Molecular optimization using computational multi-objective methods
Christos A Nicolaou \(^1,2\)*, Nathan Brown \(^3\) & Constantinos S Pattichis \(^1\)

Addresses
\(^1\) Department of Computer Science
University Of Cyprus
PO Box 20537
75 Kallipoleos Str
Nicosia
CY-1678
Cyprus
Email: cnicolaou@cs.ucy.ac.cy

\(^2\) Noesis Cheminformatics
66 Mestochiou Ave
2407-Engomi
Cyprus

\(^3\) Novartis Institutes for BioMedical Research
Basel
CH-4002
Switzerland

*To whom correspondence should be addressed

Current Opinion in Drug Discovery & Development 2007 10(3):316-324 © The Thomson Corporation ISSN 1367-6733

Improving the profile of a molecule for the drug-discovery process requires the simultaneous optimization of numerous, often competing objectives. Traditionally, standard chemoinformatics methods ignored this problem and focused on the sequential optimization of each single biological or chemical property (ie, a single objective). This approach, known as single-objective optimization (SOOP), strives to discover a single optimal solution to the optimization problem. Implicitly, SOOP-based methods assume that the optimal solution for an objective will also be the optimum for any other objectives involved in the profiling of a molecule. However, when these objectives are conflicting, as is often the case in drug discovery, the individual optima corresponding to the numerous objectives may vary substantially. Multi-objective optimization (MOOP) methods introduce a new approach for gaining optimality based on compromises and trade-offs among the various objectives. MOOP aims to discover a set of satisfactory compromises that can in turn be used to discover the global optimal solution by optimizing numerous dependent properties simultaneously. MOOP methods have only recently been introduced to the field of chemoinformatics. This paper first presents a brief introduction to issues related to MOOP and then surveys the application of MOOP methods in the field of chemoinformatics.

Keywords Chemoinformatics, multi-criterion optimization, multi-objective evolutionary algorithm, multi-objective optimization

Abbreviations
3-D three-dimensional, ADMET absorption, distribution, metabolism, excretion and toxicity, EA evolutionary algorithm, GALAHAD Genetic Algorithm with Linear Assignment for the Hypermolecular Alignment of Datasets, GAMMA Genetic Algorithm for Multiple Molecular Alignment, GASP Genetic Algorithm Similarity Program, LAMDA Linear Assignment for Molecular Dataset Alignment, MCSS maximum common substructures, MOEA multi-objective evolutionary algorithm, MOGA multi-objective genetic algorithm, MOOP multi-objective optimization, MoQSAR multi-objective QSAR, MoSELECT multi-objective SELECT, NCE new chemical entity, PSO particle swarm optimization, PUMP-RP partially unified multiple property recursive partitioning, QSAR quantitative structure-activity relationship, QSPR quantitative structure-property relationship, RSD residual standard deviation, SOOP single-objective optimization

Introduction
Drug discovery and development is a complex, lengthy process and failure of a candidate molecule at the development stage can occur as a result of a combination of reasons, such as poor pharmacokinetics, lack of efficacy, and/or toxicity [1-3]. Improving the pharmacological profile of a candidate molecule requires the optimization of numerous, often competing objectives (ie, biological or chemical properties), to discover the few improved molecules that represent the best compromise of the multiple criteria important for a successful drug [4]. Traditionally, when a series of compounds with adequate potency had been identified and the remaining objectives had to be taken into account, the pharmaceutical industry strived to optimize one objective at a time – starting with the binding affinity of a molecule – as part of a process involving the sequential optimization of each biological property, each time followed by screening and a large amount of ‘tweaking’ [2]. This approach often results in cycles of trial and error and can be a waste of resources and time.

Standard chemoinformatics methods for optimization, which are modeled on the traditional experimental optimization procedures, ignored the multi-objective nature of the problem and focused on the optimization of each single biological or chemical property as they became available during the drug-discovery process. Multi-objective optimization (MOOP) methods introduce a new approach for optimization that is founded on compromises and trade-offs among the various objectives. The aim of MOOP methods is to discover a set of satisfactory compromises and, through them, the global optimal solution by optimizing numerous dependent properties simultaneously. The major benefit of MOOP methods is that local optima corresponding to one objective can be avoided by consideration of all the objectives simultaneously, thereby escaping single objective dead-ends and leading to a more efficient overall process (Figure 1).

This review introduces the fundamental MOOP concepts, and the numerous applications of MOOP methodology to chemoinformatics are reviewed and organized by application category. The review concludes with comments about the current status and the future of the field.

Multi-objective optimization basics
Optimization problems can be divided into two broad categories: single-objective or multi-objective, depending on
The dashed line represents the sequential single-objective optimization (SOOP) of conflicting objectives. The continuous straight line represents the ideal optimization solution (not achievable in practice). The continuous wavy line represents the multi-objective optimization (MOOP) of conflicting objectives, whereby the solution space for satisfactory compromises to all objectives is searched simultaneously, resulting in a more direct route to the drug candidate compared to the SOOP method. (Figure adapted with permission from Wiley-VCH Verlag GmbH & Co KGaA and Baringhaus K-H, Matter H: Efficient strategies for lead optimization by simultaneously addressing affinity, selectivity and pharmacokinetic parameters. In: Chemoinformatics in Drug Discovery. Oprea T (Ed), Wiley-VCH, Weinheim, Germany (2004):333-379. © 2004 Wiley-VCH Verlag GmbH & Co KGaA.)

methods into the following three groups: (i) a priori methods that take into account the preferences of the user before the optimization process is conducted; (ii) progressive methods that enable the user to interact with the optimization process to guide the search; and (iii) a posteriori methods that produce a Pareto-front and allow the user to choose the most appropriate solution subset [5].

The straightforward approach to finding compromise solutions when numerous objectives are present is to transform the problem to a single-objective one by combining the multiple objectives. An example of this approach is the weighted-sum-of-objective-functions method [6]. According to this method, a weight is associated with each objective function and the weighted sum of the functions is taken as the new composite (or fitness) function, as defined by the following equation:

\[
f(n) = w_1(\text{Objective}_1) + w_2(\text{Objective}_2) + \ldots + w_n(\text{Objective}_n)
\]

where \(f(n)\) is the fitness function, and \(w\) is the user's defined weights. Any algorithm that is capable of solving single-optimization problems may be used to find the single best solution that minimizes the fitness function. A major drawback of the method is the selection of the most appropriate weighting, because it is often not clear how the different objectives should be ranked. A further drawback is that the method ignores the presence of the Pareto-front of the objectives, leading to unpredictability about where the solutions will lie on this surface. Finally, the method is limited in its ability to find solutions to problems involving competing objectives.

Pareto-based methods are capable of optimizing numerous properties simultaneously and thus avoid the pitfalls...
associated with methods combining multiple objectives into a single one. Pareto-based methods produce a set of solutions representing various compromises among the objectives and allow the user to choose the solutions that are most suitable for the task. The challenge facing Pareto-based methods is to ensure the convergence of well-dispersed solutions to guarantee the effective coverage of the true optimal front (Figure 3A) [7]. The Pareto ranking algorithm is of the order $O(MN^2)$, where $M$ is the number of objectives, and $N$ is the number of data points. Such an order can therefore be computationally expensive for large numbers of objectives and data points, and lead to non-convergence of the solutions (Figure 3B). A further potential issue with Pareto ranking is that the non-dominated frontier of solutions may be vast, particularly in circumstances with large numbers of objectives. The distribution of solutions on the Pareto-front may also 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 (Figure 3C). Therefore, it is often prudent to employ techniques, known as niching, to ensure appropriate coverage of the Pareto-front and avoid such occurrences.

Among the most popular algorithms used in Pareto-based approaches are evolutionary algorithms (EAs) [6]. EAs are a class of stochastic, heuristic-based approaches to objective optimization that are designed with biological evolutionary principles in mind and are especially suitable for exploring large search spaces [8]. Typically, such algorithms are based on populations of individuals that are evolved through a set of genetic operators such as reproduction, mutation, crossover (an analog of biological recombination) and selection of the fittest for further evolution. In the case of single objectives, selection of solutions involves ranking the individual solutions according to their fitness and choosing a subset. Multi-objective EAs (MOEAs) are an extension of traditional EAs that can address multiple objectives simultaneously. MOEAs exploit the availability of a population of individual solutions to map the entire Pareto-front in a single run [9••]. Selection of the solutions involves fitness assessment of each individual solution to all objectives and Pareto ranking. MOEAs employ several techniques to achieve faster convergence to the Pareto-front and often employ niching to identify solutions representative of the Pareto-optimal set.

### Multi-objective optimization in chemoinformatics

The *in silico* optimization of molecular entities has long been identified as being a multi-objective problem [1,4,10]. However, it is only in recent years that methods have been developed that account for this multi-objective aspect of the drug-design process, enabling scientists to optimize in multiple spaces simultaneously.

The first applications used MOOP methods in their simplest form, that is, the weighted-sum-of-objective-functions approach [11], with Pareto-based methods only being employed more recently. Currently, several applications have appeared in the literature making use of MOOP rationale and methodology. Table 1 summarizes selected applications of MOOP, and a detailed overview of MOOP technology applications is presented below.

### Substructure mining: Molecular alignments and pharmacophore definition

Substructure mining methods are used frequently in chemoinformatics for the discovery of structural commonalities in compound sets. These commonalities may be referred to as scaffolds, privileged substructures, motifs or pharmacophores, depending on their representation and the availability of supporting biological information [12]. Although substructure mining is often treated as single objective, with either the quality of the molecular alignment or conserved substructure size being the only objectives, the problem is in reality multi-objective, as demonstrated by the possibility of a compound set to exhibit several nearly

Figure 3. Solution sets produced by a Pareto-based multi-objective method.

(A) Effective coverage of the true Pareto-front. (B) Non-convergence of solutions to the Pareto-front. (C) Lack of effective coverage along the Pareto-front.
Table 1. A synopsis of MOOP applications in the chemoinformatics field.

<table>
<thead>
<tr>
<th>Field of application</th>
<th>Name of application</th>
<th>MO type</th>
<th>Search method</th>
<th>Objectives</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Substructure mining</td>
<td>GAMMA</td>
<td>Pareto-based, a posteriori</td>
<td>EA/GA</td>
<td>Size (number of atoms), geometric fit</td>
<td>[13]</td>
</tr>
<tr>
<td>Substructure mining</td>
<td>–</td>
<td>Pareto-based, a posteriori</td>
<td>EA/GA</td>
<td>Conformational energy, volume, feature scores</td>
<td>[14]</td>
</tr>
<tr>
<td>Substructure mining</td>
<td>GALAHAD</td>
<td>Pareto-based, a posteriori</td>
<td>EA/GA</td>
<td>Pharmacophore and steric parameters, energy</td>
<td>[17,18]</td>
</tr>
<tr>
<td>Substructure mining</td>
<td>–</td>
<td>Pareto-based, a posteriori</td>
<td>EA/GA</td>
<td>Classification accuracy, distance to known motifs</td>
<td>[19]</td>
</tr>
<tr>
<td>QSAR</td>
<td>MoQSAR</td>
<td>Pareto-based, a posteriori</td>
<td>EA/GP</td>
<td>QSAR model accuracy</td>
<td>[21]</td>
</tr>
<tr>
<td>QSAR</td>
<td>PUMP-RP</td>
<td>Lexicographic, a priori, non-Pareto-based</td>
<td>Decision tree</td>
<td>Selectivity between two (or more) biological endpoints</td>
<td>[22]</td>
</tr>
<tr>
<td>Library design</td>
<td>SELECT</td>
<td>weighted-sum-of-objective-functions, a priori, aggregative</td>
<td>EA/GA</td>
<td>Diversity, drug-likeness</td>
<td>[8]</td>
</tr>
<tr>
<td>Library design</td>
<td>PICCOLO</td>
<td>weighted-sum-of-objective-functions, a priori, aggregative</td>
<td>Simulated annealing</td>
<td>Reagent diversity, product novelty, similarity to ligands, pharmacokinetics</td>
<td>[26]</td>
</tr>
<tr>
<td>Library design</td>
<td>MoSELECT</td>
<td>Pareto-based, a posteriori</td>
<td>EA/GA</td>
<td>Product properties (eg, diversity, similarity to ligands, cost, bioavailability)</td>
<td>[25]</td>
</tr>
<tr>
<td>Library design</td>
<td>MoSELECT II</td>
<td>Pareto-based, a posteriori</td>
<td>EA/GA</td>
<td>Same as MoSELECT plus library size and configuration</td>
<td>[29]</td>
</tr>
<tr>
<td>Library design</td>
<td>–</td>
<td>Pareto-based, a posteriori</td>
<td>Stochastic method (interactive construction)</td>
<td>Similarity to known active compounds</td>
<td>[24]</td>
</tr>
<tr>
<td>Molecular docking</td>
<td>–</td>
<td>Pareto-based, a posteriori</td>
<td>Particle swarm optimization</td>
<td>Protein-ligand intermolecular energy; ligand intramolecular energy</td>
<td>[32]</td>
</tr>
<tr>
<td>Molecular docking</td>
<td>–</td>
<td>Pareto-based, a posteriori</td>
<td>EA/GA</td>
<td>Ligand energy, protein-ligand complex energy of interaction, shape match</td>
<td>[7]</td>
</tr>
<tr>
<td>De novo design and inverse QSPR</td>
<td>CoG</td>
<td>Pareto-based, a posteriori</td>
<td>EA/GA</td>
<td>Structural similarity to extant molecules of interest</td>
<td>[41]</td>
</tr>
<tr>
<td>De novo design and inverse QSPR</td>
<td>Molecule Evoluator</td>
<td>Progressive</td>
<td>EA/GP variation</td>
<td>User-directed molecule score; physicochemical property profile, drug-likeness</td>
<td>[42]</td>
</tr>
<tr>
<td>De novo design and inverse QSPR</td>
<td>CoG</td>
<td>Pareto-based, a posteriori</td>
<td>EA/GA</td>
<td>Three and six objectives, that is, predicted physicochemical property, residual standard deviation</td>
<td>[41]</td>
</tr>
</tbody>
</table>

equivalent common substructures [13•]. This possibility is especially likely when: (i) a mining process is used in an attempt to extract 3-D common substructures, where molecular flexibility is an issue and conformational space needs to be simultaneously explored; or (ii) mining a set of diverse molecules which may or may not have the same mode of action, and hence choosing a single alignment may lead to spurious conclusions. Typically, in such cases the multiple distinct objectives are combined into one objective using a weighted-sum-of-objective function. For example, the pharmacophore elucidation method GASP (Genetic Algorithm Similarity Program) uses a fitness function that is dependent on the number and similarity of the molecular features overlaid, the van der Waals energy of the individual conformers and the volume integral of the superimposed conformations [11]. The following section focuses on substructure mining applications using Pareto-based MOOP methods.

Pareto-based MOOP methods
In what was probably the first application of a Pareto-based approach in chemoinformatics, Handschuh et al used an EA-based method termed GAMMA (Genetic Algorithm for Multiple Molecular Alignment) for the identification of maximum common substructures (MCSS) in molecules [13•]. The performance of the solutions determined by the GAMMA method was defined by the size of the MCSS (number of atoms) and the geometric fit of the matching atoms of the overlaid structures. The method generated a set of compromising MCSS solutions. One solution, maximizing the geometric fit, was produced for each possible MCSS size. The authors applied the method to a set of known angiotensin II antagonists and produced MCSS corresponding to known pharmacophoric elements.

Subsequently, Cottrell et al described a method using Pareto ranking for the alignment of multiple molecules of interest to develop families of potential pharmacophore hypotheses [14]. The method aimed to satisfy three objectives of the alignment: the conformational energy of the molecules, the overlap volume integral, and a score based on the number of features in the pharmacophore and the goodness of the overlay. A well-established MOEA, the multi-objective genetic algorithm (MOGA) [15], was used in this method. The solutions returned by a single run of this method were compared to the solutions generated by multiple runs of the GASP method. The results demonstrated that a single run of the MOOP approach by Cottrell et al could identify a wider range of pharmacophore hypotheses than multiple runs of the GASP method. This Pareto-ranking-based method has been further developed to incorporate partial substructure matches using the conformational energy of the molecules, and both the quality and the volume of the overlay as objectives [16].

More recently, the GALAHAD (Genetic Algorithm with Linear Assignment for the Hypermolecular Alignment of Datasets) system has been reported for pharmacophore identification with the inclusion of a Pareto ranking procedure [17]. The system is in turn built upon the LAMDA (Linear Assignment for Molecular Dataset Alignment) algorithm [18]. The GALAHAD system identifies a set of ligand conformations that minimizes energy while also maximizing the consensus of the pharmacophoric and steric parameters of the conformation overlay.

Rajapakse et al described an application to discovering peptide motifs within conserved regions of protein sequences that are linked to specific biological functions [19]. The aim of the method was to discover a consensus motif that compromises classification accuracy and similarity objectives, to a set of known, experimentally derived motifs. The motifs were produced using an MOEA and the most appropriate motif was chosen by the user. The selected motif was found to outperform both experimental motifs and motifs extracted using standard computational approaches.

Quantitative structure-activity relationships
The development of quantitative structure-activity relationship (QSAR) methods and their related approaches have long been of interest to the drug discovery process [20]. QSAR methods attempt to generate computational models that can identify the structural and physicochemical features of molecules contributing to a certain biological outcome, which is typically associated with a single objective. Despite the availability of numerous optimization objectives, MOOP techniques have only recently been applied to the building of QSAR models.

Nicolotti et al employed a variant of an EA called multi-objective genetic programming that used Pareto ranking to optimize QSAR models [21•]. A number of conflicting objectives including model accuracy, number of terms, internal complexity and desirability, that is, interpretability of the descriptors used in the model, were considered. The technique, termed multi-objective QSAR (MoQSAR), represents QSAR models by using a so-called tree data structure, where the leaves represent constant values or molecular descriptor variables and the internal nodes represent mathematical operators, such as plus, minus or multiplication. In each iteration of the algorithm, models are evolved, the tree representations are translated into standard QSAR equations and the performance of these equations is calculated for each objective. The models are Pareto-ranked and a fitness score is determined. The result is a collection of models representing different compromises of the objectives. Nicolotti et al reported that the MoQSAR approach provided a family of solutions representing various compromises of the objectives, allowing the user to select the equation most suitable for a specific application. In addition, the objective functions did not require further calibration for each new dataset with this Pareto ranking approach.

Stockfisch proposed a non-evolutionary multi-objective technique called the partially unified multiple property recursive partitioning (PUMP-RP) method for generating QSAR models [22]. The PUMP-RP method is an extension of the decision tree classification method that is able to accommodate multiple objectives. The process generates a decision tree in a two-step growth process. If progress toward optimization can be made on all objectives, a generic decision tree is grown by performing splits on descriptors...
that are associated to all objectives. When this is no longer possible, a split on property type is performed and property-specific sub-trees are grown. The PUMP-RP method was successfully used to construct models to analyze selectivity relationships [23]. This approach resembles a lexicographic-type analysis and requires an a priori ordering of the objectives. The objectives are optimized sequentially until no more progress can be achieved. It is worth noting that ordering of the objectives is crucial to the method because different orderings produce different models. This reliance on the order of the objectives is a potential drawback to the method because the relationship between objectives is not always clear and therefore ranking objectives is often subjective.

**Library design**

The advent of combinatorial chemistry and high-throughput screening systems has enabled the synthesis and screening of large numbers of compounds for the purposes of drug discovery. However, simply synthesizing and screening large compound libraries may not increase the probability of discovering promising hits [9••]. In order to maximize the probability of discovering such compounds, combinatorial chemistry libraries must be carefully planned and a number of design objectives should be simultaneously optimized [10].

Initial efforts in designing combinatorial libraries were directed at generating diverse libraries representative of the full possible combinatorial population [24]. More recent efforts shifted toward the design of focused libraries occupying restricted regions of the chemistry space with the boundaries defined by the available knowledge of the biological target [25]. Regardless of the library design focus, it was realized early on that a single property or measure alone seldom defines the quality of the molecular library [10]. Other factors such as absorption, distribution, metabolism, excretion and toxicity (ADMET) properties, selectivity and cost are now considered crucial to successful library design. Consequently, the focus of combinatorial library design has now shifted toward designing libraries based on a number of properties simultaneously [3,9••].

The first applications to address the multi-objective nature of the combinatorial library design problem resorted to the weighted-sum-of-objective-functions method. The SELECT method was proposed to design combinatorial libraries containing compounds with high diversity and drug-likeness by using a SOOP method based on an EA [8]. Zheng et al also described a weighted-sum-of-objective-functions method termed PICCOLO in which various objectives were considered including, for example, reagent diversity, product novelty, similarity to known ligands, and pharmacokinetics [26]. The various objectives were combined into a single objective, and the PICCOLO method used simulated annealing, an established stochastic optimization technique, to search for optimal solutions. A similar approach, proposed by Rassokhin and Agrafiotis, allows common design objectives, such as diversity, to be combined with additional criteria in order to bias the design toward more pharmacologically relevant regions of chemical space [10,27]. More recently, a method based on a desirability index comprising eight compound properties was proposed by Le Bailly de Tilleghem et al [28]. The method starts by using a random set of reagents to form an initial library. Subsequently, the reagents are ejected and replaced by other reagents in a systematic manner by evaluating the contribution of each reagent to the fitness of the designed library.

Despite the reported successes of the methods based on the weighted-sum-of-objective-functions, compromising conflicting objectives in combinatorial library design remained a problem [9••]. To address this limitation, the multi-objective SELECT (MoSELECT) method [9••,25] was developed to introduce the Pareto concept to combinatorial library design. The MoSELECT method employs the MOGA algorithm to simultaneously handle multiple objectives such as diversity, physiochemical properties and ease of synthesis [15]. The method was proved to be effective at finding families of equivalent solutions for both diverse and focused library designs [29]. Comparisons between MoSELECT and SELECT demonstrated that the MoSELECT method could both locate more optimal solutions and expand along the Pareto-front of the objectives compared with the SELECT method. An extension to the MoSELECT method, MoSELECT II, has been designed to treat library size (e g , number of compounds) and configuration (e g , number of reagents at each position) as additional objectives [29].

Soltanshahi et al proposed a method where the objectives in combinatorial library design correspond to a shape-based similarity of the generated compounds to a set of known ligands [24]. The incremental construction method, a stochastic approach, was used to generate libraries based on a supplied scaffold and a set of reagents. At each step of this method a small random number of reagents were considered. Appropriate reagents were chosen based on the similarity of their virtual products to the query molecules. Multiple similarities were calculated for each virtual product (one for each ligand query) and the algorithm calculated the domination relationships among the products and defined their Pareto ranking, which was used for reagent selection. Validation was performed by generating libraries from small sets of known ligands.

**Docking**

Docking describes a process by which a ligand and a protein target fit together in 3-D space [30]. A complete and exhaustive search of all possible binding configurations of a ligand and protein is not possible because of the complexity of the problem, thus optimization techniques are frequently used [7]. Scoring functions are used to measure accurately the quality of the calculated protein-ligand complex [31]. Often, the scoring function combines several terms describing the goodness of fit and the energy requirements of the docking configuration into a single function, thus reducing the problem from multi- to single-objective [32•]. For example, the weighted-sum-of-objective-functions approach was used by Jones et al to combine several energy terms of a docking scoring function into a single term [33]. An EA was then used to explore the space of possible protein-ligand poses. AutoDock, a widely used docking software package, uses a similar approach [34]. Recently,
applications of MOOP methodology have been gaining acceptance in the docking community; a selection of such applications are described below.

**Application of MOOP to docking**

Janson *et al* described a docking optimization application termed ClustMPSO (based on the particle swarm optimization [PSO] algorithm) that minimizes two objectives: the intermolecular energy between the protein and the ligand and the intramolecular energy of the ligand [32\textsuperscript{•},35]. The PSO algorithm, the generation of which was inspired by the behavior of flocking birds, defines a swarm of particles representing candidate solutions each with its own position and velocity. The particles move through hyperspace and communicate good positions among themselves in order to adjust their position and velocity. In the ClustMPSO application the particles are clustered in every iteration and divided into several swarms on the objectives’ compromise surface. The goal of this approach is to find several interesting, near-optimal docking configurations in addition to the best docking configuration. Janson *et al* reported that this application was tested on two docking experiments using real data and was found to outperform standard techniques.

A method of multi-objective protein-ligand docking has been presented in the computer science literature [7]. The reported docking system optimizes three objectives: the internal energy of the ligand, the van der Waals and electrostatic energy of interaction of the protein-ligand complex, and the shape complementarities of the protein and the ligand. Three state-of-the-art MOEA methods were available to the system. The methods tweaked the docking compound to find new protein-ligand configurations, and evolved the population of configurations to approximate the Pareto surface compromising the objectives. Validation on protein-ligand complexes whose docking configurations are known showed promising results for this approach.

Several additional research groups are active in the field of Pareto-based docking optimization applications. Mardikian *et al* described a docking method, based on a MOGA capable of simultaneously optimizing individual interaction energy types, which revealed insights into the relative importance of the various objectives [36]. A MOEA is also used by Zoete *et al* in their docking program EADock [37].

**De novo molecular design and inverse QSPR**

While the design of new chemical entities (NCEs) has long been recognized as being of high importance for drug discovery, it is only relatively recently that methods have been proposed in which the concept of multiple objectives has been introduced to the optimization of NCEs.

Regarded as the first multi-objective *de novo* design method [38\textsuperscript{•}], the system proposed by Brown *et al* optimizes NCEs via Pareto ranking a family of candidate solutions according to their molecular similarity to a set of existing molecules of interest [39]. This method is based on the hypothesis that structurally similar molecules will exhibit similar properties and also provide a controlled method of exploring the chemistry space of interest. This molecular similarity optimization method for designing NCEs has been applied successfully to fill gaps in property space by using external quantitative structure-property relationship (QSPR) models to predict NCEs [40\textsuperscript{•}]. However, a recognized limitation of optimizing in similarity space is that the optimized NCEs are necessarily highly similar to existing molecules. To overcome this problem, Brown *et al* have more recently published a similar method using the inverse QSPR methodology that optimizes molecules directly in property space, allowing multiple molecular properties to be optimized simultaneously [41]. In this approach to solve the inverse QSPR problem, Brown *et al* use not only the predicted properties, but also two indicators of prediction quality – residual standard deviation (RSD) and leverage – to ensure the resultant molecules still conform to the QSPR model space.

In addition to the methods employed by Brown *et al*, two software programs have been released by Cidrux BV [42\textsuperscript{•}] and Coalesix Inc [43] that provide a progressive, user-directed *de novo* design approach to NCEs. For both programs, the user selects a score for each of the candidate molecules based on the molecular structure itself together with a physicochemical property profile of the molecule covering properties such as molecular weight, number of hydrogen-bond acceptors/donors, and polar surface area. In this approach, the user is the multi-object optimizer and can, therefore, focus the area of exploration to those regions deemed to be of most interest for the particular application.

The most significant challenges that remain in *de novo* NCE design, whether the problem is single- or multi-objective, are estimating molecular stability and synthetic accessibility. The optimization systems that incorporate human interaction implicitly take these two factors into account as objectives. However, such objectives may also be included as components of automated optimization systems once reliable and robust computational methods have been developed for this purpose.

**Conclusion**

Many endeavors in modern drug discovery are hindered by the lack of consideration of multiple molecular properties in the early stages of lead identification and optimization. Indeed, one of the common causes for lead compounds to fail in the later stages of drug discovery is the lack of consideration of multiple objectives (eg, ADMET properties) at the early stage of optimization of candidate compounds [4]. Standard chemoinformatics approaches still largely ignore multiple objectives and optimize each biological property sequentially. Most pharmaceutical companies readily adopt computational approaches to predict biological properties of molecules but fail to take advantage of the availability of multiple compound properties to guide molecular optimization. In order to fully exploit this multitude of available properties the industry needs to realize the fundamental shift required in the optimization paradigm and invest in the transition from single- to multi-objective technologies [1].
Encouragingly, the chemoinformatics community recognizes that the parallel optimization of potency, selectivity and ADME/T properties via predictive computational models shows great promise for drug discovery [44]. In order to address this new significant challenge to the field, the community has begun to explore MOOP methodology, and some first applications showing promising results can already be found in the literature. The authors of this review anticipate that in the near future the pace of MOOP methodology adoption will increase considerably, and that standard chemoinformatics techniques will regularly be applied with the simultaneous optimization of multiple properties of interest to the discovery of new drugs.

References

• of outstanding interest
• of special interest


• The first application of Pareto MOOP to QSAR.


- A detailed explanation of a novel particle swarm-based MOOP algorithm for docking.


- An excellent review of the state of de novo design in chemoinformatics.


- The first application of Pareto MOOP to de novo molecular design.


- A de novo design MOOP application where users assume the role of the fitness function. Users select candidate molecules for further evolution after each iteration.
