ELSEVIER

Evolving molecules using multi-objective optimization: applying to ADME/Tox

Sean Ekins[1,2,3,4], J. Dana Honeycutt[5] and James T. Metz[6]

[1] Collaborations in Chemistry, 601 Runnymede Avenue, Jenkintown, PA 19046, USA
[2] Department of Pharmaceutical Sciences, University of Maryland, Baltimore, MD 21201, USA
[3] Department of Pharmacology, Robert Wood Johnson Medical School, University of Medicine & Dentistry of New Jersey, Piscataway, NJ 08854, USA
[4] Collaborative Drug Discovery, 1633 Bayshore Highway, Suite 342, Burlingame, CA 94010, USA
[5] Accelrys, 10188 Telesis Court, Suite 100, San Diego, CA 92121, USA
[6] GPRD R4DG, Department of Fragment Screening and Lead Characterization, Abbott Laboratories, 100 Abbott Park Road, Abbott Park, IL 60064, USA

Modern drug discovery involves the simultaneous optimization of many physicochemical and biological properties that transcends the historical focus on bioactivity alone. The process of resolving many requirements is termed 'multi-objective optimization', and here we discuss how this can be used for drug discovery, focusing on evolutionary molecule design to incorporate optimal predicted absorption, distribution, metabolism, excretion and toxicity properties. We provide several examples of how Pareto optimization implemented in Pareto Ligand Designer can be used to make trade-offs between these different predicted or real molecular properties to result in better predicted compounds.

Introduction

When we think of evolution, we tend to think of the continual process of gradual change or adaption that an organism undergoes, overcoming multiple challenges acting on the population to continue the survival of the fittest, a process of natural selection. This well-known principle can help us in developing new molecules with optimal physicochemical properties that have survived filtering and molecular transformation. The awareness has increased that successful drug discovery increasingly requires more than just finding a molecule that is highly potent at the target: it needs to be as close to optimal for these other desired physicochemical properties too. This is perhaps important for compounds that have activity against multiple targets or are promiscuous [1–3], which has enabled molecule repurposing in some cases [4–6]. The molecule of interest might also need to be orally available to be absorbed and is required to reach its target before being cleared from the body. At each step (from absorption to reaching its target and elimination), the molecule has to cross multiple membranes. Each of these physical processes requires different physicochemical properties – for example, moderate hydrophobicity is good for membrane penetration [7] but not for solubility. Pharmaceutical scientists have at their command an array of real and virtual data on their molecules of interest (e.g. predicted physicochemical properties alongside measured metabolic stability) and data on published molecules that are structurally similar or active at the same target(s). Drug discovery scientists, therefore, need to consider the many diverse requirements for a molecule alongside its bioactivity, any of which might in be in conflict with one another. The process of resolving these conflicting requirements is termed 'multi-objective optimization' or 'multidimensional optimization'. It is not unique in being applied to drug discovery but has been used previously in such varied domains as optics, electronics, cancer treatment and product development [8–11]. Some have suggested the need for a simultaneous, multi-objective optimization of various molecular properties with efficacy data [12–15]. Similarly, we need to optimize the absorption, distribution, metabolism and excretion (ADME) and toxicity data that are generated for compounds [16] much earlier in the process using high-throughput screening (HTS).

With increased compound throughput in drug discovery, we have seen lead compounds derived from HTS hits frequently having undesirable properties [17–19], such as increased hydrophobicity (log *P*) and decreased solubility. New leads should be

Corresponding author:. Ekins, S. (ekinssean@yahoo.com)

1359-6446/06/$ - see front matter doi:10.1016/j.drudis.2010.04.003

more stringently selected in terms of their molecular properties (MW < 350, log P < 3 and affinity ~0.1 μM [20] or other desirable properties [21], e.g. the rule of five [22]). Marketed drugs that are inhaled have been found to possess more hydrogen bonds and have generally lower c log P than drugs that are not inhaled [23]. In parallel, many companies have instituted computational filters to remove undesirable molecules from their HTS or from vendor libraries. Examples include removal of swill (REOS) from Vertex [24] and filters from GSK [25] and BMS [26]. Abbott reported a sensitive assay to detect thiol-reactive molecules by NMR (ALARM NMR) [27,28]. The data from this have also been used to create a Bayesian classifier model to predict reactivity [29].

The understanding of the importance of structural transformations that can be made can also facilitate modifications to molecules that might increase activity [30] or enable the optimization of off-target activities [31,32]. For example, subtle chemical modifications can dramatically alter pharmacological and ADME/Tox profiles, and physicochemical properties can be thought of as driving these differences (e.g. the slight modification of a clinical candidate for cancer, tipifarnib, resulted in a potent inhibitor for Chagas disease active with more predictable drug-like qualities) [33]. Understanding which parts of the molecule are important for activity has led to measures of ligand efficiency or fit quality that balance potency with size or other properties (e.g. molecular weight, number of heavy atoms or polar surface area). The result is that smaller, more efficient molecules might have better drug-like properties, and this is particularly prominent in fragment-based drug design [34–37].

Simultaneous multi-objective drug discovery is clearly necessary [38] to improve drug discovery output and increase efficiency. The most cost-effective approach for drug discovery is to simulate as much as possible using computational methods [39–41]. These have a greater throughput than *in vitro* and could impact the quality of the molecules generated by helping to address potential liabilities that lead to later-stage failures. Now we are seeing repeatedly that large volumes of *in vitro* data are being used as inputs into computational models for oral bioavailability [42], human ether-a-go-go-related gene (hERG), Cytochrome P450 (CYPs) [43] or other ADME properties using an array of machine learning algorithms such as support vector machines [44–46], Bayesian modeling [47], Gaussian processes [48] and others [49]. A wide variety of computational methods (e.g. ligand-based, structure-based and hybrid methods) can have an immediate impact on the prediction of metabolic transformations that can guide synthesis of more metabolically stable compounds [50] involving many different enzymes and potential sites for metabolism [51–57]. It is, therefore, important to block labile sites but consider retaining or improving the bioactivity, solubility and other properties of a molecule.

We and others have indicated the need for integrated simulation tools [39,58] that bring together different types of models to improve the drug discovery decision-making process. We initially suggested a system of multidimensional scoring using many ADME/Tox filters in decision-making [40], which seems to have also been indicated by others [59]. ADME filters have led to the derivation of rules for most of these properties, which are heavily influenced by molecular weight and c log P [60].

When there are multiple endpoints – experimental data, computational predictions or both – then trade-offs have to occur and, ideally, we should use a method that can consider each variable and enable the selection of the probable best compounds. Multi-criteria decision methods [61], which are a type of multi-objective optimization [62], are one approach to the simultaneous optimization of several variables [63] based on desirability functions [64] or desirability indexes [65]. In contrast to the prevailing trend focusing on simple rules to filter or select compounds, we would rather consider many more variables without hard cut-offs. Previously, we indicated how such multi-objective optimization approaches could work in drug discovery simultaneously rather than optimizing single properties sequentially [13] to derive a set of Pareto-optimal (see description below) compounds. We will now expand greatly on this to show the developments in the field over nearly a decade and new ways to apply these technologies to evolve better quality compounds. Although there is some review of the use of multi-objective optimization in bioinformatics and computational biology [66], our focus is limited to cheminformatics.

Approaches to multi-objective optimization

Broadly speaking, there are at least two approaches to numerically solving a multi-objective optimization problem. In the simplest case, we will assume that there are just two objectives, but all of the available techniques can be generalized to any number of objectives (with some caveats). We will also assume that we seek to optimize the properties of individual compounds, as opposed to properties of compound libraries (such as structural diversity) or mixtures of compounds. We discuss library optimization in the next section.

A basic first approach to multi-objective optimization is to somehow combine all of the objectives into a single objective function. This allows classical single-objective optimization techniques to be applied to the problem. One variant of this approach is weighted-sum optimization, in which the overall objective is a weighted arithmetic mean of values representing the individual objectives [67]. For example, suppose we have developed models to predict a specific activity and the toxicity of chemical compounds. To find the compounds that have the greatest activity and the lowest toxicity, we construct the following objective function:

$$O = \frac{w_a A}{s_a} - \frac{w_t T}{s_t},$$

where A is the predicted activity of a compound, T is its predicted toxicity, w_a and w_t are positive weighting coefficients whose values we assign according to the relative importance we place on maximizing activity versus minimizing toxicity, and s_a and s_t are scaling factors to correct for possible differences in range for the A and T values. (A common approach is to use as the scaling parameter the standard deviation of the variable value over the set of compounds we are considering.)

Another variant of the single-objective approach to multi-objective optimization is to use as our objective function a weighted geometric mean of desirability functions based on the individual objectives. This approach is known as 'desirability optimization methodology' and has certain advantages over the weighted-sum approach [63].

Having established an objective function by either means, we then seek to maximize it (or minimize it, depending on how the objective is defined). If our optimization is over a fixed set of

compounds, this simply requires sorting the compounds according the value of *O*. If, instead, we wish to explore a space of compounds whose structures are not predefined, we need to iteratively 'mutate' the structure, beginning from one or more seed compounds, until the objective has been maximized. This, of course, will also require software and a rule base for generating synthetically and chemically reasonable compounds. In the end, we are left with a single compound that is 'best' according to our criteria.

The single-objective approach to multi-objective optimization has two major drawbacks: first, ordinary single-objective optimization yields only a single 'best' solution because suboptimal solutions are not retained (or if they are, as possible in evolutionary algorithms, there is no guarantee that the extra retained solutions meet any sort of optimality criterion themselves – e.g. by being solutions to an objective function defined by different weights). Information on near-optimal solutions is lacking. Nor does single-objective optimization provide a way, short of running multiple optimizations with differing weights, of answering such questions as, 'In return for a slight decrease in activity, is it possible to greatly reduce the toxicity?'

Second, the meaning of the weights is vague. Although their effect is mathematically well defined, their meaning is hard to grasp intuitively. For example, what exactly does it mean to say, 'Low toxicity is twice as important as high activity'? The only way to see how the weights affect the results is to vary them and run multiple optimizations.

The second broad approach to multi-objective optimization does not require the prioritization or weighting of individual objectives and is a focus of this review. This approach is known as Pareto or trade-off optimization [67,68] and is an alternative method to the desirability function approach. One of the advantages of Pareto optimization is that all objectives are put on an equal footing [69]. Figure 1 illustrates this in practice.

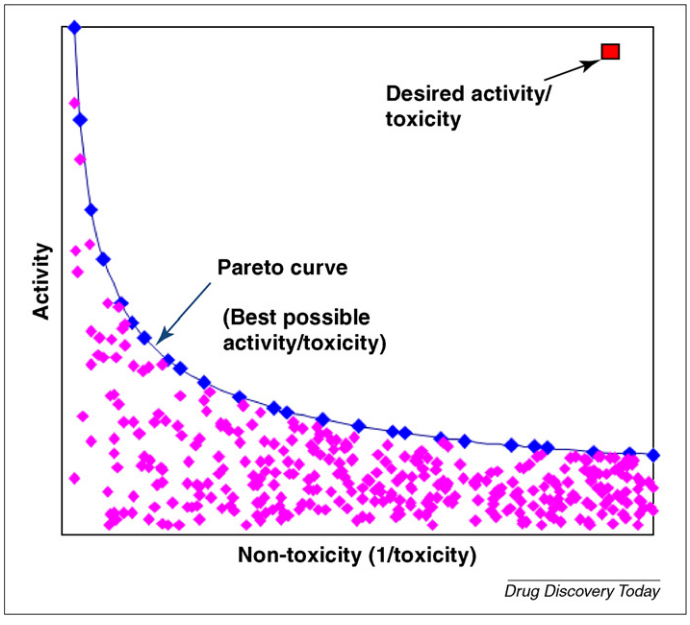

FIGURE 1

Schematic illustration of Pareto optimization with hypothetical data.

The data in this graph are hypothetical but show the typical pattern found in multi-objective optimization problems. We seek to maximize both the activity and the nontoxicity of our compounds. (We define nontoxicity here as simply the reciprocal of toxicity, assuming the latter is always greater than zero.) Each point on the graph represents a different compound. Ideally, we would like an activity and toxicity profile corresponding to the red point in the upper right; however, according to the displayed data, the most active compounds tend to be the most toxic and the least toxic compounds tend to have low activity.

Observe, however, that many compounds (those represented by the magenta points) have relatively high toxicity, yet low activity compared with other compounds with an equal or even lower toxicity. Inspection of the graph should convince you that the only compounds of even tentative interest for this optimization problem are those represented by the blue points. To be precise, for every magenta point, at least one blue point is better with regard to toxicity or activity, while being at least as good with regard to the other property. (In the jargon of Pareto optimization, we say that the blue point dominates the magenta point.)

The blue points define the Pareto-optimal curve (with more than two objectives, this curve becomes a surface or hypersurface). These are the points with the best possible trade-off between the two objectives. Note that we have not yet indicated any weights for the objectives – that is, any preference for lower toxicity over greater activity or vice versa (e.g. a preference for greater activity would have us focus on points toward the upper left of the curve).

Pareto optimization is the process of generating a set of points on the Pareto-optimal curve or surface. This accomplishes two things. First, it eliminates from consideration the vast number of compounds that are not of interest, irrespective of any weighting or priority of one objective over another. Second, it retains multiple compounds that might be of interest, enabling one to visually inspect the trade-offs involved in improving some properties at the expense of others and to choose the best Pareto-optimal compound accordingly. Because Pareto optimization generates multiple solutions, this visual analysis is a key follow-up step to the optimization.

Running a single Pareto optimization is equivalent to running multiple weighted-sum optimizations with varying weights. Under certain conditions, the equivalence is exact. That is, under certain conditions, every Pareto-optimal point is equivalent to the solution of a weighted-sum optimization with a different (unspecified) set of weights [70]. Thus, a weighted-sum single-objective optimization yields one of the blue points in Fig. 1. A multi-objective Pareto optimization yields all of the blue points.

Applications of Pareto multi-objective optimization: library optimization and beyond

Desirability functions and Pareto optimization [71] have been applied to numerous problems, including those in the area of compound and library optimization [72–76]. In library optimization, we seek to optimize not only properties of individual compounds in the library (although that might be one component of the problem) but also properties of the library as a whole, such as structural diversity and scaffold coverage to ensure that different chemotypes are represented [77].

For example, we might wish to find a subset of 1000 compounds, taken from a library of ten million, that maximizes both diversity and the degree to which the compounds are drug-like on average. Here, the unit being optimized is the subset rather than the individual compound, and the Pareto curve is defined by a group of these subsets. Some subsets along the Pareto curve will be more diverse and less drug-like on average, and others will be less diverse and more drug-like on average. The Pareto approach has been used by others to solve this and similar problems [72,78,79].

One drawback of the Pareto approach is that as the number of properties to be optimized increases, the number of samples on the Pareto-optimal surface tends to increase exponentially. The reason for this can be seen by considering that this 'surface' is a one-dimensional curve for two optimization properties, a two-dimensional surface for three properties, a three-dimensional hypersurface for four properties, and so on. This implies a great increase in the required memory and computation time with the number of properties to be optimized. It also suggests that the researcher can be faced with hundreds or thousands of optimal samples to choose from at the end. To mitigate these problems and reduce the number of properties to be optimized, it makes sense to combine correlated properties into composite properties and then Pareto-optimize the composite properties (e.g. multiple measures of toxicity might be combined into an overall toxicity property and multiple ADME properties might be combined). The approach then becomes a hybrid of the single-objective and Pareto methods. Such an approach has been described recently with a case study for automated drug design for estrogen receptor antagonists using desirability indexes to reduce the number of objectives [80].

A recent review has summarized many computational multi-objective methods for molecule optimization [81], some of which are discussed below. Desirability-based multi-objective optimization (MOOP-DESIRE) was proposed for filtering combinatorial libraries [82] and global QSAR studies studying NSAIDS with analgesic, anti-inflammatory and ulcerogenic properties, which all needed optimization [83]. The multi-objective evolutionary graph algorithm (MEGA) is a new method for *de novo* design of molecules that bind a target. MEGA uses multi-objective optimization to trade off between conflicting objectives (e.g. selectivity of one protein versus another, such as estrogen receptor beta over alpha selectivity) [84]. Similarly, this approach has been used to optimize compounds' antifungal profiles [85]. Multi-objective genetic QSAR uses the Pareto ranking to produce a family of models representing a different compromise in the objectives [86]. Multi-objective optimization has also been used in pharmacophore identification to explore conformational space for multiple ligands simultaneously and align them using a genetic algorithm [87]. Another pharmacophore method uses hierarchical multiple objective ranking, which trades off internal strain, pharmacophoric overlap and steric overlap [88]. Variants of Pareto optimization have also been used in protein design [89] and in docking-based virtual screening [90], docking with EADock [91], fragment-based *de novo* ligand design by multi-objective evolutionary optimization [92], the inverse quantitative structure property relationship (QSPR) problem [93] and evolving interpretable SAR [94].

This is in addition to bioinformatics applications, such as the use of a multi-objective genetic algorithm followed by support vector machine (SVM) used with microarray data to better find clusters of co-expressed genes that were biologically relevant [95]. Multi-objective approaches have also been used for sequence analysis [96], optimization of 2D-GC/MS data in metabolomics [97] and evolutionary search using multiple optimization algorithms [98].

One commercially available set of software tools for performing Pareto optimization of compounds and compound libraries is found in the Accelrys Pipeline Pilot™ and Discovery Studio® programs (http://www.accelrys.com). These can be used to optimize a set of compound libraries to be both maximally drug-like and maximally diverse (Supplementary Fig. S1). The first objective is to maximize the mean value of the 'drug-like' property, which is the prediction of a Bayesian model trained to distinguish drug-like from baseline compounds. The second objective is to maximize structural diversity. As a measure of diversity, we use the number of distinct structural features found within a subset, based on the FCFP_4 molecular fingerprint [99], but any other diversity measure could be used instead. To begin, the optimizer randomly assigns 100 compounds to each of 40 subsets. Then, using the NSGA-II algorithm for Pareto optimization [100], the population of subsets evolves over several hundred generations. The subsets that are most diverse and most drug-like are the ones that tend to survive. Figure 2 shows the progress of the optimization.

The graph displays the population of subsets every 25 generations, with a different color for each generation. The initial random population at the lower left has both low diversity and low drug-like character. As the optimization proceeds, both quantities increase, until the optimized subsets begin to converge at the upper right of the figure. Once convergence has occurred, one can choose one or more subsets considered best from the final optimal population.

Another commercial implementation of Pareto optimization is found in SAS (http://support.sas.com/documentation/cdl/en/orlsoug/59688/HTML/default/ga_sect65.htm). Desirability-based multi-objective optimization is implemented in the commercial packages JMP (http://www.jmp.com), Minitab (http://www.minitab.com), STATISTICA (http://www.statsoft.com) and Stat-Ease (http://www.statease.com). The 'desirability' package [101] for performing desirability-based optimization is available for the open-source R statistics program. Some other tools incorporate Pareto optimization as part of their functionality, including the incremental molecule construction method OptDesign [102] and the pharmacophore method GALAHAD [88,103] (http://www.tripos.com). It is probable that as Pareto optimization increases in popularity, we will see it implemented in more software tools used in drug discovery and data mining.

Pareto Ligand Designer in practice

There have been numerous methods published for *de novo* molecule design, including genetic algorithms that mimic to a great extent Darwinian evolution [104–107] and particle swarm optimization methods [108] with various fitness functions to direct the design of further molecules (e.g. physicochemical properties, docking, similarity and QSAR).

To illustrate how Pareto optimization could be applied in *de novo* molecule design, as a 'proof of concept', a Pipeline Pilot protocol has been constructed at Abbott utilizing the Pareto Sort component to perform simultaneous, multi-objective optimizations of a known, active CCK antagonist, 1 (Fig. 3a), reported by

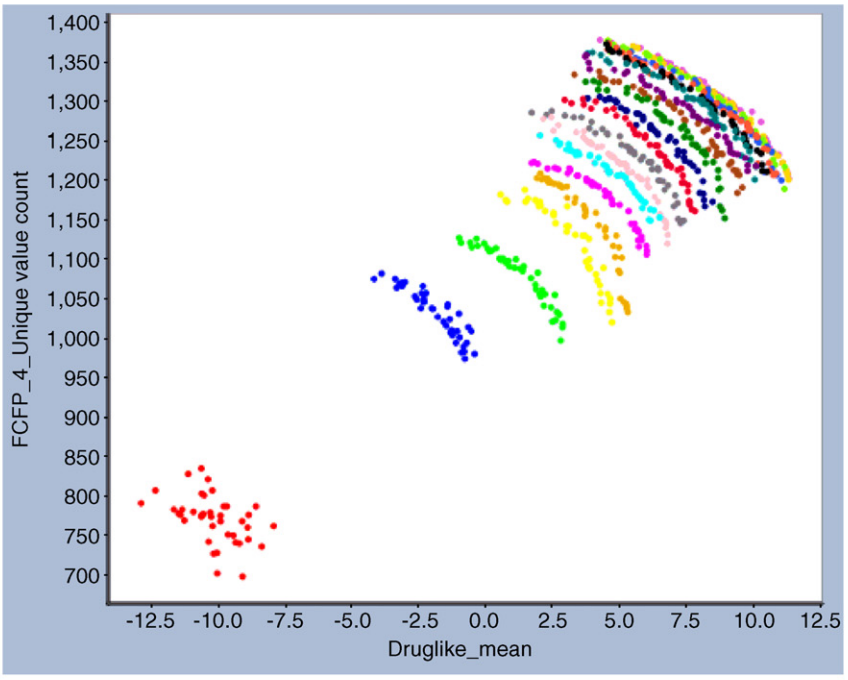

FIGURE 2

A graph showing the progress of Pareto optimization. The graph displays the population of subsets every 25 generations, with a different color for each generation. The initial random population at the lower left has both low diversity and low drug-like character.

Evans *et al.* [109]. Although the CCK antagonist has measured biological activity (IC_{50} = 0.30 μM), the compound is predicted to have poor blood brain barrier (BBB) penetration, poor aqueous solubility, medium CYP2D6 binding probability and a high hepatotoxicity probability using models available in the Pipeline Pilot ADMET component package [110–113]. The Accelrys Pipeline Pilot ADMET BBB model has not yet been published, yet the product notes indicate that the regression model was derived from a training set of 102 compounds and applied to a test set of 86 compounds (RMSE for the training set was 0.36, and root mean square error (RMSE) for the test set was 0.31). The optimizations were performed as three separate 'scenarios' in which the goals were to simultaneously improve the predicted values of two, three and four variables, while maintaining biological activity. Tanimoto similarity calculated using Accelrys ECFP_6 fingerprints calibrated using Belief Theory was used as a surrogate predictor for maintaining biological activity [114].

It should be emphasized that only the initial compound had measured activity, whereas the activities and properties of all molecules generated by the algorithm were not experimentally measured. It is assumed that the various models employed during the optimization had sufficient accuracy and domain applicability to drive the optimization toward sets of molecules with a reasonable probability of having the desired composite set of improved activities and properties; however, one might want to consider in future that measures of domain applicability could be included in the optimization (e.g. a distance from model training set).

In each of the optimization scenarios described next, a Pipeline Pilot protocol was created that performs the following operations:

Initialization. $i = 0$. An initial set of one or more seed molecules are read in. The initial molecules can be assigned as reference molecules for the computation of Tanimoto similarity (Belief Theory). The ligand properties for the set are computed and the values are written to an overall statistics file, which is used to keep track of the progress of the optimization.

Main loop. $i = i + 1$. The ligands are then passed into a Pipeline Pilot Pareto Sort component where optimal compounds along the Pareto front are identified and stored in a file. (The Pareto Sort component performs only the fast nondominated sort from the NSGA-II algorithm [100].) The optimal and nonoptimal compounds are then subjected to an extensive set of molecular transformations, some of which are included in the Drug Guru program [30]. Compounds resulting from the molecular transformations are then passed through several property and structure filters, including 'orange alerts' [29]. Molecules that survive the filters are then assigned to the next iteration number and passed into the Pareto Sort component, statistics are computed, and the cycle begins again (top of main loop).

Optimal Pareto compounds are recycled back into the transformation/optimization selection process. They are not completely removed from the optimization loop. If a previously optimized molecule generates a new molecule with even better properties, the molecule will probably be retained in the next generation of optimized molecules. If the previously optimized molecule generates a new molecule with poorer properties, it will probably be removed because it will not be along the Pareto front. If the previously optimized molecule generates a new molecule that has already been generated (a duplicate), the new molecule will be removed.

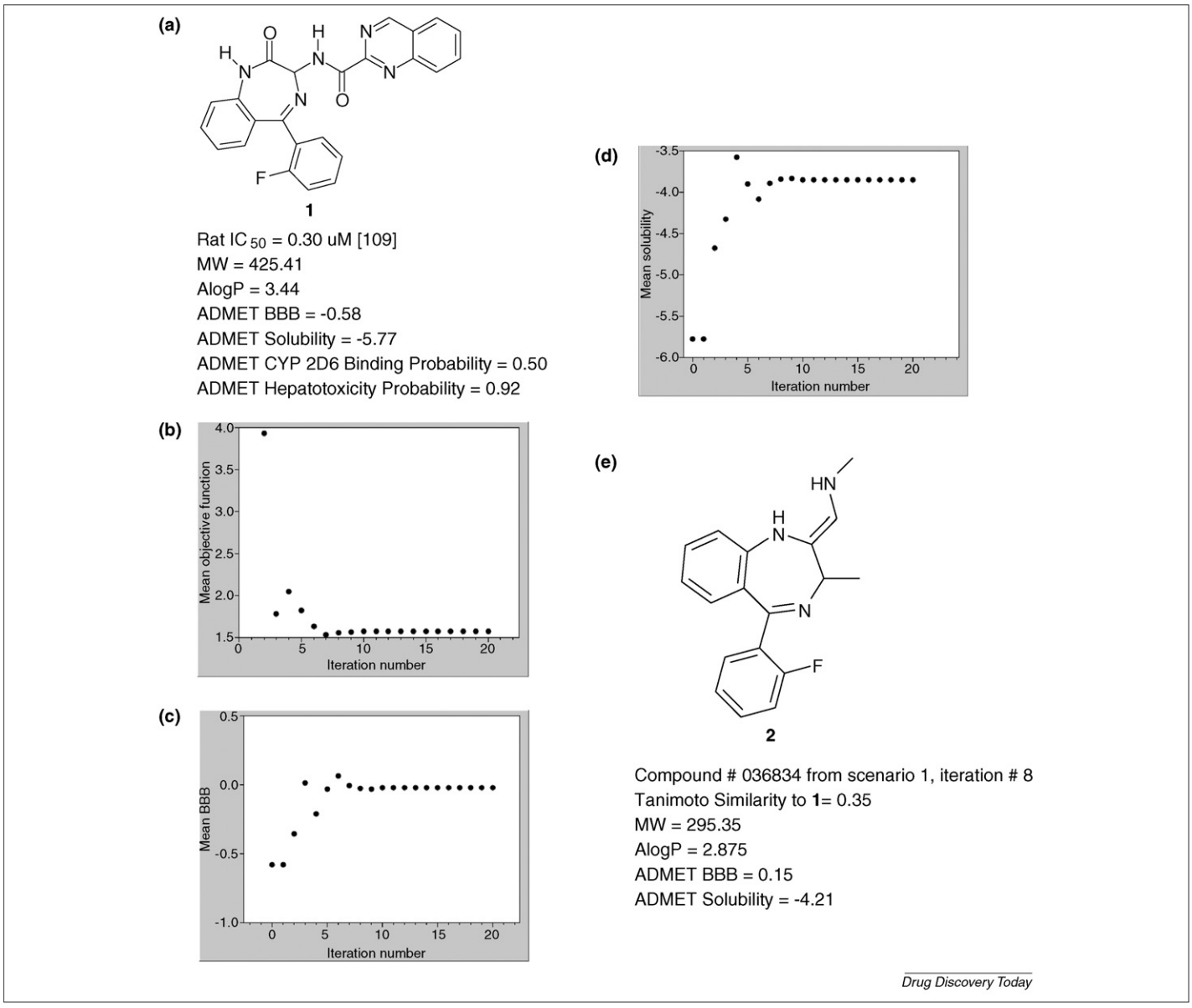

FIGURE 3

(a) Molecule 1 used in scenario 1 and associated properties. **(b)** Iteration number versus the mean value of the objective function for scenario 1. **(c)** Iteration number versus the mean value of BBB for scenario 1. **(d)** Iteration number versus the mean value of solubility for scenario 1. **(e)** Compound 2 from scenario 1, iteration number 8.

Table 1 in Ref. [30] lists ten examples of transformations from Drug Guru. Some of the transformations from Drug Guru have been included in Pareto Ligand Designer (kindly provided by Dr Kent Stewart). It should be noted that some of the Drug Guru transformations implemented in Pareto Ligand Designer (PLD) give identical results to Drug Guru, whereas other transformations give different results. PLD, however, incorporates several hundred additional transformation rules that generate novel molecules beyond the capabilities of Drug Guru. Additional transformations are added as optimized molecules are found in the literature that cannot be generated in a few iterations by PLD.

BBB and solubility optimization

In the first scenario, the optimization goals are to begin with the known, active CCK antagonist, 1, and generate a set of optimized molecules with the following properties: (i) maintain or improve the biological activity, (ii) maintain or decrease the molecular weight, (iii) maintain *A* log *P* within a reasonable range, (iv) improve (increase) the BBB partitioning and (v) improve (increase) the aqueous solubility.

Biological activity was maintained using a minimum ECFP_6 fingerprint Tanimoto similarity filter of 0.35, corresponding to an activity belief of 16.6% [114]. The molecular weight was maintained or decreased using a filter set to 500 Da. Log *P* was maintained in the range of 0.00–5.00 using minimum and maximum *A* log *P* filters. BBB partitioning was calculated using the Accelrys ADMET BBB component. The component calculates the value of $\log_{10}$([brain concentration]/[blood concentration]). Aqueous solubility was calculated using the Accelrys ADMET solubility component. The component calculates the value of $\log_{10}$(molar

solubility). Compounds with Pareto-optimal maximum values of BBB and solubility were saved and written to files at each iteration.

Figure 3b shows the iteration number versus the overall objective function. At each iteration, the Pareto optimizer generates five sets of compounds. The number of Pareto sets is a user-defined parameter. We then compute the mean value of the objective function over all the compounds in all five Pareto sets. Structure (e.g. 'orange alerts' [29]) and property filters are turned on at various iterations during the optimization to guide the algorithm toward the creation of new chemical matter that is likely to be of interest to practicing organic chemists. This can cause an abrupt increase in the objective function, which then – typically – starts to decrease after a few additional iterations. It should be noted that the objective function tends toward zero as the properties become optimized. Clearly, the overall objective function improves dramatically within the first five iterations and then begins to level off at about ten iterations. Figure 3c shows the iteration number versus the mean BBB value. The BBB improves for the first five iterations and then shows no further improvement after about ten iterations. Figure 3d shows the iteration number versus the mean solubility value. The solubility increases and then begins to level off at about ten iterations.

Structure 2 (Fig. 3e) is an example of a compound generated at iteration number 8. The values of the properties, including the Pareto-optimized BBB and solubility, are listed below the structure. For the sake of consistency, an example structure has been taken from iteration number 8 from each optimization scenario. Note that comparisons between scenarios are not entirely valid because the optimization conditions were not identical for each scenario.

BBB, solubility and ADMET CYP2D6 binding optimization

In the second scenario, the optimization goals are to begin with known, active CCK antagonist 1 (Fig. 3a) and generate a set of optimized molecules with the following properties: (i) maintain or improve the biological activity, (ii) maintain or decrease the molecular weight, (iii) maintain *A* log *P* within a reasonable range,

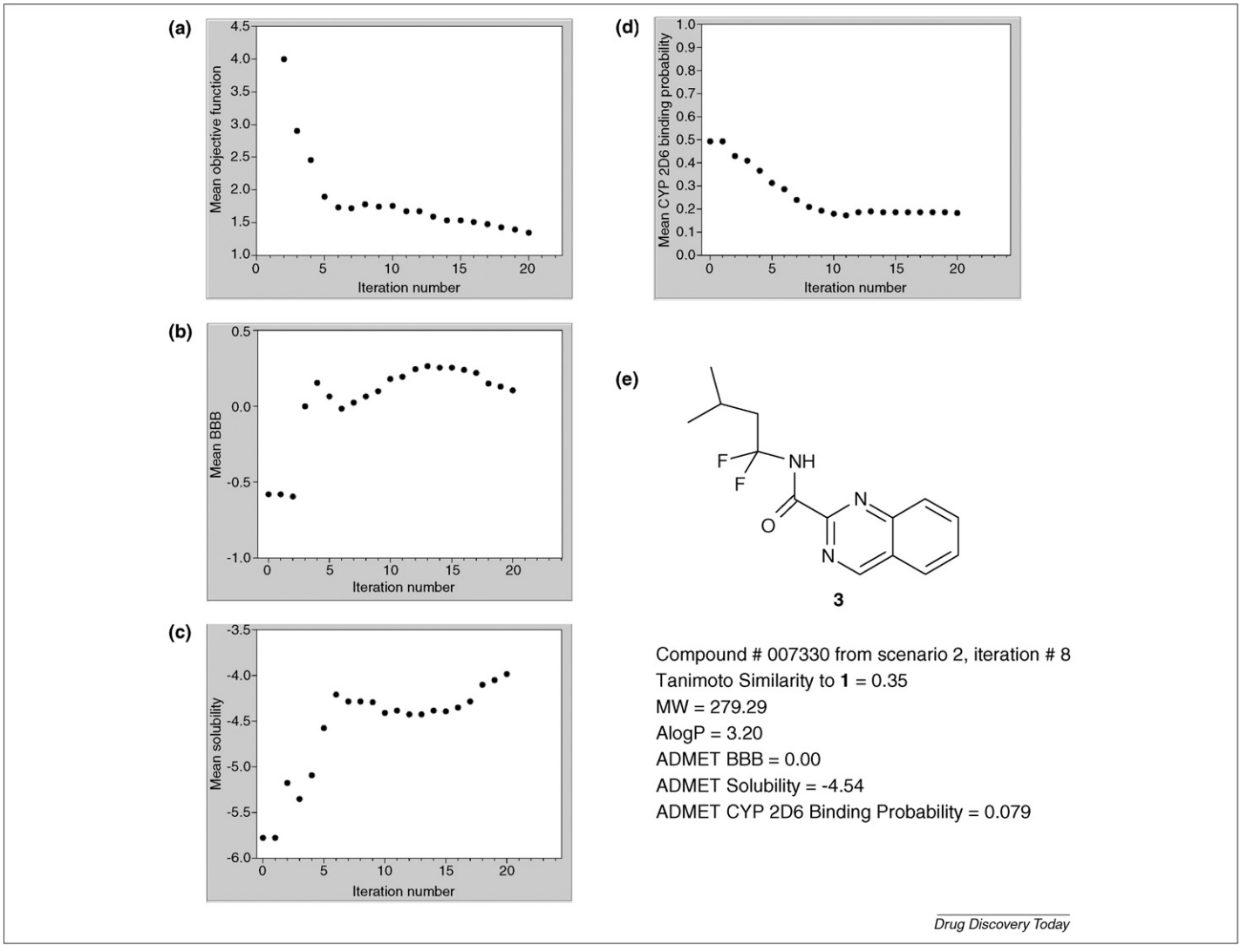

FIGURE 4

(a) Iteration number versus the mean value of the objective function for scenario 2. **(b)** Iteration number versus the mean value of BBB for scenario 2. **(c)** Iteration number versus the mean value solubility for scenario 2. **(d)** Iteration number versus the mean CYP2D6 binding probability for scenario 2. **(e)** Compound 3 from scenario 2, iteration number 8.

(iv) improve (increase) the BBB partitioning, (v) improve (increase) the aqueous solubility and (vi) decrease the CYP2D6 binding probability. The CYP2D6 binding probability was calculated using the CYP2D6 binding model in the Pipeline Pilot ADMET component. The component calculates the probability that a compound will be an inhibitor (probability = 1.0) or not (probability = 0.0). The CYP2D6 binding probability of compound 1 was 0.5, suggesting an intermediate probability of CYP2D6 binding.

Figure 4a shows the iteration number versus the mean value of the objective function. The objective function has sharply decreased after five iterations and continues a more gradual decrease beyond ten iterations, implying slower, gradual improvements in the desired qualities of the Pareto-optimized molecules. Figure 4b shows the iteration number versus the mean value of BBB. The BBB has increased at about five iterations but does not show a consistent improvement afterwards. This suggests that BBB optimization in combination with solubility and reduction of CYP2D6 binding might be difficult to achieve. Figure 4c shows the iteration number versus the mean value of the solubility. There is a sharp increase in solubility near iteration 5, followed by another increase of solubility at a slower rate. Figure 4d shows the CYP2D6 binding probability, which begins near 0.5 and slowly decreases to 0.2 but does not decrease much further. Compound 3 (Fig. 4e) is an example of a Pareto-optimized ligand from iteration number 8.

Current practical issues and challenges of PLD

The technology presented demonstrates that multi-objective optimization of calculated ligand properties is clearly possible, resulting in ligands of reasonable quality based on predicted properties. There are several remaining challenges to be addressed, in addition to actually validating the predictions experimentally.

(i) Development of additional transformation reactions that can potentially get around problems of either slow convergence or poor optimization. This might require continual surveying of the medicinal chemistry literature to ensure that structures that have undergone lead optimization, as well as known optimized structures, can be readily generated by PLD. This might be possible using a test set of known drugs with known ADME/Tox properties, for example.

(ii) As has been suggested by others [108], it will be desirable to replace the chemical transformations with high-yielding 'real' chemical reactions and track the complete synthetic pathway or the least costly reactions. There are several approaches for evaluating synthetic accessibility, including methods that use a combined scoring method incorporating structural complexity, similarity to available starting materials (e.g. SYLVIA [115]) or relative atomic electronegativity and bond parameters (SMCM [116]).

(iii) Conditions for parallel processing will need to be optimized for maximum throughput.

(iv) It will be important to continue to develop and test structure filters to remove compounds that are undesirable to practicing organic chemists for a variety of reasons (e.g. reactivity).

(v) The optimization process relies on the existence of accurate predictive models with a sufficient applicability domain to cover the structures that are generated in the transformation reactions. Hence, there will need to be development and use of accurate predictive models with broad applicability domains or some use of tools for navigating chemical space when the SAR is limited [117].

(vi) We could imagine using approaches to ensure that molecules suggested by PLD look more like endogenous metabolites than commercially available chemicals [118] to bias the physicochemical properties to those that are likely to have improved absorption.

Discussion

For well over 20 years, we have seen the dramatic increase in costs for the discovery and development of new drugs [119,120] – a trend that seems likely to continue. Drug discovery increasingly requires the simultaneous optimization of many measured and calculated properties. As we suggest in this overview (and many others are also showing with their various studies described above), it is possible to achieve such optimizations in principle using methods such as Pareto optimization. One challenge is that as the number of properties to be optimized increases, the required calculation time greatly increases, such that efforts to combine properties into composites (e.g. as desirability functions) [64,65] should be made to decrease the number of properties undergoing Pareto optimization. In the examples we have suggested above, all of the predictions are ADME related, but one could also imagine optimizing other non-ADME related properties such as costs and molecular complexity alongside predicted properties.

Tools like the PLD suggested here, MEGA [84] and other methods [108] represent new approaches to *de novo* design that have the potential to consider many properties (e.g. ADME/Tox) besides bioactivity (or bioactivity at multiple targets), while rapidly generating ideas for potentially synthesizable molecules. This builds on the research in predictive models for ADME/Tox properties [13] and other areas. Ultimately, such approaches might represent an additional source of molecule ideas for lead discovery or, as we have indicated, they might assist in lead optimization and beyond.

Evolving molecules with more ideal properties (using Pareto optimization) alone is not a panacea and might not ensure the survival of companies focused solely on small-molecule drug discovery; however, its use alongside other methods (such as genetic algorithms) could help suggest some nonobvious molecules that have bioactivity balanced with other properties of interest. As a consequence, the development of future software that learns from the molecules it suggests and their properties might represent another example of the 'survival of the fittest'.

Conflict of interest statement

Sean Ekins is a consultant for Collaborations In Chemistry and Collaborative Drug Discovery, Inc. J. Dana Honeycutt is an employee of Accelrys, and James T. Metz is an employee of Abbott.

Acknowledgements

S.E. dedicates this review to the memory of Dr Chad Stoner, a friend and colleague, who published several papers on multidimensional analysis of ADME data. S.E. gratefully acknowledges Maggie A.Z. Hupcey and Peter W. Swaan for earlier discussions over several years and Accelrys for providing Discovery Studio. J.T.M. would like to thank the support staff at Accelrys for

many helpful discussions, suggestions, and corrections to the Pipeline Pilot protocols. The authors would also like to thank the reviewers for their constructive suggestions.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.drudis.2010.04.003.

References

1 Azzaoui, K. *et al.* (2007) Modeling promiscuity based on *in vitro* safety pharmacology profiling data. *ChemMedChem* 2, 874–880
2 Ecker, G.F. (2005) *In silico* screening of promiscuous targets and antitargets. *Chem. Today* 23, 39–42
3 Ekins, S. (2004) Predicting undesirable drug interactions with promiscuous proteins *in silico*. *Drug Discov. Today* 9, 276–285
4 Strachan, R.T. *et al.* (2006) Screening the receptorome: an efficient approach for drug discovery and target validation. *Drug Discov. Today* 11, 708–716
5 O'Connor, K.A. and Roth, B.L. (2005) Finding new tricks for old drugs: an efficient route for public-sector drug discovery. *Nat. Rev. Drug Discov.* 4, 1005–1014
6 Chong, C.R. and Sullivan, D.J., Jr (2007) New uses for old drugs. *Nature* 448, 645–646
7 Egan, W.J. *et al.* (2000) Prediction of drug absorption using multivariate statistics. *J. Med. Chem.* 43, 3867–3877
8 Sun, J.H. *et al.* (2009) Optical design and multiobjective optimization of miniature zoom optics with liquid lens element. *Appl. Opt.* 48, 1741–1757
9 Fu, B. *et al.* (2006) Piezoelectric transducer design via multiobjective optimization. *Ultrasonics* 44 (Suppl. 1), e747–e752
10 Hula, A. *et al.* (2003) Multi-criteria decision-making for optimization of product disassembly under multiple situations. *Environ. Sci. Technol.* 37, 5303–5313
11 Lahanas, M. *et al.* (2003) A hybrid evolutionary algorithm for multi-objective anatomy-based dose optimization in high-dose-rate brachytherapy. *Phys. Med. Biol.* 48, 399–415
12 Ekins, S. *et al.* (2000) Progress in predicting human ADME parameters *in silico*. *J. Pharmacol. Toxicol. Methods* 44, 251–272
13 Ekins, S. *et al.* (2002) Towards a new age of virtual ADME/TOX and multidimensional drug discovery. *J. Comput. Aid. Mol. Des.* 16, 381–401
14 Ekins, S. *et al.* (2000) Present and future *in vitro* approaches for drug metabolism. *J. Pharmacol. Toxicol. Methods* 44, 313–324
15 van De Waterbeemd, H. *et al.* (2001) Property-based design: optimization of drug absorption and pharmacokinetics. *J. Med. Chem.* 44, 1313–1333
16 Balani, S.K. *et al.* (2005) Strategy of utilizing *in vitro* and *in vivo* ADME tools for lead optimization and drug candidate selection. *Curr. Top. Med. Chem.* 5, 1033–1038
17 Oprea, T.I. (2002) Current trends in lead discovery: are we looking for the appropriate properties? *J. Comput. Aid. Mol. Des.* 16, 325–334
18 Oprea, T.I. *et al.* (2001) Is there a difference between leads and drugs? A historical perspective. *J. Chem. Inf. Comput. Sci.* 41, 1308–1315
19 Keseru, G.M. and Makara, G.M. (2009) The influence of lead discovery strategies on the properties of drug candidates. *Nat. Rev. Drug Discov.* 8, 203–212
20 Teague, S.J. *et al.* (1999) The design of leadlike combinatorial libraries. *Angew. Chem. Int. Ed. Engl.* 38, 3743–3748
21 Rishton, G.M. (2008) Molecular diversity in the context of leadlikeness: compound properties that enable effective biochemical screening. *Curr. Opin. Chem. Biol.* 12, 340–351
22 Lipinski, C.A. *et al.* (2001) Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Deliv. Rev.* 46, 3–26
23 Ritchie, T.J. *et al.* (2009) Analysis of the calculated physicochemical properties of respiratory drugs: can we design for inhaled drugs yet? *J. Chem. Inf. Model.* 49, 1025–1032
24 Walters, W.P. and Murcko, M.A. (2002) Prediction of 'drug-likeness'. *Adv. Drug Deliv. Rev.* 54, 255–271
25 Hann, M. *et al.* (1999) Strategic pooling of compounds for high-throughput screening. *J. Chem. Inf. Comput. Sci.* 39, 897–902
26 Pearce, B.C. *et al.* (2006) An empirical process for the design of high-throughput screening deck filters. *J. Chem. Inf. Model.* 46, 1060–1068
27 Huth, J.R. *et al.* (2005) ALARM NMR: a rapid and robust experimental method to detect reactive false positives in biochemical screens. *J. Am. Chem. Soc.* 127, 217–224
28 Huth, J.R. *et al.* (2007) Toxicological evaluation of thiol-reactive compounds identified using a la assay to detect reactive molecules by nuclear magnetic resonance. *Chem. Res. Toxicol.* 20, 1752–1759
29 Metz, J.T. *et al.* (2007) Enhancement of chemical rules for predicting compound reactivity towards protein thiol groups. *J. Comput. Aid. Mol. Des.* 21, 139–144
30 Stewart, K.D. *et al.* (2006) Drug Guru: a computer software program for drug design using medicinal chemistry rules. *Bioorg. Med. Chem.* 14, 7011–7022
31 Wermuth, C.G. (2004) Selective optimization of side activities: another way for drug discovery. *J. Med. Chem.* 47, 1303–1314
32 Wermuth, C.G. (2006) Selective optimization of side activities: the SOSA approach. *Drug Discov. Today* 11, 160–164
33 Kraus, J.M. *et al.* (2009) Rational modification of a candidate cancer drug for use against Chagas disease. *J. Med. Chem.* 52, 1639–1647
34 Hopkins, A.L. *et al.* (2004) Ligand efficiency: a useful metric for lead selection. *Drug Discov. Today* 9, 430–431
35 Abad-Zapatero, C. and Metz, J.T. (2005) Ligand efficiency indices as guideposts for drug discovery. *Drug Discov. Today* 10, 464–469
36 Reynolds, C.H. *et al.* (2008) Ligand binding efficiency: trends, physical basis, and implications. *J. Med. Chem.* 51, 2432–2438
37 Bembenek, S.D. *et al.* (2009) Ligand efficiency and fragment-based drug discovery. *Drug Discov. Today* 14, 278–283
38 Abou-Gharbia, M. (2009) Discovery of innovative small molecule therapeutics. *J. Med. Chem.* 52, 2–9
39 Swaan, P.W. and Ekins, S. (2005) Reengineering the pharmaceutical industry by crash-testing molecules. *Drug Discov. Today* 10, 1191–1200
40 Shimada, J. *et al.* (2002) Integrating computer-based *de novo* drug design and multidimensional filtering for desirable drugs. *Targets* 1, 196–205
41 Delaney, J. (2009) Modelling iterative compound optimisation using a self-avoiding walk. *Drug Discov. Today* 14, 198–207
42 Stoner, C.L. *et al.* (2004) Integrated oral bioavailability projection using *in vitro* screening data as a selection tool in drug discovery. *Int. J. Pharm.* 269, 241–249
43 O'Brien, S.E. and de Groot, M.J. (2005) Greater than the sum of its parts: combining models for useful ADMET prediction. *J. Med. Chem.* 48, 1287–1291
44 Chekmarev, D.S. *et al.* (2008) Shape signatures: new descriptors for predicting cardiotoxicity *in silico*. *Chem. Res. Toxicol.* 21, 1304–1314
45 Kortagere, S. *et al.* (2009) Hybrid scoring and classification approaches to predict human pregane X receptor activators. *Pharm. Res.* 26 (4), 1001–1011
46 Kortagere, S. *et al.* (2008) New predictive models for blood brain barrier permeability of drug-like molecules. *Pharm. Res.* 25, 1836–1845
47 Klon, A.E. *et al.* (2006) Improved naive Bayesian modeling of numerical data for absorption, distribution, metabolism and excretion (ADME) property prediction. *J. Chem. Inf. Model.* 46, 1945–1956
48 Obrezanova, O. *et al.* (2007) Gaussian processes: a method for automatic QSAR modeling of ADME properties. *J. Chem. Inf. Model.* 47, 1847–1857
49 Zhang, L. *et al.* (2008) QSAR modeling of the blood–brain barrier permeability for diverse organic compounds. *Pharm. Res.* 25, 1902–1914
50 Trunzer, M. *et al.* (2009) Metabolic soft spot identification and compound optimization in early discovery phases using MetaSite and LC–MS/MS validation. *J. Med. Chem.* 52, 329–335
51 Boyer, S. *et al.* (2007) Reaction site mapping of xenobiotic biotransformations. *J. Chem. Inf. Model.* 47, 583–590
52 Boyer, S. and Zamora, I. (2002) New methods in predictive metabolism. *J. Comput. Aid. Mol. Des.* 16, 403–413
53 Darvas, F. *et al.* (2000) Diversity measures for enhancing ADME admissibility of combinatorial libraries. *J. Chem. Inf. Comput. Sci.* 40, 314–322
54 Darvas, F. *et al.* (1999) MetabolExpert: its use in metabolism research and in combinatorial chemistry. In *Drug Metabolism: Databases and High-throughput Testing During Drug Design and Development*. (Erhardt, P.W., ed.), International Union of Pure and Applied Chemistry and Blackwell Science
55 Embrechts, M.J. and Ekins, S. (2007) Classification of metabolites with kernel-partial least squares (K-PLS). *Drug Metab. Dispos.* 35, 325–327
56 Stranz, D.D. *et al.* (2008) Combined computational metabolite prediction and automated structure-based analysis of mass spectrometric data. *Toxicol. Mech. Methods* 18, 1–8
57 Yamashita, F. *et al.* (2008) 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.* 48, 364–369
58 Chadwick, A. *et al.* (2006) Improving the pharmaceutical R & D process: how simulation can support management decision making. In *Computer Applications in Pharmaceutical Research and Development* (Ekins, S., ed.), pp. 247–273, John Wiley & Sons

59 Wunberg, T. *et al.* (2006) Improving the hit-to-lead process: data-driven assessment of drug-like and lead-like screening hits. *Drug Discov. Today* 11, 175–180
60 Gleeson, M.P. (2008) Generation of a set of simple, interpretable ADMET rules of thumb. *J. Med. Chem.* 51, 817–834
61 Derringer, G. and Suich, R. (1980) Simultaneous optimization of several response variables. *J. Qual. Technol.* 12, 214–219
62 Rassokhin, D.N. and Agrafiotis, D.K. (2000) Kolmogorov–Smirnov statistic and its application in library design. *J. Mol. Graph. Model.* 18, 368–382
63 Derringer, G.C. (1994) A balancing act: optimizing a product's properties. *Qual. Prog.* 27, 51–58
64 Govaerts, B. and Le Bailly de Tilleghem, C. (2005) *Distribution of Desirability Index in Multicriteria Optimization using Desirability Functions based on the Cumulative Distribution Function of the Standard Normal.* http://www.stat.ucl.ac.be/ISpub/dp/2005/dp0531.pdf
65 Govaerts, B. and Le Bailly de Tilleghem, C. (2005) *Uncertainty Propagation in Multiresponse Optimization using a Desirability Index.* http://www.stat.ucl.ac.be/ISpub/dp/2005/dp0532.pdf
66 Handl, J. *et al.* (2007) Multiobjective optimization in bioinformatics and computational biology. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 4, 279–292
67 Andersson, J. (2000) *A Survey of Multiobjective Optimization in Engineering Design.* Dept. of Mechanical Engineering, Linköping University
68 Fonseca, C.M. and Fleming, P.J. (1995) An overview of evolutionary algorithms in multiobjective optimization. *Evol. Comput.* 3, 1
69 Ortiz, M.C. *et al.* (2006) Vectorial optimization as a methodological alternative to desirability function. *Chemom. Intell. Lab. Sys.* 83, 157–168
70 Geoffrion, A.M. (1968) Proper efficiency and the theory of vector optimization. *J. Math. Anal. Appl.* 22, 618–630
71 Van Veldhuizen, D.A. and Lamont, G.B. (2000) Multiobjective evolutionary algorithms: analyzing the state-of-the-art. *Evol. Comput.* 8, 125–147
72 Agrafiotis, D.K. (2002) Multiobjective optimization of combinatorial libraries. *J. Comput. Aid. Mol. Des.* 16, 335–356
73 Wright, T. *et al.* (2003) Optimizing the size and configuration of combinatorial libraries. *J. Chem. Inf. Comput. Sci.* 43, 381–390
74 Chen, H. *et al.* (2009) ProSAR: a new methodology for combinatorial library design. *J. Chem. Inf. Model.* 49, 603–614
75 Le Bailly de Tilleghem, C. *et al.* (2005) A fast exchange algorithm for designing focused libraries in lead optimization. *J. Chem. Inf. Model.* 45, 758–767
76 Fischer, J.R. *et al.* (2010) LoFT: similarity-driven multiobjective focused library design. *J. Chem. Inf. Model.* 50, 1–21
77 Gillet, V.J. (2008) New directions in library design and analysis. *Curr. Opin. Chem. Biol.* 12, 372–378
78 Gillet, V.J. *et al.* (2002) Combinatorial library design using a multiobjective genetic algorithm. *J. Chem. Inf. Comput. Sci.* 42, 375–385
79 Chen, G. *et al.* (2005) Focused combinatorial library design based on structural diversity, druglikeness and binding affinity score. *J. Comb. Chem.* 7, 398–406
80 Kruisselbrink, J.W. *et al.* (2009) Combining aggregation with Pareto optimization: a case study in evolutionary molecular design. In *Proceedings of EMO '09, Lect. Notes Comput. Sci.*. Springer, Berlin/Heidelberg, Vol. 5467/2009 pp. 453–467
81 Nicolaou, C.A. *et al.* (2007) Molecular optimization using computational multi-objective methods. *Curr. Opin. Drug Discov. Dev.* 10, 316–324
82 Cruz-Monteagudo, M. *et al.* (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
83 Cruz-Monteagudo, M. *et al.* (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
84 Nicolaou, C.A. *et al.* (2009) *De novo* drug design using multiobjective evolutionary graphs. *J. Chem. Inf. Model* 49, 295–307
85 Saiz-Urra, L. *et al.* (2009) Global antifungal profile optimization of chlorophenyl derivatives against *Botrytis cinerea* and *Colletotrichum gloeosporioides*. *J. Agric. Food Chem.* 57, 4838–4843
86 Nicolotti, O. *et al.* (2002) Multiobjective optimization in quantitative structure-activity relationships: deriving accurate and interpretable QSARs. *J. Med. Chem.* 45, 5069–5080
87 Cottrell, S.J. *et al.* (2006) Incorporating partial matches within multi-objective pharmacophore identification. *J. Comput. Aid. Mol. Des.* 20, 735–749
88 Clark, R.D. and Abrahamian, E. (2008) Using a staged multi-objective optimization approach to find selective pharmacophore models. *J. Comput. Aid. Mol. Des.* 23, 765–771
89 Suarez, M. *et al.* (2008) Pareto optimization in computational protein design with multiple objectives. *J. Comput. Chem.* 29, 2704–2711
90 Li, H. *et al.* (2009) An effective docking strategy for virtual screening based on multi-objective optimization algorithm. *BMC Bioinformatics* 10, 58
91 Grosdidier, A. *et al.* (2007) EADock: docking of small molecules into protein active sites with a multiobjective evolutionary optimization. *Proteins* 67, 1010–1025
92 Dey, F. and Caflisch, A. (2008) Fragment-based *de novo* ligand design by multiobjective evolutionary optimization. *J. Chem. Inf. Model.* 48, 679–690
93 Brown, N. *et al.* (2006) A novel workflow for the inverse QSPR problem using multiobjective optimization. *J. Comput. Aid. Mol. Des.* 20, 333–341
94 Birchall, K. *et al.* (2008) Evolving interpretable structure-activity relationship models. 2. Using multiobjective optimization to derive multiple models. *J. Chem. Inf. Model.* 48, 1558–1570
95 Maulik, U. *et al.* (2009) Combining Pareto-optimal clusters using supervised learning for identifying co-expressed genes. *BMC Bioinformatics* 10, 27
96 Zwir, I. *et al.* (2002) Automated biological sequence description by genetic multiobjective generalized clustering. *Ann. N. Y. Acad. Sci.* 980, 65–82
97 O'Hagan, S. *et al.* (2007) Closed-loop, multiobjective optimization of two-dimensional gas chromatography/mass spectrometry for serum metabolomics. *Anal. Chem.* 79, 464–476
98 Vrugt, J.A. and Robinson, B.A. (2007) Improved evolutionary optimization from genetically adaptive multimethod search. *Proc. Natl. Acad. Sci. U. S. A.* 104, 708–711
99 Rogers, D. *et al.* (2005) Using extended-connectivity fingerprints with Laplacian-modified Bayesian analysis in high-throughput screening follow-up. *J. Biomol. Screen.* 10, 682–686
100 Deb, K. *et al.* (2000) *A Fast Elitist Non-Dominated Sorting Genetic Algorithm for Multi-Objective Optimization: NSGA-II.* Indian Institute of Technology
101 Kuhn, M. (2008) Desirability function optimization and ranking. http://cran.r-project.org/web/packages/desirability/vignettes/desirability.pdf.
102 Soltanshahi, F. *et al.* (2006) Balancing focused combinatorial libraries based on multiple GPCR ligands. *J. Comput. Aid. Mol. Des.* 20, 529–538
103 Richmond, N.J. *et al.* (2006) GALAHAD. 1. Pharmacophore identification by hypermolecular alignment of ligands in 3D. *J. Comput. Aid. Mol. Des.* 20, 567–587
104 Douguet, D. *et al.* (2000) A genetic algorithm for the automated generation of small organic molecules: drug design using an evolutionary algorithm. *J. Comput. Aid. Mol. Des.* 14, 449–466
105 Pegg, S.C-H. *et al.* (2001) A genetic algorithm for structure-based *de novo* design. *J. Comput. Aid. Mol. Des.* 15, 911–933
106 Lameijer, E.W. *et al.* (2006) The molecule evoluator. An interactive evolutionary algorithm for the design of drug-like molecules. *J. Chem. Inf. Model.* 46, 545–552
107 Lameijer, E.W. *et al.* (2005) Evolutionary algorithms in drug design. *Nat. Comput.* 4, 177–243
108 Schneider, G. *et al.* (2009) Voyages to the (un)known: adaptive design of bioactive compounds. *Trends Biotechnol.* 27, 18–26
109 Evans, B.E. *et al.* (1988) Methods for drug discovery: development of potent, selective, orally effective cholecystokinin antagonists. *J. Med. Chem.* 31, 2235–2246
110 Cheng, A. and Merz, K.M., Jr (2003) Prediction of aqueous solubility of a diverse set of compounds using quantitative structure–property relationships. *J. Med. Chem.* 46, 3572–3580
111 Susnow, R.G. and Dixon, S.L. (2003) Use of robust classification techniques for the prediction of human cytochrome P450 2D6 inhibition. *J. Chem. Inf. Comput. Sci.* 43, 1308–1315
112 Dixon, S.L. and Villar, H.O. (1998) Bioactive diversity and screening library selection via affinity fingerprinting. *J. Chem. Inf. Comput. Sci.* 38, 1192–1203
113 Cheng, A. and Dixon, S.L. (2003) In silico models for the prediction of dose-dependent human hepatotoxicity. *J. Comput. Aid. Mol. Des.* 17, 811–823
114 Muchmore, S.W. *et al.* (2008) Application of belief theory to similarity data fusion for use in analog searching and lead hopping. *J. Chem. Inf. Model.* 48, 941–948
115 Boda, K. *et al.* (2007) Structure and reaction based evaluation of synthetic accessibility. *J. Comput. Aid. Mol. Des.* 21, 311–325
116 Allu, T.K. and Oprea, T.I. (2005) Rapid evaluation of synthetic and molecular complexity for *in silico* chemistry. *J. Chem. Inf. Model.* 45, 1237–1243
117 Bajorath, J. *et al.* (2009) Navigating structure–activity landscapes. *Drug Discov. Today* 14, 698–705
118 Dobson, P.D. *et al.* (2009) 'Metabolite-likeness' as a criterion in the design and selection of pharmaceutical drug libraries. *Drug Discov. Today* 14, 31–40
119 Dimasi, J.A. (2001) New drug development in the United States from 1963 to 1999. *Clin. Pharmacol. Ther.* 69, 286–296
120 DiMasi, J.A. *et al.* (2003) The price of innovation: new estimates of drug development costs. *J. Health Econ.* 22, 151–185