive strategy in library design was based on a desirability index comprising eight properties to systematically assess, after an initial random stage, the likelihood of replacement of the potential reactants in order to maximize the fitness of the entire designed library [75].

However, the above cited methods reduced the library design from a multi-objective to a single optimization problem. The potential of MOO was brilliantly reported in the milestone works of Gillet et al. [76,77] who proposed the MOGA algorithm to simultaneously handle multiple objectives such as diversity, physicochemical properties and ease of synthesis [78]. The method disclosed impressive performances in the detection of equivalent solutions for the design of diverse and focused libraries. A later paper by the same author introduced even other objectives such as the library size (i.e., the number of compounds) and the configuration (i.e., the number of reactants at each position) [79].

Another valuable approach in the design of combinatorial libraries comprised objectives accounting for the shape based similarity of the generated compounds with respect to a set of known query molecules [80]. Using a stochastic search, molecules were generated by an incremental construction method based on a given scaffold and a pool of reagents that were chosen on the basis of the similarity of their products to the reference molecules. Multiple similarities were computed for each virtual product (one for each ligand query). As a result, the algorithm calculated the values of the dominance to enumerate the products on the basis of their Pareto ranking, which was used for reagent selection. Subsequently, a validation phase was performed by generating libraries from small sets of known ligands.

An interesting paper describes the application of a multi-objective evolutionary algorithm for the design of peptidic mimotopes that can be interesting drug candidates, as such peptides are able to mimic protein epitopes [81]. However, the design of attractive mimotopes is a challenging task as peptides are highly sensitive to chemical degradation, are enormously flexible and have to pay a considerable loss of entropy to engage a given biological target. With this in mind, the design of mimotopes involves the concurrent optimization of several properties and can be hence considered a typical multi-objective problem. A Pareto-based approach was thus pursued to design peptides able to mimic antibody epitopes of thrombin and factor VIII proteins, respectively. Relevant objectives were the degree of similarity of a newly designed peptide sequence to a predefined structural motif of appropriate size; the conformational stability assessed by the estimated probability computed for a Boltzmann-distributed ensemble; the peptide length which was minimized as longer chains involve lower bioavailability, harder synthesis, increased costs and higher computational time for conformational sampling.

A new approach was also recently proposed [82]. In this work, Fischer et al. present LoFT (library optimizer using feature trees), a novel method for the design of focused libraries. Basically, LoFT uses a core with several links to join potential reagents that are taken from a fragment library. The molecular similarity is assessed by using a feature tree descriptor, which transforms a chemical structure into a tree where the nodes represent their steric and physicochemical properties (i.e., the features). In contrast to other approaches for library design, the similarity comparison prevents the full enumeration of the entire designed focused library as the comparison is performed at the product level. In addition, the optimization can evolve by using simultaneously some molecules as queries while others as anti-queries.

4. Conclusion

The transition from single to multi-objective strategies can indeed represent a big leap forward in medicinal chemistry research. The main advantage in the use of MOO methods is that several different properties can be concurrently optimized to find a family of fair compromises, which are equivalent with respect to the considered objectives. Such an approach permits to overcome the classical limitation of sequential optimization methods as there is no need to use any a priori weighting scheme to calibrate the relevance of the different objectives. On the other hand, the multiple
optimization prevents the occurrence of any potential bias towards certain properties as a wider search space is performed. In addition, the simultaneous optimization gives the chance to always have a range of potential and valid solutions, which can be profitably used for better direct discovery and development of a potential bioactive molecule. Indeed, MOO involved a paradigmatic change in drug discovery and development because, as it is today widely accepted there is no magic bullet in drug discovery while, conversely, a multi-pronged approach makes by far the best sense. In view of this, MOO approach would be, by its nature, well suited to promote a real integration of chemistry, biology and medicine with the promise of creating a new supportive and predictive environment (i.e., translational informatics, such a term is now taking place) for drug discovery, by joining chemoinformatics with chemical biology.

5. Expert opinion

For a long time, the major efforts in the drug discovery process have been primarily focused on the attempt of increasing potency, while other key physicochemical and pharmacokinetic properties (e.g., ADMET properties) have been relegated only to secondary importance. This behavior was nearly elected to a paradigma until some years ago when the accumulated evidence of the flop in the preclinical or clinical trial of promising candidate drugs rehabilitated the role of an earlier evaluation of ADMET properties for a rational design of novel drugs [7]. It is today clear that drug discovery and development are an inventive and, more importantly, multifactorial process whose success or failure are dependent on the chance of finding a fair, although difficult, compromise between diverse and, sometimes even conflicting, objectives. Other interesting and frequent real-life examples of MOO are those concerned with the need of maximizing profit while minimizing the product cost; of maximizing performance while minimizing a vehicle fuel consumption; of minimizing weight while maximizing the strength of a particular component and, eventually, of speeding up calculation time without losing accuracy. In the field of drug discovery and development, the successful rational design of new biologically active molecules is on the control of a large and often conflicting number of objectives. A few examples are molecular diversity and drug-likeness in designing combinatorial libraries; steric similarity and energy in pharmacophore generation; statistical accuracy and chemical desirability in QSAR modeling; scores and ligand displacements in the selection of docking poses; optimization of ADME properties and potency. In principle, not one, but a variety of solutions could potentially represent valuable answers to highly complex problems. And in fact, the existence of multiple properties can often involve the existence of multiple solutions whose number grows up at the increase of the number of properties under optimization. However, the challenge is to find a family of equivalent solutions that could profitably be adopted as an idea generator to facilitate an expert understanding in the rational design of new drugs. Among others, the Pareto ranking is indeed a precious strategy to find non-dominated solutions that are those representing the optimal compromise among all of the objectives under optimization. As a result, MOO methods represent a very valuable approach to give the medicinal chemist the option to have a privileged viewpoint in the critical step of decision making at the different stages of the research project development. The advantage of benefiting equivalent solutions is indubitably helpful as the medicinal chemist can easier and more confidently overcome the typical personal biases mostly determined by his or her own research experience. Interestingly, a side benefit is that the existence of multiple responses can represent a valid strategy to adjust on the fly the relevance of properties that for some reason were not sufficiently considered till that moment. Moreover, the eventual incorporation of other additional objectives is not a hard task. Optionally, the derivation of a general consensus model obtained by combining different, although equivalent, solutions can be conceived as a strategy to enhance the confidence with respect to any possible solution. The widespread diffusion of multi-objective techniques in the medicinal chemistry community is testified by the large number of applications, developed and currently used for chemoinformatics purposes. Encouragingly, a real and always increasing interest has so far emerged not only in academia but mostly in pharmaceutical companies that are fully committed to the simultaneous optimization of many biological and physicochemical properties in the prospective design of new promising drug candidates.

Declaration of interest

The authors state no conflict of interest and have received no payment in preparation of this manuscript.
Strategies of multi-objective optimization in drug discovery and development

Bibliography

28. Cabalero J. 3D-QSAR (CoMFA and CoMSIA) and pharmacophore (GALAHAD) studies on the differential inhibition of aldose reductase by flavonoid compounds. J Mol Graph Model 2010;29:363-71


Strategies of multi-objective optimization in drug discovery and development


77. Gillet VJ, Willett P, Fleming PJ, Green DVS. Designing focused libraries using MoSELECT. J Mol Graph Model 2002;20:491-8


Affiliation
Orazio Nicolotti†, Ilenia Giangreco, Antonellina Introcaso, Francesco Leonetti, Angela Stefanachi & Angelo Carotti
†Author for correspondence
Dipartimento Farmaco-Chimico, University of Bari, via Orabona 4, I 70125 Bari, Italy
E-mail: nicolotti@farmchim.uniba.it