COMBINATORIAL INFORMATICS IN THE POST-GENOMICS ERA

Dimitris K. Agrafiotis, Victor S. Lobanov and F. Raymond Salemme

The multitude of potential drug targets emerging from genome sequencing demands new approaches to drug discovery. A chemogenomics strategy, which involves the generation of small-molecule compounds that can be used both as tools to probe biological mechanisms and as leads for drug-property optimization, provides a highly parallel, industrialized solution. Key to the success of this strategy is an integrated suite of chemi-informatics applications that can allow the rapid and directed optimization of chemical compounds with drug-like properties using 'just-in-time' combinatorial chemical synthesis. An effective embodiment of this process requires new computational and data-mining tools that cover all aspects of library generation, compound selection and experimental design, and work effectively on a massive scale.

PROBE LIBRARY
A collection of diverse compounds that is aimed at discovering hits across a wide variety of biological targets.

3-Dimensional Pharmaceuticals, Inc., 665 Stockton Drive, Exton, Pennsylvania 19341, USA. Correspondence to D.K.A. e-mail: agrafiotis@3dp.com
DOI: 10.1038/nrd791

Genomic and proteomic approaches to the identification of new targets for drug intervention present unprecedented opportunities for the discovery of new agents with novel therapeutic modes of action[1]. Nevertheless, some daunting difficulties and risks will need to be overcome to realize this potential. Historically, the proteins that the pharmaceutical industry has targeted for drug discovery have generally been well understood from a mechanistic and biological standpoint. By contrast, relatively little or nothing might be known about the mechanism or biological function of a 'genomics' target, which might be rendered interesting in the first instance simply by virtue of its appearance in a disease context or apparent effects in a shotgun gene-knockout experiment. So, genomics-based target discovery is typically followed by a laborious process of target validation, which generally produces useful, although often ambiguous, information about the potential therapeutic relevance of the target.

An alternative, 'chemogenomics' approach to target validation uses the basic information that is provided by the target sequence to make the protein and subsequently discover a small-molecule 'tool compound' that interacts with that target. The tool compound can in turn be evaluated in a biological disease model to directly test a therapeutic hypothesis. This approach can be implemented as a highly parallel process, and is particularly well suited to the discovery of drugs in broad families, in which inter-target specificity might be a crucial factor in the ultimate development of therapeutic agents with minimal side effects. Although this chemistry-orientated approach does not eliminate the need for biological target validation, it defers the required investment to a later stage in the discovery cycle, when these resources can be deployed more efficiently and with a higher probability of success. In this article, we describe a chemogenomics strategy for drug discovery, and overview the key role of chemi-informatics in this approach.

A chemogenomics strategy for drug discovery

A practical and cost-effective embodiment of a chemogenomics strategy for isolated molecular targets is outlined in FIG. 1 (REFS 2–5). Gene sequences for targets that have been identified by genomics approaches (FIG. 1a) are cloned and expressed as target proteins (FIG. 1b) that are suitable for screening with a PROBE LIBRARY of small, drug-like chemical compounds (FIG. 1c). These compounds are screened to find active hits using a quantitative, universal binding assay (FIG. 1d) that has a wide dynamic range and does not require the development of custom protocols or reagents. Initial hits or quantitative structure–activity data that emerge from the binding assay are analysed (FIG. 1e), and are used to formulate

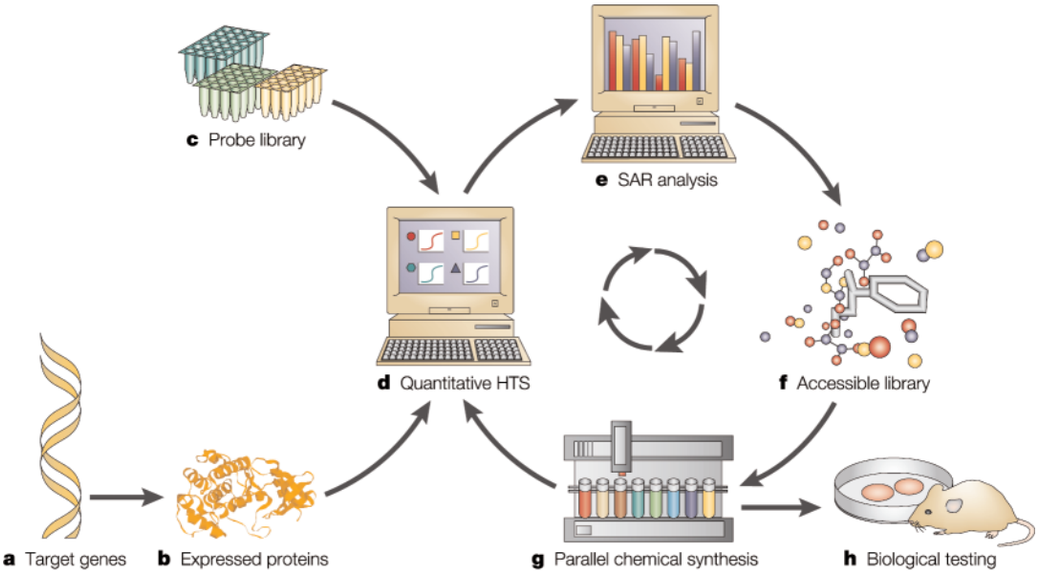

Figure 1 | **A practical and cost-effective embodiment of a chemogenomics strategy. a**,**b** | Gene sequences for targets that have been identified by genomics approaches are cloned and expressed as target proteins that are suitable for screening with **c** | a probe library of small, drug-like chemical compounds. **d** | These compounds are screened to find active hits using a quantitative, universal binding assay. **e** | Initial hits or quantitative structure–activity data that emerge from the binding assay are analysed and used to formulate a selection strategy for the synthesis of further compounds with improved properties. **f** | These compounds are selected from a computer database of synthetically accessible analogues of the initial probe library, **g** | synthesized by parallel-synthesis methods, and **d** | tested to elaborate the structure–activity profile of the target under investigation and to refine the selection criteria for further rounds of chemical synthesis and biological testing. In each iteration, priority is assigned to the synthetic candidates using a multiobjective optimization process that is designed to ensure that compounds are not only optimized for target binding affinity, but also have drug-like characteristics that will allow them to **h** | be used directly as tool compounds in appropriate cellular or biological model systems. HTS, high-throughput screening; SAR, structure–activity relationship.

a selection strategy for the synthesis of further compounds with improved properties. These compounds are selected from a computer database of synthetically accessible analogues of the initial probe library (FIG. 1f), which is constructed using verified synthetic protocols and is characterized by an extensive set of computed, drug-related molecular properties. The selected compounds are synthesized by parallel-synthesis methods (FIG. 1g), and are subsequently tested (FIG. 1d) to elaborate the structure–activity profile of the target under investigation, and to refine the selection criteria for further rounds of chemical synthesis and biological testing. In each iteration, priority is assigned to the synthetic candidates using a MULTIOBJECTIVE OPTIMIZATION process that is designed to ensure that compounds are not only optimized for target binding affinity, but also have drug-like characteristics that will allow them to be used directly as tool compounds in appropriate cellular or biological model systems (FIG. 1h).

To implement this process in an efficient and practical manner, several novel technological components must be developed and optimized, including: first, a quantitative, universal assay system; and second, a comprehensive chemi-informatics platform to provide process control and decision support for lead discovery and optimization. A powerful universal assay system that simplifies initial target characterization (for example, for proper three-dimensional folding) and the development of screening assays was recently described[6]. This approach uses fluorescence-based detection of protein thermal stability, both as a means to ensure integrity of protein folding, and as a means to screen hits and optimize lead compounds. The method approximates measurements that are made using differential scanning calorimetry (DSC), but unlike DSC, it has been developed in a highly miniaturized format that is suitable for rapid, high-throughput screening of large chemical libraries. It is based on the biophysical observation that a ligand that is bound to a protein stabilizes the protein native state by an amount that is proportional to its binding affinity, and, consequently, causes the protein to melt at a higher temperature. Because the method measures intrinsic binding affinity, it can be used with virtually any soluble protein or receptor. The observable range of accessible melting temperatures allows detection of compounds that bind in the 10 μM to nM range in a single measurement, making the method suitable for the assessment of target 'drugability', lead identification and lead optimization.

Of course, a potent, non-toxic and bioavailable lead compound might not emerge directly from the existing probe libraries and/or their virtual analogues. Depending on the chemistry involved, directed libraries typically exhaust readily available reagents within a few iterations, and — for the most promising cases — are followed up by second- and third-generation iterable custom libraries that are specifically tailored to the target under investigation.

MULTIOBJECTIVE OPTIMIZATION
The solution to a problem that involves the simultaneous optimization of multiple design objectives.

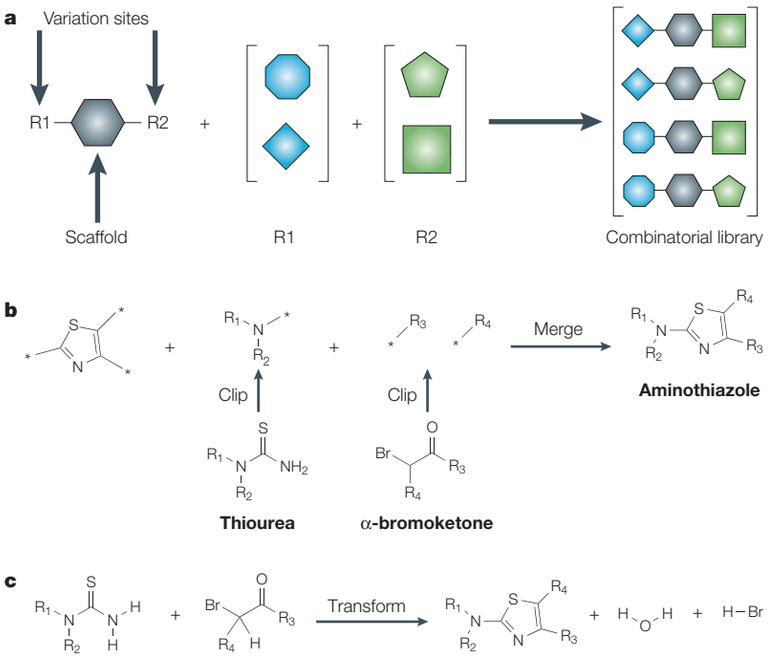

Figure 2 | **Virtual-library generation.** Two approaches are commonly used for generating virtual libraries. **a** | The first is based on the use of a Markush structure, which represents a common scaffold with variation sites labelled as R-groups, each of which is associated with a list of alternatives. **b** | The virtual library is assembled by systematic attachment of clipped reagents to the respective variation sites of the core scaffold, shown here using the example of the synthesis of aminothiazoles from thioureas and α-bromoketones. Although enumeration is reduced to simple concatenation of the corresponding connection tables, the lists of clipped reagents must be carefully constructed from the monomers by removing the parts of the structure that are discarded during the reaction. **c** | A more flexible approach, again illustrated by using the synthesis of aminothiazoles, encodes a reaction as a chemical transform, which specifies the parts of the reacting molecules that undergo chemical transformations and the nature of these transformations. Adapted from REF. 13.

VIRTUAL LIBRARY
A computer representation of a collection of chemical compounds.

COMBINATORIAL LIBRARY
A collection of compounds that are derived from the systematic application of a synthetic sequence on a prescribed set of building blocks.

ENUMERATION
The process of constructing the connection tables of the combinatorial products from their respective building blocks, as prescribed by the reaction sequence.

CLIPPED REAGENT
The (potentially modified) part of a reagent that becomes part of the final product.

CONNECTION TABLE
A computer representation of the atoms and bonds that comprise a molecule. This is the computer equivalent of a chemical sketch of a molecule.

The role of chemi-informatics

The potential for improved performance using the strategy outlined in FIG. 1 lies in the ability to rapidly follow up on initial hits through intelligent selection of related compounds from a computer database (or VIRTUAL LIBRARY) of synthetically accessible analogues with predefined synthesis recipes and predicted property profiles. This approach introduces a high level of parallelism and process automation at all stages of the design, synthesis, quality control and testing of compounds. Experience indicates that, to address the full spectrum of targets emerging from genomics-based target-identification efforts, it will be necessary to physically screen probe libraries that span a wide range of chemotypes and contain hundreds of thousands of compounds. These libraries will derive from synthetic strategies that could, in theory, produce billions of related analogues, which far exceeds the capabilities of conventional chemical-database management systems and data-modelling tools.

This raises several important questions. How can huge combinatorial libraries be generated, represented, accessed, searched and manipulated? What are the most appropriate chemical-property spaces, and how can they best be computed, sampled, visualized and validated? And what are the most effective ways to design, execute and analyse a combinatorial-chemistry experiment? Fortunately, combinatorial libraries are not random collections of molecules, but have a highly constrained intrinsic structure. An effective chemi-informatics system must capitalize on this structure and defer any form of computation until it is absolutely necessary. The remaining sections address the questions above — providing a general overview of the current state-of-the-art in computer-assisted library design — and describe some recent advances that allow rapid analysis of combinatorial compounds without necessitating their virtual synthesis. More detailed reviews can be found elsewhere[7–12].

Virtual-library generation

The construction of a virtual COMBINATORIAL LIBRARY involves three basic steps: reaction encoding, selection of reagents and ENUMERATION. Two approaches are commonly used (FIG. 2). The first is based on the use of a Markush structure, which represents a common scaffold with variation sites labelled as R-groups, each of which is associated with a list of alternatives[13]. In this case, the virtual library is assembled by systematic attachment of CLIPPED REAGENTS to the respective variation sites of the core scaffold. Although enumeration is reduced to simple concatenation of the corresponding CONNECTION TABLES, the lists of clipped reagents must be carefully constructed from the monomers by removing the parts of the structure that are discarded during the reaction. Moreover, this 'fragment-making' approach is not suitable for reactions that cause modification of the building blocks, such as the Diels–Alder reaction, or oligomeric libraries for which the core scaffold is poorly defined (for example, peptide and peptoid libraries).

A more flexible approach encodes a reaction as a chemical 'transform'[14,15]. The transform specifies the parts of the reacting molecules that undergo chemical transformations and the nature of these transformations. This approach mimics more closely the steps that are involved in actual synthesis, does not require a common template or the generation of clipped reagents, and can be applied to a broad spectrum of chemical reactions. It is important that the reaction language accommodates multicomponent reactions, ring cyclizations, protecting-group removal and core-structure modification, and offers the ability to specify multiple products, designate stereochemistry and differentiate functional-group reactivity if a reagent has more than one potential reactive site[15].

In order to be scalable, virtual-library generation should avoid explicit enumeration and storage of every product unless it is specifically requested, and should be able to access any desired structure in a rapid time frame. This approach is often referred to as 'LAZY' (or implicit) ENUMERATION. Different implementations vary greatly in their storage requirements and enumeration speeds. Performance can be enhanced through careful algorithmic design — for example, by compiling product-assembly instructions into machine code, circumventing explicit generation of intermediates, pre-storing key MOLECULAR-PERCEPTION flags to be used for subsequent

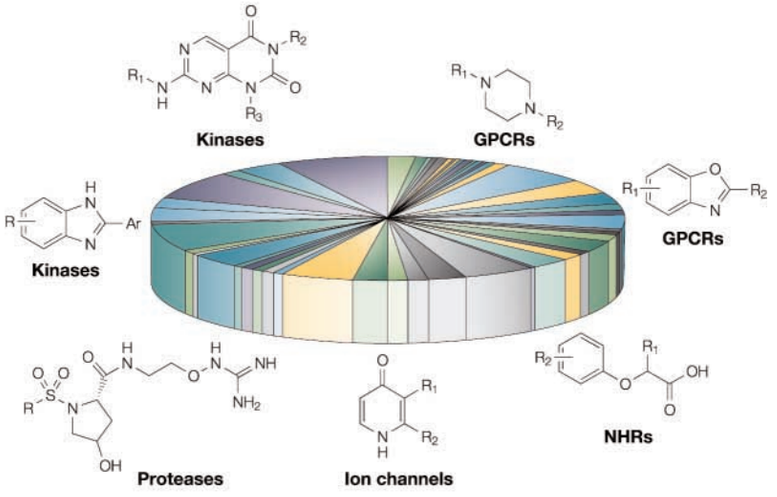

Figure 3 | **Representative chemical classes in the 3DP probe library.** Representative chemical scaffolds that comprise the probe library of 3-Dimensional Pharmaceuticals (3DP). The library contains many pharmacophoric and target-family motifs, including G-protein-coupled receptors (GPCRs), proteases, nuclear hormone receptors (NHRs), ion channels and so on.

computational tasks and so on. The most advanced implementations can reach enumeration speeds in excess of 20,000 products per second, and can compress a one-billion-member library into a 1–5-MB data stream in just a few seconds on a modern PC[15].

Although synthetic accessibility cannot be guaranteed for every possible combination of reactants, careful reagent filtering can minimize the probability of generating compounds that cannot be synthesized or are unstable, and can considerably reduce the size of the virtual library. Typical filters include substructural screens that check for the presence of interfering functional groups or other non-drug-like structural patterns (see below)[14,16]. Reagent filtering is computationally efficient, and is typically followed by visual inspection and, if necessary, further manual pruning.

Appropriate chemical spaces

Molecular descriptors. Screening-library design has traditionally been guided by MOLECULAR DIVERSITY[17], which basically represents a generalization of the concept of molecular similarity from individuals to collections. Molecular similarity is typically quantified by a numerical index that is derived either by direct computation, or by the measurement of a set of characteristic features (descriptors) that are subsequently combined using a dissimilarity or distance measure[18]. Three types of molecular descriptor are in common use: one-dimensional descriptors, which encode chemical composition; two-dimensional descriptors, which encode chemical topology; and three-dimensional descriptors, which encode three-dimensional shape and functionality[19].

The first two categories are computed directly from the connection table, and combine elements of GRAPH THEORY with some form of chemical knowledge. One-dimensional and two-dimensional descriptors can be further classified into: descriptors that capture the constitutional, branching and ring character of a molecule, and focus predominantly on topology[20] (for example, molecular-connectivity indices, which include size, saturation, hetero-atom content, topological shape and symmetry, and counts of characteristic subgraphs, such as rings, paths, clusters and so on); descriptors that represent abstract patterns that are discovered by systematic traversal of the molecular graph (for example, hashed FINGERPRINTS[21]); descriptors that represent occurrences of specific atom types or functional groups that are chemically or biologically important (for example, substructure keys, atom and fragment counts[22], and topological PHARMACOPHORES[23]); descriptors that incorporate atomic properties that are relevant in ligand binding and DRUG LIKENESS, such as volume, hybridization, partial atomic charge, electronegativity, polarizability, hydrophobicity and hydrogen-bonding potential (for example, atom pairs[24], topological torsions[25,26], autocorrelation functions[27,28] and approximate surface-area descriptors[29]); and global physicochemical properties, such as molecular weight, LOG P, molar refractivity and so on[30]. The numerical representation can be a vector of binary, integer or real numbers, and the underlying mathematics range from simple counting schemes[22] to complex matrix-diagonalization routines[31]. Once the space is established, molecular similarity is defined using a distance function that is appropriate for the underlying data representation (for example, Euclidean distance in real space, Tanimoto coefficient in binary space and so on). One-dimensional and two-dimensional descriptors can usually be computed very quickly, and have a successful history in similarity searching and structure–activity correlation. This success is often attributed to the fact that they seem to provide the right level of chemical 'resolution' — they are general enough to relate compounds in diverse chemical classes, but specific enough to distinguish between closely related analogues.

Three-dimensional descriptors attempt to capture three-dimensional shape and functionality, which have an important role in the recognition of drugs by macromolecules. Examples of such descriptors include geometric atom pairs and topological torsions[32], spatial autocorrelation vectors[33], WHIM indices[34], molecular hashkeys[35], BCUTs[36] and pharmacophore fingerprints[37–40]. The pharmacophore is typically represented as a set of three or four pharmacophore centres that form a triangle or tetrahedron. These centres include macromolecular recognition sites, such as charged centres, hydrogen-bond donors and acceptors, hydrophobic centres and aromatic-ring centres. To generate the key, the pharmacophores (that is, the pharmacophoric centres and their respective distances) that are exposed by a particular conformation are mapped onto specific bits in a bitmap (or fingerprint). Individual fingerprints can be combined into 'molecular fingerprints', which represent the union across all conformations of a particular molecule, and 'library fingerprints', which represent the union across all molecules in a library.

Three-dimensional descriptors can be derived from a single or multiple conformers, the latter being generally more realistic in representing the similarity or diversity

LAZY ENUMERATION
The on-demand virtual synthesis of combinatorial products.

MOLECULAR PERCEPTION
The computational detection of important structural features, such as rings, aromaticity, stereochemistry and topological symmetry, from the molecule's connection table.

MOLECULAR DIVERSITY
The chemical-information content of a collection of compounds. The concept is often context dependent.

GRAPH THEORY
Formally, a connection table for a molecule records its chemical structure as a graph — a set of vertices (the atoms) linked by edges (the bonds). This allows mathematical analyses to be used to classify the structure or calculate molecular properties.

FINGERPRINT
A set of binary numbers (1s and 0s) that are used to characterize a molecular structure. Each bit signifies the presence (1) or absence (0) of one or more structural features in the target molecule.

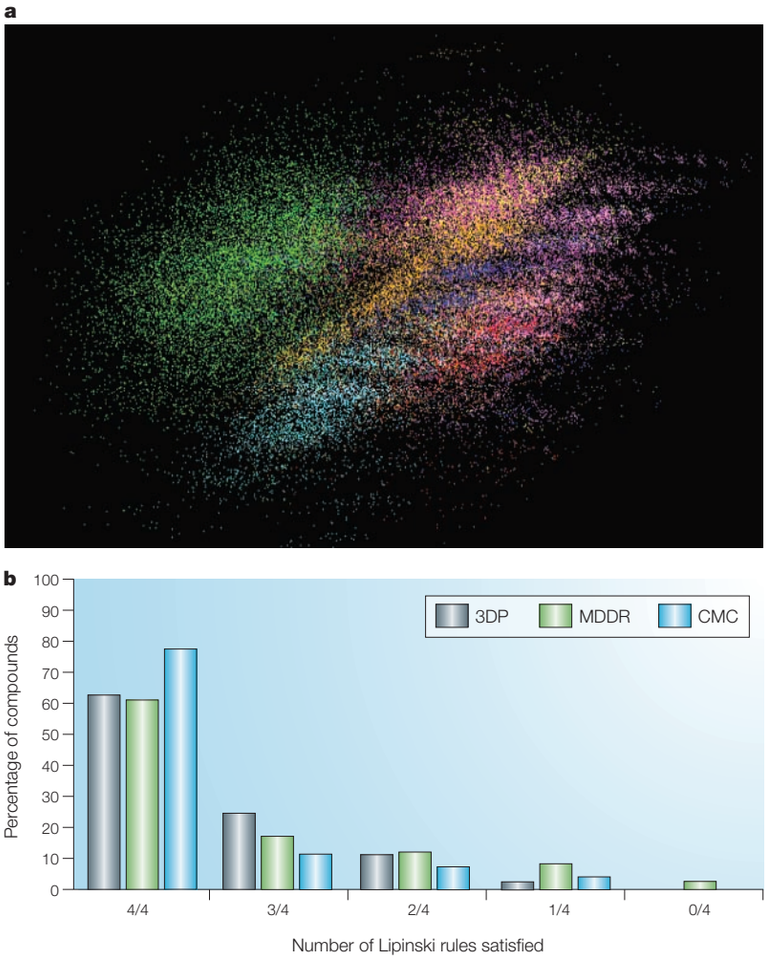

Figure 4 | **Diversity and drug likeness of the 3DP probe library. a** | Three-dimensional, nonlinear map of ~200,000 compounds that comprise the probe library of 3-Dimensional Pharmaceuticals (3DP). The map is constructed in such a way that the distances between the compounds on the map approximate as closely as possible to the corresponding similarities of the respective compounds. Each structural class (scaffold) is highlighted with a different colour, and was designed to be both internally and externally diverse (that is, the individual libraries consist of diverse compounds and complement each other in diversity space). **b** | Percentage of compounds from 3DP's probe library, MDDR (MDL Drug Data Report) and the CMC (Comprehensive Medicinal Chemistry) database that satisfy the Lipinski constraints.

of a set of compounds. However, their use for analysing combinatorial libraries is limited by computational complexity, difficulties arising from conformational flexibility, and information loss that occurs when the descriptors of the individual conformations are combined to produce the descriptors of the ensemble. Several substituent-based methods have been devised to address some of these shortcomings, but they are typically based on the assumption that the conformations that are adopted by a particular substituent do not depend on the other substituents or the scaffold, which is not always valid[41–43]. More importantly, these descriptors are specific to a particular library and cannot be used for cross-library comparisons. Experience indicates that many three-dimensional features are captured implicitly when a fairly comprehensive set of one-dimensional and two-dimensional descriptors are used[44]. Explicit three-dimensional approaches[45,46], including receptor docking[47,48], are more appropriate for filtering relatively small collections (~10^5 compounds, although more scalable methods have been described[49]), and are applicable only to targets with known (or predictable) structures.

Efficient navigation of chemical space requires that the relationships between compounds be represented in an intuitive manner that is easily understood by the chemist involved in a drug-optimization programme. The diversity of chemical space encourages the use of large descriptor sets to provide adequate structural and/or biological discrimination. However, the more descriptors that are used to describe the data, the greater the likelihood that they are correlated. Redundant variables can be a serious threat in many data-mining applications: they can distort similarities by over-emphasizing certain molecular characteristics at the expense of others, cause over-fitting in QSAR (quantitative structure–activity relationship) modelling, increase storage and computing requirements, and even limit the number of available analysis options. Dimensionality-reduction techniques fall into two broad categories: first, methods that preserve some of the original descriptors (such as fast random elimination of descriptors[50], cluster significance analysis[51] and information theory[52]); and second, methods that generate alternative latent features that are based on the original descriptors. The latter can be accomplished using linear methods, such as principal-components analysis[53], singular-value decomposition[54,55] and factor analysis[56], and nonlinear methods, such as multi-dimensional scaling[57] (MDS) and nonlinear mapping[58] (NLM). Linear methods transform a set of vectors that are described by partially cross-correlated variables into a smaller number of orthogonal variables. By contrast, nonlinear methods attempt to extract low-dimensional representations that preserve the relationships of the original data objects. MDS is particularly valuable, because it can also be used to produce Cartesian coordinate vectors from data that are supplied directly in the form of proximities, which simplifies their analysis using conventional statistical and data-mining techniques. Although MDS is a computationally intensive procedure, a recently published technique that involves the use of neural networks allows the scaling of data sets that are orders of magnitude larger than those that are accessible with conventional algorithms[59]. This general strategy was subsequently extended to use local learning techniques[60], generalized to handle complex distance functions and input data supplied in non-vectorial form[61], and modified to allow the scaling of combinatorial libraries in a way that circumvents explicit enumeration[62] (see below).

Assessing diversity. Diversity-profiling techniques fall into three general categories: first, cell-based methods, which divide chemical space into (hyper)rectangular regions and measure the occupancy of the resulting cells; second, variance-based methods, which measure the degree of correlation between the pertinent features

PHARMACOPHORE
The ensemble of steric and electronic features that are necessary to ensure optimal supramolecular interactions with a specific biological target and to trigger (or block) its biological function. Only molecules that interact at the same receptor site in the same way share a common pharmacophore.

DRUG LIKENESS
The thesis that drugs have certain common properties that differentiate them from other ordinary chemicals.

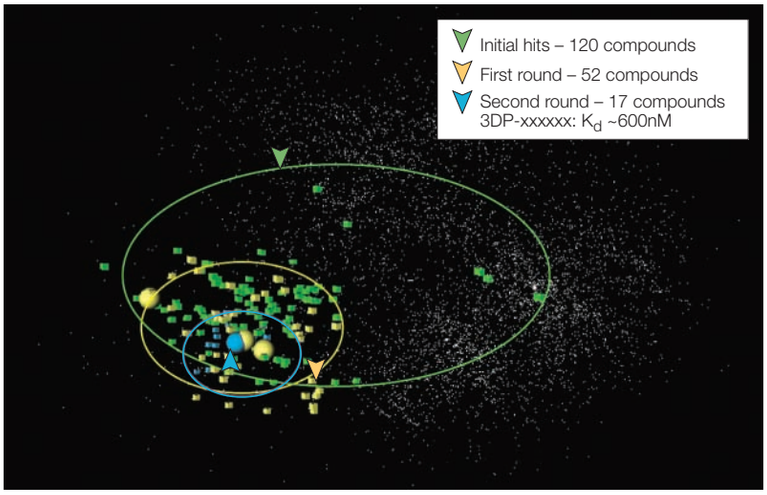

Figure 5 | **Iterative library design.** Exemplary application of the approach that is outlined in FIG. 1 on a proprietary target. The process begins by screening the probe library (or any diverse subset thereof) against the target, and designing increasingly focused arrays around the most promising hits to emerge from the screens. Depending on the number and quality of the hits, the selection of the compounds for the next iteration might be based on molecular similarity or formal statistical structure–activity models. The number of compounds that are synthesized in each iteration ranges from a few tens to a few thousands.

of molecules; and third, distance-based methods, which express diversity as a function of the pairwise molecular dissimilarities.

Cell-based methods encode absolute position in space, and can typically be computed very quickly. Several cell-based diversity functions have been proposed, ranging from simple counts of occupied cells to more elaborate measures[36,63]. Cell-based methods can be applied only to data spaces of modest dimensionality (typically no more than five or six), are sensitive to OUTLIERS, and have other limitations related to the CURSE OF DIMENSIONALITY[64].

Variance-based methods[65] attempt to find a subset of compounds with descriptors that show the least possible correlation, and which can optimally test the significance of each descriptor in predicting a relevant dependent variable (for example, biological activity). The most widely used method is D-optimal design[66], but the results are model dependent, and tend to favour the extremes of the feature space[67].

Distance-based methods are the most general, as they can work with any measure of molecular similarity and do not require a vectorial representation of chemical space. They can be used to produce 'spread' designs, which attempt to maximize the distances between the selected compounds to eliminate redundancy, or 'coverage' designs, which attempt to maximize representativeness[68–71]. Their main disadvantage is their QUADRATIC COMPLEXITY, but this problem can be alleviated using multidimensional search trees[72] or probabilistic approaches that measure the distributions of intermolecular dissimilarities[73].

Drug likeness. Choosing descriptors and diversity metrics solely on the basis of early-stage discovery objectives (for example, increasing the hit rate of high-throughput-screening experiments, increasing the clustering and separation of active from inactive compounds, or constructing predictive QSAR models on the basis of historical structure–activity data[74–79]) carries substantial risk. Successful drug development requires that compounds have many other properties besides potency and specificity. Indeed, most drug failures are related to development issues, such as drug solubility, uptake and distribution, metabolism, pharmacokinetics, toxicity, and chemical and metabolic stability. Recently, Lipinski *et al.*[80] analysed drugs on the market and developed a simple set of heuristic rules for determining the solubility and permeability of compounds that are being considered as drug candidates — the LIPINSKI RULE OF 5. This study caused a profound shift in the way in which combinatorial libraries are designed, and in some ways gave birth to the idea of 'drug likeness'. This concept has subsequently been elaborated to establish a broader range of structural and physicochemical properties that distinguish drugs from non-drugs. These characteristics are determined by analysing large databases of marketed drugs or advanced clinical candidates, and comparing them with collections of ordinary chemicals. Models range from simple, qualitative rule-based systems that are based on property distributions[81], pharmacophore[82] and BIOISOSTERIC preferences[83], and commonly occurring substructures[84], to more elaborate classification schemes that are based on unsupervised[56] and supervised[22,85,86] pattern-recognition techniques, or specific oral-bioavailability models[87]. These models can be used to detect potential ADME (absorption, distribution, metabolism and excretion) liabilities, and enrich combinatorial libraries with compounds that have an increased probability of leading to a successful preclinical candidate[88]. In practice, compound selection might be implemented through a 'hard' filter that eliminates problematic compounds from further consideration[16], or a 'soft' bias that skews the selection towards favourable ADME regions[89–93]. The latter approach is supported by a recent review, which noted that three of the top-selling GlaxoWellcome drugs would be considered inappropriate by many drug-likeness filters[10].

Although debate continues about the optimal size and diversity that is appropriate for initial screening, our laboratory has had good overall success by screening libraries of ~300,000 drug-like compounds based on ~50 underlying chemical scaffolds (FIGS 3 and 4).

Design of experiments

Combinatorial libraries are almost invariably synthesized in the form of arrays that represent all combinations of a prescribed set of building blocks. The fundamental problem in library design is to identify the monomers that, when combined, will produce the library that best satisfies a predefined set of objectives. Depending on the available information, two stages in the design cycle can be identified: first, lead discovery, which involves the screening of large, diverse chemical libraries in search of novel hits; and second, lead optimization, which involves the synthesis of smaller, focused libraries that

LOG P
The octanol/water partition coefficient is the ratio of the compound's solubility in octanol to its solubility in water. The logarithm of this partition coefficient is called log P. It provides an estimate of the compound's ability to pass through a cell membrane.

OUTLIER
A point that, because of observation noise, does not follow the characteristics of the input (or desired response) data.

CURSE OF DIMENSIONALITY
The sparsity of data in higher dimensions.

QUADRATIC COMPLEXITY
Quadratic complexity means that if the size of the problem doubles, the computational time that is required by the algorithm to solve it quadruples. The complexity (or order) of an algorithm is an important criterion for comparing algorithms that involve the analysis of large data sets.

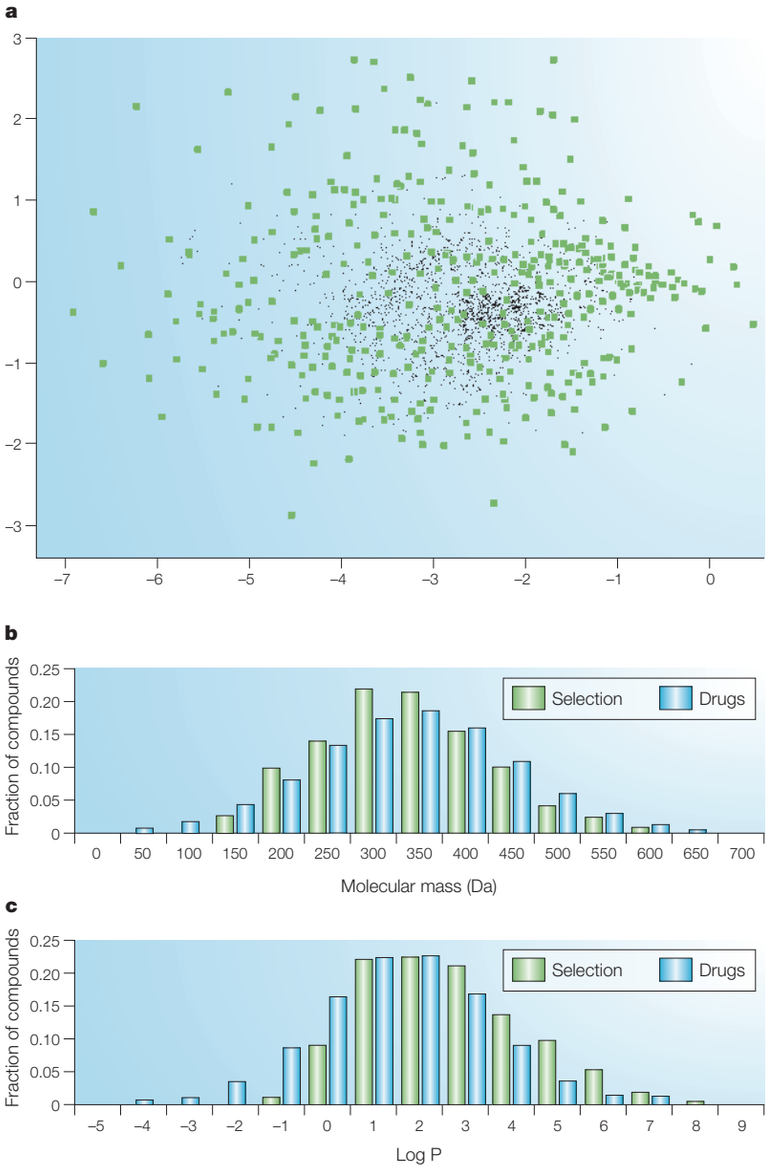

Figure 6 | **Multiobjective library design.** Multiobjective selection of a 20×20 combinatorial array from a 300×300 virtual library based on the reductive-amination reaction. The selection was designed to simultaneously maximize molecular diversity, satisfy matrix-synthesis constraints and enforce drug-like molecular weight and log P distributions. **a** | Nonlinear map that represents diversity space. **b** | Molecular-mass distribution of selected compounds plotted against that of known drugs. **c** | Log P distribution of selected compounds plotted against that of known drugs. Reproduced from REF. 91 © (2000) with permission from Elsevier Science.

are designed to explore the structure–activity space around these hits.

In either case, library design is a complex task that requires the simultaneous optimization of multiple, frequently conflicting objectives. There are no formal rules for setting these objectives, and the decision often involves a mixture of mathematics and chemical intuition. Regardless of the specific goals, monomer selection is a COMBINATORIAL OPTIMIZATION problem of formidable proportions. Two different approaches have been developed: reagent-based design, in which each variation site is considered independently of all the others[66,94]; and product-based design, for which the selection is ultimately based on the properties of the enumerated products. The latter is generally the method of choice[95,96], but, until recently (see below), it was believed to be applicable to only explicitly enumerated libraries of relatively small size.

The selection problem can be viewed as a heuristic search for which each state in the search space represents a particular subset of the virtual library. Design strategies include clustering or partitioning methods, greedy optimization heuristics, and advanced stochastic optimization schemes. Clustering methods divide the population into disjointed sets, and select several representatives from each set[97,98]. Greedy algorithms carry out the selection in a stepwise manner by making decisions that make sense at the time without regard for future consequences. Such methods include forward selection, backward elimination and greedy replacement[65], and other variants that have been specifically tailored for combinatorial arrays[99,100]. Depending on the scale of the problem and the complexity of the cost function, these algorithms can be prohibitively slow and often converge to suboptimal local minima.

Stochastic algorithms attempt to circumvent the multiple-minima problem by allowing the generation of successor states that might be inferior to their predecessors. Several methods have been investigated, including simulated annealing[101–106], genetic algorithms[90,107–111] and particle swarms[112]. These methods can accommodate virtually any conceivable design criterion, including diversity, similarity to known active compounds, predicted activity and/or selectivity as determined by a QSAR or receptor-binding model, enforcement of drug-like property distributions, reagent cost and availability, and many others. Examples of how these criteria can evolve as a function of time, and how multiple design objectives can be optimized simultaneously, are illustrated in FIGS 5 and 6, respectively.

Beyond enumeration

Descriptor calculation for explicitly enumerated compounds typically proceeds at a rate of a few hundred compounds per second (at best), and can be used with only small or medium-sized virtual libraries. Two different approaches have been devised to address this problem. The first is to perform selective enumeration, and the second is to use descriptors that do not require explicit construction of the connection tables of the products. An example of the former strategy is the similarity-searching algorithm that is outlined in REF. 113. The algorithm is based on the observation that the structural diversity of a combinatorial library stems from a limited number of building blocks, so that it is possible — through random sampling — to identify reagents that lead to products that are most closely related to the query structure. This algorithm provides optimal or nearly optimal solutions in rapid time frames, but is applicable to only a relatively narrow class of optimization tasks (mainly similarity searching and other types of focused design).

LIPINSKI RULE OF 5
For compounds that are not substrates of biological transporters, poor absorption and permeation are more likely to occur when there are more than 5 hydrogen-bond donors, more than 10 hydrogen-bond acceptors, the molecular mass is greater than 500 Da, or the log P is greater than 5.

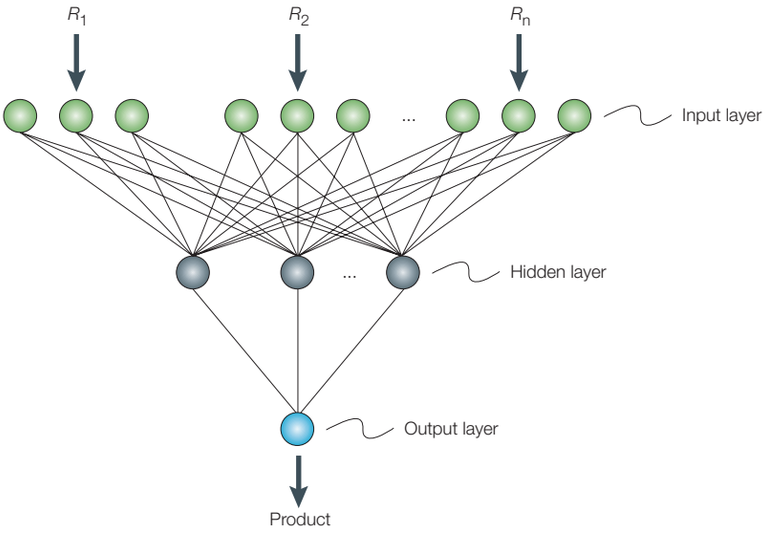

Figure 7 | **Combinatorial networks.** Combinatorial neural networks (CNNs) are multilayer perceptrons (MLPs) that are trained to reproduce properties of combinatorial products from pertinent features (descriptors) of their respective building blocks[117]. CNNs typically comprise an input layer that contains $r \times n$ neurons — for which r is the number of variation sites in the combinatorial library and n is the number of reagent descriptors — one or more hidden layers, and an output layer that has a single neuron for each product property that is predicted by the neural network.

The alternative approach is to use 'decomposable' descriptors, which can be computed in an additive or nearly additive manner from the corresponding descriptor values of the clipped reagents[114–116]. Although the products need not be assembled, the most useful descriptors are not additive, and are therefore not amenable to this approach. So, both of the aforementioned approaches limit either the types of selection that can be carried out, or the types of descriptor that can be used in the design.

Recently, however, there has been evidence that direct calculation of descriptors might be unnecessary[117]. Indeed, it was found that most of the descriptors that are commonly used in library design can be estimated accurately from properties of their respective building blocks, including non-decomposable descriptors that cannot be computed by simple addition of fragment contributions. The approach that is outlined in REF. 117 attempts to construct models that are specific to a particular library and a particular descriptor (or set of descriptors), and use those models to generate approximate descriptors 'on the fly', as needed by the application.

This approach requires a training set, which consists of the descriptors of all the reagents that make up the combinatorial library, along with the properties of a relatively small number of randomly chosen products. The latter are computed in a conventional way; that is, by running a computer program or subroutine on the fully enumerated structures. These data are then used as input to a COMBINATORIAL NEURAL NETWORK (CNN) (FIG. 7), which is trained to predict the properties of the products from the pre-computed descriptors of their respective building blocks. Once the network is trained, properties of other products can be calculated by simply extracting the descriptors of the corresponding reagents and feeding them through the neural network. As it is not possible to know *a priori* which reagent descriptors are most relevant for a particular learning task, the training process is often guided by a FEATURE-SELECTION algorithm, such as simulated annealing, evolutionary programming or some other alternative. This approach limits the expensive enumeration and descriptor calculation to only a small fraction of products (the training set), and can estimate the descriptors of most of the compounds in the virtual library without generating their connection tables. More importantly, it does not require the use of clipped reagents, and can be applied to a wide range of molecular properties regardless of origin and complexity with minimal programming effort. In essence, the method encodes complex computer programs in the synaptic parameters of a neural network, and allows descriptor calculations for virtual libraries at a rate of hundreds of thousands of compounds per second.

This approach is broadly applicable, and can be easily extended to predict latent variables, such as principal components or nonlinear map coordinates[62], and other complex molecular properties that require a much more substantial effort to compute than simple chemical descriptors (for example, ADME-related physicochemical properties, such as log P, pKa, solubility, docking scores and so on; D. K. A. and V. S. L., unpublished observations).

Concluding remarks

Realizing the potential benefits of the genomics revolution, and particularly the potential to tailor new drugs to specific genotypic backgrounds, will require significant advances in the efficiency with which new drugs are discovered and developed. The clear path is one that actively incorporates any and all of the information that is required to specify the ultimate development compound throughout each stage of its evolution — from initial screening-library design to final clinical-candidate optimization and ADME refinement. In this article, we have described a set of chemi-informatics tools aimed at library design for general screening and early-stage-compound optimization that are based primarily on the measured binding properties and statistically observed ADME profiles of known marketed drugs. We note, however, that the underlying approaches that have been developed are readily geared to the incorporation of a much more diverse array of both experimental high-throughput ADME data to better address clinical-development issues, and pharmacogenomics data to better match drug properties and patient genotype. Indeed, the prime motivation for the development of highly efficient chemi-informatics tools for drug discovery lies not in the process as it is perceived at present, but in the need to build flexible and scalable informatics tools to accommodate the expansion in both the available chemistry and the diversity of ADME and pharmacogenomics data. Incorporation of such data can productively focus the discovery and development process, and probably defines the only systematic course to economically creating medicines that are tailored to specific genotypes.

BIOISOSTERISM
The idea that a chemical group in a biologically active molecule can be replaced by another chemical group without loss of activity.

COMBINATORIAL OPTIMIZATION
The number of different combinations of k objects out of a set of n objects is given by the binomial coefficient $C_k^n = n!/(n-k)!k!$. This can be used to calculate the number of distinct $k_1 \times k_2 \times \ldots \times k_d$ combinatorial arrays in a $n_1 \times n_2 \times \ldots \times n_d$ combinatorial library. For example, there are approximately 10^{40} different 10×10×10 arrays in a 100×100×100 library.

COMBINATORIAL NEURAL NETWORK
(CNN). A neural network that is trained to predict molecular properties of combinatorial products from pertinent features of their respective building blocks.

FEATURE SELECTION
A computational technique that attempts to identify a small subset of features that are most relevant to a particular machine learning task.

1. International Human Genome Sequencing Consortium. Initial Sequencing and Analysis of the Human Genome. *Nature* **409**, 860–921 (2001).
2. Agrafiotis, D. K., Bone, R. F., Salemme, F. R. & Soll, R. M. System and method for automatically generating chemical compounds with desired properties. US Patent 5,463,564 (1995).
3. Agrafiotis, D. K., Bone, R. F., Salemme, F. R. & Soll, R. M. System and method for automatically generating chemical compounds with desired properties. US Patent 5,574,656 (1996).
4. Agrafiotis, D. K., Bone, R. F., Salemme, F. R. & Soll, R. M. System, method and computer program for at least partially automatically generating chemical compounds having desired properties. US Patent 5,684,711 (1997).
5. Agrafiotis, D. K., Bone, R. F., Salemme, F. R. & Soll, R. M. System, method and computer program for at least partially automatically generating chemical compounds with desired properties from a list of potential chemical compounds to synthesize. US Patent 5,901,069 (1999).
6. Pantoliano, M. P. *et al*. High density miniaturized thermal shift assay as a general strategy for drug discovery. *J. Biomol. Screen.* **6**, 492–440 (2001).
   This article describes the use of a high-throughput, fluorescence-based method for detecting thermal phase transitions in proteins as a means to evaluate their stability and the effects of bound ligands.
7. Martin, E. J., Spellmeyer, D. C., Critchlow, R. E. Jr & Blaney, J. M. in *Reviews in Computational Chemistry* Vol. 10 (eds Lipkowitz, K. B. & Boyd, D. B.) 75–100 (VCH, Weinheim, 1997).
8. Agrafiotis, D. K. in *The Encyclopedia of Computational Chemistry* (eds Schleyer, P. V. R. *et al*.) 742–761 (John Wiley and Sons, Chichester, 1998).
9. Bures, M. G. & Martin, Y. C. Computational methods in molecular diversity and combinatorial chemistry. *Curr. Opin. Chem. Biol.* **2**, 376–380 (1998).
10. Agrafiotis, D. K., Myslik, J. C. & Salemme, F. R. Advances in diversity profiling and combinatorial series design. *Mol. Divers.* **4**, 1–22 (1999).
    An in-depth review of computational methods that are used in diversity analysis and combinatorial-library design.
11. Drewry, D. H. & Young, S. S. Approaches to the design of combinatorial libraries. *Chemometr. Intell. Lab. Syst.* **48**, 1–20 (1999).
12. Leach, A. R. & Hann, M. M. The *in silico* world of virtual libraries. *Drug Discov. Today* **5**, 326–336 (2000).
13. Leland, B. A. *et al*. Managing the combinatorial explosion. *J. Chem. Inf. Comput. Sci.* **37**, 62–70 (1997).
14. Leach, A. R., Bradshaw, J., Green, D. V. S., Hann, M. M. & Delany, J. J. Implementation of a system for reagent selection and library enumeration, profiling & design. *J. Chem. Inf. Comput. Sci.* **39**, 1161–1172 (1999).
15. Lobanov, V. S. & Agrafiotis, D. K. Scalable methods for the construction and analysis of virtual combinatorial libraries. *Combin. Chem. High-Throughput Screen.* **5**, 167–178 (2002).
16. Walters, W. P., Stahl, M. T. & Murcko, M. A. Virtual screening — an overview. *Drug Discov. Today* **3**, 160–178 (1998).
17. Agrafiotis, D. K., Lobanov, V. S., Rassokhin, D. N. & Izrailev, S. in *Virtual Screening for Bioactive Molecules* (eds Böhm, H.-J. & Schneider, G.) 265–300 (Wiley–VCH, Weinheim, 2000).
18. Johnson, M. A. & Maggiora, G. M. *Concepts and Applications of Molecular Similarity* (Wiley, New York, 1990).
    An authoritative overview of the theoretical and practical aspects of molecular similarity as it applies to chemical and biological research.
19. Livingston, D. J. The characterization of molecular structures using molecular properties. A survey. *J. Chem. Inf. Comput. Sci.* **40**, 195–209 (2000).
20. Hall, L. H. & Kier, L. B. in *Reviews of Computational Chemistry* (eds Boyd, D. B. & Lipkowitz, K. B.) 367–422 (VCH, Weinheim, 1991).
    Describes a class of important molecular-connectivity indices and their use in predicting molecular properties.
21. James, C. A., Weininger, D. & Delaney, J. Daylight Theory Manual. *Daylight Chemical Information Systems* [online] (cited 12 Mar 02) <http://www.daylight.com/>.
22. Sadowski, J. & Kubinyi, H. A scoring scheme for discriminating between drugs and nondrugs. *J. Med. Chem.* **41**, 3325–3329 (1998).
    Describes the application of neural networks for discriminating drugs from non-drugs by using simple atom-type descriptors.
23. Schneider, G., Neidhart, W., Giller, T. & Schmid, G. Scaffold-hopping by topological pharmacophore search: a contribution to virtual screening. *Angew. Chem. Int. Edn Engl.* **38**, 2894–2896 (1999).
24. Carhart, R. E., Smith, D. H. & Venkataraghavan, R. Atom pairs as molecular features in structure–activity studies: definition and application. *J. Chem. Inf. Comput. Sci.* **25**, 64–73 (1985).
25. Nilakantan, R., Bauman, N., Dixon, J. S. & Venkataraghavan, R. Topological torsions: a new molecular descriptor for SAR applications. Comparison with other descriptors. *J. Chem. Inf. Comput. Sci.* **27**, 82–85 (1987).
26. Kearsley, S. K. *et al*. Chemical similarity using physicochemical property descriptors. *J. Chem. Inf. Comput. Sci.* **36**, 118–127 (1996).
27. Moreau, G. & Broto, P. The autocorrelation of a topological structure: a new molecular descriptor. *Nouv. J. Chim.* **4**, 359–360 (1980).
28. Bauknecht, H. *et al*. Locating biologically active compounds in medium-sized heterogeneous datasets by topological autocorrelation vectors: dopamine and benzodiazepine agonists. *J. Chem. Inf. Comput. Sci.* **36**, 1205–1213 (1996).
29. Labute, P. A widely applicable set of descriptors. *J. Mol. Graph. Model.* **18**, 464–467 (2000).
30. Kubinyi, H. in *Methods and Principles in Medicinal Chemistry* Vol. 1 (eds Manhold, R., Krogsgaard-Larsen, P. & Timmermann, H.) 21–36 (VCH, Weinheim, 1993).
31. Burden, F. R. Molecular identification number for substructure searches. *J. Chem. Inf. Comput. Sci.* **29**, 225–227 (1989).
32. Sheridan, R. P., Miller, M. D., Underwood, D. J. & Kearsley, S. K. Chemical similarity using geometric atom pair descriptors. *J. Chem. Inf. Comput. Sci.* **36**, 128–136 (1996).
33. Wagener, M., Sadowski, J. & Gasteiger, J. Autocorrelation of molecular surface properties for modeling corticosteroid binding globulin and cytosolic Ah receptor activity by neural networks. *J. Am. Chem. Soc.* **117**, 7769–7775 (1995).
34. Todeschini, R., Lasagni, M. & Marengo, E. New molecular descriptors for 2D and 3D structures. Theory. *J. Chemom.* **8**, 263–272 (1994).
35. Ghuloum, A. M., Sage, C. R. & Jain, A. N. Molecular hashkeys: a novel method for molecular characterization and its application for predicting important pharmaceutical properties of molecules. *J. Med. Chem.* **42**, 1739–1748 (1999).
36. Pearlman, R. S. & Smith, K. M. Metric validation and the receptor-relevant subspace concept. *J. Chem. Inf. Comput. Sci.* **9**, 28–35 (1999).
37. Sheridan, R. P. *et al*. 3Dsearch; a system for three-dimensional substructure searching. *J. Chem. Inf. Comput. Sci.* **29**, 255–260 (1989).
38. Murrall, N. W. & Davies, E. K. Conformational freedom in 3-D databases. 1. Techniques. *J. Chem. Inf. Comput. Sci.* **30**, 312–316 (1990).
39. Guner, O. F. *Pharmacophore Perception, Development and Use in Drug Design* (International Univ. Line, La Jolla, 2000).
    A collection of articles that describe the use of pharmacophore modelling in drug design.
40. Mason, J. S. *et al*. New 4-point pharmacophore method for molecular similarity and diversity applications: overview of the method and applications, including a novel approach to the design of combinatorial libraries containing priviledged substructures. *J. Med. Chem.* **42**, 3251–3264 (1999).
41. Leach, A. R., Green, D. V. S., Hann, M. M., Judd, D. B. & Good, A. C. Where are the GaPs? A rational approach to monomer acquisition and selection. *J. Chem. Inf. Comput. Sci.* **40**, 1262–1269 (2000).
42. Martin, E. J. & Hoeffel, T. J. Oriented substituent pharmacophore property space (OSPPREYS): A substituent-based calculation that describes combinatorial library products better than the corresponding product-based selection. *J. Mol. Graph. Model.* **18**, 383–403 (2000).
    This paper describes the use of substituent-based pharmacophore descriptors to encode conformation-dependent properties of combinatorial products.
43. Cramer, R. D., Clark, R. D., Patterson, D. E. & Ferguson, A. M. Bioisosterism as a molecular diversity descriptor: steric fields of single topomeric conformers. *J. Med. Chem.* **39**, 3060–3069 (1996).
44. Matter, H. & Potter, T. Comparing 3D pharmacophore triplets and 2D fingerprints for selecting diverse compound subsets. *J. Chem. Inf. Comput. Sci.* **39**, 1211–1225 (1999).
45. Salemme, F. R., Spurlino, J. & Bone, R. Serendipity meets precision: the integration of structure based drug design and combinatorial chemistry for efficient drug discovery. *Structure* **5**, 319–324 (1997).
46. Graybill, T. L. *et al.* in *Molecular Diversity and Combinatorial Chemistry* (eds Chaiken, I. M. & Janda, K. D.) 16–26 (ACS, Washington DC, 1996).
47. Jones, G., Willett, P., Glen, R. C., Leach, A. R. & Taylor, R. Further development of a genetic algorithm for ligand docking and its application to screening combinatorial libraries. *ACS Symp. Ser.* **719**, 271–291 (1999).
48. Waszkowycz, B., Perkins, T. D. J., Sykes, R. A. & Li, J. Large-scale virtual screening for discovering leads in the post-genomics era. *IBM Syst. J.* **40**, 360–376 (2001).
49. Sun, Y., Ewing, T. J. A., Skillman, A. G. & Kuntz, I. D. CombiDock: structure-based combinatorial docking and library design. *J. Comput. Aided. Mol. Des.* **12**, 597–604 (1998).
50. Waller, C. L. & Bradley, M. P. Development and validation of a novel variable selection technique with application to multidimensional quantitative structure-activity relationship studies. *J. Chem. Inf. Comput. Sci.* **39**, 345–355 (1999).
51. Rose, V. S. & Wood, J. Generalized cluster significance analysis with conditional probabilities. *Quant. Struct. Activ. Rel.* **17**, 348–356 (1998).
52. Godden, J. W. & Bajorath, J. Differential Shannon entropy as a sensitive measure of differences in database variability of molecular descriptors. *J. Chem. Inf. Comput. Sci.* **41**, 1060–1066 (2001).
53. Cooley, W. & Lohnes, P. *Multivariate Data Analysis* (Wiley, New York, 1971).
54. Xie, D., Tropsha, A. & Schlick, T. An efficient projection protocol for chemical databases: singular value decomposition combined with truncated Newton minimization. *J. Chem. Inf. Comput. Sci.* **40**, 167–177 (2000).
55. Hull, R. D. *et al*. Latent semantic structure indexing (LASSI) for defining chemical similarity. *J. Med. Chem.* **44**, 1177–1184 (2001).
56. Cummins, D. J., Andrews, C. W., Bentley, J. A. & Cory, M. Molecular diversity in chemical databases: comparison of medicinal chemistry knowledge bases and databases of commercially available compounds. *J. Chem. Inf. Comput. Sci.* **36**, 750–763 (1996).
57. Kruskal, J. B. Non-metric multidimensional scaling: a numerical method. *Phychometrika* **29**, 115–129 (1964).
58. Sammon, J. W. A nonlinear mapping for data structure analysis. *IEEE Trans. Comput.* **C18**, 401–409 (1969).
59. Agrafiotis, D. K. & Lobanov, V. S. Nonlinear mapping networks. *J. Chem. Inf. Comput. Sci.* **40**, 1356–1362 (2000).
60. Rassokhin, D. N., Lobanov, V. S. & Agrafiotis, D. K. Nonlinear mapping of massive data sets by fuzzy clustering and neural networks. *J. Comput. Chem.* **22**, 373–386 (2001).
61. Agrafiotis, D. K., Rassokhin, D. N. & Lobanov, V. S. Multidimensional scaling and visualization of large molecular similarity tables. *J. Comput. Chem.* **22**, 488–500 (2001).
62. Agrafiotis, D. K. & Lobanov, V. S. Multidimensional scaling of combinatorial libraries without explicit enumeration. *J. Comput. Chem.* **22**, 1712–1722 (2001).
63. Jamois, E. A., Hassan, M. & Waldman, M. Evaluation of reagent-based and product-based strategies in the design of combinatorial library subsets. *J. Chem. Inf. Comput. Sci.* **40**, 63–70 (2000).
64. Agrafiotis, D. K. & Rassokhin, D. N. A fractal approach for selecting an appropriate bin size for cell-based diversity estimation. *J. Chem. Inf. Comput. Sci.* **42**, 117–122 (2002).
65. Montgomery, D. C. *Design and Analysis of Experiments* 4th edn (John Wiley and Sons, New York, 1996).
66. Martin, E. J. *et al*. Measuring diversity: Experimental design of combinatorial libraries for drug discovery. *J. Med. Chem.* **38**, 1431–1436 (1995).
    This paper describes the use of statistical experimental-design techniques to select building blocks for combinatorial libraries using a rich set of molecular descriptors.
67. Hassan, M., Bielawski, J. P., Hempel, J. C. & Waldman, M. Optimization and visualization of molecular diversity of combinatorial libraries. *Mol. Divers.* **2**, 64–74 (1996).
68. Kennard, R. W. & Stone, L. A. Computer-aided design of experiments. *Technometrics* **11**, 137–148 (1969).
69. Higgs, R. E., Bemis, K. G., Watson, I. A. & Wikel, J. H. Experimental designs for selecting molecules from large chemical databases. *J. Chem. Inf. Comput. Sci.* **37**, 861–870 (1997).
70. Snarey, M., Terrett, N. K., Willett, P. & Wilton, D. J. Comparison of algorithms for dissimilarity-based compound selection. *J. Mol. Graph. Model.* **15**, 372–385 (1997).
71. Mount, J., Ruppert, J., Welch, W. & Jain, A. N. IcePick: a flexible surface-based system for molecular diversity. *J. Med. Chem.* **42**, 60–66 (1999).
72. Agrafiotis, D. K. & Lobanov, V. S. An efficient implementation of distance-based diversity metrics based on k-d trees. *J. Chem. Inf. Comput. Sci.* **39**, 51–58 (1999).
73. Agrafiotis, D. K. A constant time algorithm for estimating the diversity of large chemical libraries. *J. Chem. Inf. Comput. Sci.* **41**, 159–167 (2001).
74. Downs, G. M. & Willett, P. Similarity searching and clustering of chemical-structure databases using molecular property data. *J. Chem. Inf. Comput. Sci.* **34**, 1094–1102 (1994).
75. Brown, R. D. & Martin, Y. C. Use of structure–activity data to compare structure-based clustering methods and descriptors for use in compound selection. *J. Chem. Inf. Comput. Sci.* **36**, 572–584 (1996).
    A comparison of several two-dimensional and three-dimensional descriptors, which is based on their ability to discriminate active from inactive compounds.

76. Brown, R. D. & Martin, Y. C. The information content of 2D and 3D structural descriptors relevant to ligand–receptor binding. *J. Chem. Inf. Comput. Sci.* **37**, 1–9 (1997).
77. Patterson, D. E., Cramer, R. D., Ferguson, A. M., Clark, R. D. & Weinberger, L. E. Neighborhood behavior: a useful concept for validation of molecular diversity descriptors. *J. Med. Chem.* **39**, 3049–3059 (1996).
78. Matter, H. Selecting optimally diverse compounds from structure databases: a validation study of two-dimensional and three-dimensional molecular descriptors. *J. Med. Chem.* **40**, 1219–1229 (1997).
79. Martin, Y. C., Bures, M. G. & Brown, R. D. Validated descriptors for diversity measurements and optimization. *Pharm. Pharmacol. Commun*. **4**, 147–152 (1998).
80. Lipinski, C. A., Lombardo, F., Dominy, B. W. & Feeny, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Deliv. Rev.* **23**, 3–25 (1997).
A discussion of the importance of ADME screening in early-stage drug discovery.
81. Oprea, T. I. Property distribution of drug-related chemical databases. *J. Comput. Aided Mol. Des.* **14**, 251–264 (2000).
82. Muegge, I., Heald, S. L. & Brittelli, D. Simple selection criteria for drug-like chemical matter. *J. Med. Chem.* **44**, 1841–1846 (2001).
83. Sheridan, R. P. The most common chemical replacements in drug-like compounds. *J. Chem. Inf. Comput. Sci.* **2**, 103–108 (2002).
84. Wang, J. & Ramnarayan, K. Toward designing drug-like libraries: a novel computational approach for prediction of drug feasibility of compounds. *J. Combin. Chem.* **1**, 524–533 (1999).
85. Ajay, A., Walters, W. P. & Murcko, M. A. Can we learn to distinguish between drug-like and nondrug-like molecules? *J. Med. Chem.* **41**, 3314–3324 (1998).
86. Wagener, M. & van Geerestein, V. J. Potential drugs and nondrugs: prediction and identification of important structural features. *J. Chem. Inf. Comput. Sci.* **40**, 280–292 (2000).
87. Yu, L. X., Lipka, E., Crison, J. R. & Amidon, G. L. Transport approach to the biopharmaceutical design of oral drug delivery systems: prediction of intestinal absorption. *Adv. Drug Deliv. Rev.* **19**, 359–376 (1996).
88. Teague, S. J., Davis, A. M., Leeson, P. D. & Oprea, T. I. The design of leadlike combinatorial libraries. *Angew. Chem. Int. Edn Engl.* **38**, 3743–3748 (1999).
Based on an analysis of 18 lead-drug pairs, the authors point out that traditional medicinal chemistry optimization tends to drive initial high-throughput screening (HTS) hits outside the "rule-of-five" range, and suggest that combinatorial libraries should have lower molecular masses and lower log P profiles than those originally proposed by Lipinski.
89. Koehler, R. T., Dixon, S. L. & Villar, O. H. LASSOO: a generalized directed diversity approach to the design and enrichment of chemical libraries. *J. Med. Chem.* **42**, 4695–4704 (1999).
90. Gillet, V. J., Willet, P., Bradshaw, J. & Green, D. V. S. Selecting combinatorial libraries to optimize diversity and physical properties. *J. Chem. Inf. Comput. Sci.* **39**, 169–177 (1999).
91. Rassokhin, D. N. & Agrafiotis, D. K. Kolmogorov–Smirnov statistic and its applications in library design. *J. Mol. Graph. Model.* **18**, 370–384 (2000).
92. Brown, R. D., Hassan, M. & Waldman, M. Combinatorial library design for diversity, cost efficiency and drug-like character. *J. Mol. Graph. Model.* **18**, 427–437 (2000).
93. Shi, S., Peng, Z., Kostrowicki, J., Paderes, J. & Kuki, A. Efficient combinatorial filtering for desired molecular properties of reaction products. *J. Mol. Graph. Model.* **18**, 478–496 (2000).
94. Martin, E. & Wong, A. Sensitivity analysis and other improvements to tailored combinatorial library design. *J. Chem. Inf. Comput. Sci.* **40**, 215–220 (2000).
95. Gillet, V. J., Willett, P. & Bradshaw, J. The effectiveness of reactant pools for generating structurally-diverse combinatorial libraries. *J. Chem. Inf. Comput. Sci.* **37**, 731–740 (1997).
96. Jamois, E. A., Hassan, M. & Waldman, M. Evaluation of reagent-based and product-based strategies in the design of combinatorial library subsets. *J. Chem. Inf. Comput. Sci.* **40**, 63–70 (2000).
97. Graham, E. T., Jacober, S. P. & Cardoso, M. G. A novel frequency distribution selection method for efficient plate layout of a diverse combinatorial library. *J. Chem. Inf. Comput. Sci.* **41**, 1508–1516 (2001).
98. Bayada, D. M., Hamersma, H. & van Geerestein, V. J. Molecular diversity and representativity in chemical databases. *J. Chem. Inf. Comput. Sci.* **39**, 1–10 (1999).
99. Agrafiotis, D. K. & Lobanov, V. S. Ultrafast algorithm for designing focused combinatorial arrays. *J. Chem. Inf. Comput. Sci.* **40**, 1030–1038 (2000).
100. Stanton, R. V. *et al*. Combinatorial library design: maximizing model fitting compounds with matrix synthesis constraints. *J. Chem. Inf. Comput. Sci.* **40**, 701–705 (2000).
101. Agrafiotis, D. K. Stochastic algorithms for maximizing molecular diversity. *J. Chem. Inf. Comput. Sci.* **37**, 841–851 (1997).
102. Hassan, M., Bielawski, J. P., Hempel, J. C. & Waldman, M. Optimization and visualization of molecular diversity of combinatorial libraries. *Mol. Diversity* **2**, 64–74 (1996).
103. Good, A. C. & Lewis, R. A. New methodology for profiling combinatorial libraries and screening sets: cleaning up the design process with HARPcik. *J. Med. Chem.* **40**, 3926–3236 (1997).
104. Zheng, W., Cho, S. J. & Tropsha, A. Rational combinatorial library design: 1) Focus-2D: a new approach to the design of targeted combinatorial chemical libraries. *J. Chem. Inf. Comput. Sci.* **38**, 251–258 (1998).
105. Waldman, M., Li, H. & Hassan, M. Novel algorithms for the optimization of molecular diversity of combinatorial libraries. *J. Mol. Graph. Model.* **18**, 412–426 (2000).
106. Agrafiotis, D. K. Multiobjective optimization of combinatorial libraries. *IBM J. Res. Develop.* **45**, 545–566 (2001).
107. Sheridan, R. P. & Kearsley, S. K. Using a genetic algorithm to suggest combinatorial libraries. *J. Chem. Inf. Comput. Sci.* **35**, 310–3201 (1995).
108. Weber, L., Wallbaum, S., Broger, C. & Gubernator, K. Optimization of the biological activity of combinatorial compound libraries by a genetic algorithm. *Angew. Chem. Int. Edn Engl.* **34**, 2280–2282 (1995).
109. Singh, J. *et al*. Application of genetic algorithms to combinatorial synthesis: a computational approach for lead identification and lead optimization. *J. Am. Chem. Soc*. **118**, 1669–1676 (1996).
A description of the use of a genetic algorithm to optimize peptide-based collagenase substrates using direct experimental feedback, without constructing any intermediate models of biological activity.
110. Brown, R. D. & Martin, Y. C. Designing combinatorial library mixtures using genetic algorithms. *J. Med. Chem.* **40**, 2304–2313 (1997).
111. Sheridan, R. P., SanFeliciano, S. G. & Kearsley, S. K. Designing targeted libraries with genetic algorithms. *J. Mol. Graph. Model.* **18**, 320–334 (2000).
112. Farnum, M. & Agrafiotis, D. K. *Combinatorial Swarms* (CombiChem, London, 2001).
113. Lobarov, V. S. & Agrafiotis, D. K. Stochastic similarity selections from large combinatorial libraries. *J. Chem. Inf. Comput. Sci.* **40**, 460–470 (2000).
114. Downs, G. M. & Barnard, J. M. Techniques for generating descriptive fingerprints in combinatorial libraries. *J. Chem. Inf. Comput. Sci.* **37**, 59–61 (1997).
115. Cramer, R. D., Patterson, D. E., Clark, R. D., Soltanshahi, F. & Lawless, M. S. Virtual compound libraries: a new approach to decision making in molecular discovery research. *J. Chem. Inf. Comput. Sci.* **38**, 1010–1023 (1998).
116. Ivanciuc, O. & Klein, D. J. Computing Weiner-type indices for virtual combinatorial libraries generated from heteroatom-containing building blocks. *J. Chem. Inf. Comput. Sci.* **42**, 8–22 (2002).
117. Lobanov, V. S. & Agrafiotis, D. K. Combinatorial networks. *J. Mol. Graph. Model.* **19**, 571–578 (2001).
Describes the use of neural networks for predicting properties of combinatorial products from properties of their respective building blocks. This method allows product-based virtual screening of massive combinatorial libraries in a way that circumvents their virtual synthesis.

Online links

FURTHER INFORMATION
Chemical Informatics Societies and Professional Organizations:
http://www.indiana.edu/~cheminfo/informatics/cinformsocs.html
Access to this interactive links box is free online.

Dr Dimitris K. Agrafiotis is Executive Director of Research Informatics at 3-Dimensional Pharmaceuticals, Inc. (3DP). He holds a Ph.D. in theoretical organic chemistry from Imperial College of Science, Technology and Medicine, University of London (1988; adviser H. Rzepa). After postdoctoral training with A. Streitwieser at the University of California, Berkeley, and Nobel laureate E. J. Corey at Harvard, he joined Parke-Davis Pharmaceutical Research (now Pfizer Inc.) as a Senior Scientist in the Computer-Aided Drug Design group. In 1994, he moved to 3-Dimensional Pharmaceuticals, Inc., where he has focused on the development of intelligent computational tools for combinatorial chemistry, high-throughput screening and structure-based drug design. He serves on the Editorial Board of the *Journal of Molecular Graphics and Modelling* and is a co-inventor of 3DP's proprietary DirectedDiversity® technology.

F. Raymond Salemme founded 3-Dimensional Pharmaceuticals, Inc. (3DP) in 1993, and serves as President and Chief Scientific Officer. Before founding 3DP, he established and directed research groups for biophysics, computational chemistry and structure-based drug design at Sterling Winthrop Pharmaceuticals and DuPont Merck Pharmaceuticals. Before DuPont Merck Pharmaceuticals, he worked at DuPont Central Research, where he pioneered large-scale computational simulations of complex biological systems. Salemme joined DuPont from Genex Corporation, where he formed one of the first biotechnology groups to use X-ray crystallography for protein engineering. Before Genex, he was Professor of Chemistry and Biochemistry at the University of Arizona, where he specialized in studies of protein structure determined by X-ray crystallography and structural proteomics. He holds a B.A. in molecular biophysics from Yale University and a Ph.D. in chemistry from the University of California at San Diego. He is co-inventor of 3DP's proprietary DirectedDiversity® technology for combinatorial chemistry and ThermoFluor® technology for high-throughput screening and functional genomics. He has published extensively, and holds 14 US Patents. In addition to duties at 3DP, he serves on national advisory committees on advanced technology and biotechnology, including the NIST Visiting Committee on Advanced Technology.

Victor S. Lobanov received an M.S. in chemistry from Moscow State University and a Ph.D. in computational chemistry from the University of Tartu, Estonia, in 1995. After postdoctoral training with Alan R. Katritzky at the University of Florida, he joined 3-Dimensional Pharmaceuticals, Inc. (3DP) in 1996, where he serves as Assistant Director of the Research Informatics group. In his current position, Lobanov is involved in the development of advanced computational tools for combinatorial chemistry and analysis of high-throughput screening data obtained using 3DP's proprietary ThermoFluor® technology.

Dr Dimitris K. Agrafiotis
http://www.dimitris-agrafiotis.com

Online links

FURTHER INFORMATION

Chemical Informatics Societies and Professional Organizations

http://www.indiana.edu/~cheminfo/informatics/cinformsocs.html

ONLINE ONLY

3-Dimensional Pharmaceuticals

http://www.3dp.com/

SUMMARY

- A practical and cost-effective embodiment of a chemogenomics approach to drug discovery involves the following steps:
    - Gene sequences for targets that have been identified by genomics approaches are cloned and expressed as target proteins that are suitable for screening with a probe library of small, drug-like chemical compounds.
    - These compounds are screened to find active hits using a quantitative universal binding assay.
    - Initial hits or quantitative structure–activity data emerging from the binding assay are analysed and used to formulate a selection strategy for the synthesis of additional compounds with improved properties.
    - These compounds are selected from a computer database of synthetically accessible analogues of the initial probe library, constructed using verified synthetic protocols and characterized by an extensive set of computed drug-related molecular properties.
    - The selected compounds are synthesized by parallel-synthesis methods and are subsequently tested to elaborate the structure–activity profile of the target under investigation, and refine the selection criteria for additional rounds of chemical synthesis and biological testing.
    - In each iteration, priority is assigned to the synthetic candidates using a multiobjective optimization process designed to assure that compounds are not only optimized for target binding affinity, but also have drug-like characteristics that will allow them to be used directly as tool compounds in appropriate cellular or biological model systems.
- The potential for improved performance using such a strategy lies in the ability to rapidly follow up on initial hits through intelligent selection of related compounds from a computer database of synthetically accessible analogues with predefined synthesis recipes and predicted property profiles.
- To address the full spectrum of targets emerging from genomics-based efforts, it will be necessary to physically screen probe libraries that span a wide range of chemotypes and contain hundreds of thousands of compounds.
- These libraries will derive from synthetic strategies that could, in theory, produce billions of related analogues, which far exceeds the capabilities of conventional chemical-database management systems and data-modelling tools.

- Thus, the following key questions need to be addressed:
    - How can huge combinatorial libraries be generated, represented, accessed, searched and manipulated?
    - What are the most appropriate chemical-property spaces, and how can they best be computed, sampled, visualized and validated?
    - What are the most effective ways to design, execute and analyse a combinatorial-chemistry experiment?
- Successful deployment of such a system requires a new generation of computational tools that work effectively on a massive scale.