ELSEVIER

Chemical information management in drug discovery: Optimizing the computational and combinatorial chemistry interfaces

Tudor I. Oprea, Johan Gottfries, Vladimir Sherbukhin, Peder Svensson, and Thomas C. Kühler

Department of Medicinal Chemistry, AstraZeneca R&D Mölndal, Mölndal, Sweden

Structure-property relationships, central to many of today's drug discovery strategies, are not straightforward to deal with when trying to predict drug efficacy, that is, the combined outcome of target affinity, pharmacodynamic behavior, pharmacokinetic properties, and metabolic fate. In this article, we discuss the handling of chemical property *information in reagents-for-synthesis selection, enumeration, and virtual library construction. We describe the use of diversity assessment and/or experimental design in selection of compound-libraries-to-be-synthesized. Our overall objective was to identify good-quality drug candidates through reliable structure-activity relationship data, with the minimum number of compounds synthesized and tested. Chemical filters, property filters, scoring functions, and utilization of interactive visualization tools are discussed. The concept of chemical diversity and aspects of chemical space navigation employing a proprietary tool, Chemical Global Positioning System (ChemGPS), for mapping the drug-related chemical space are examined. Guidelines and workflow recommendations for the practicing medicinal chemist are proposed.*

Keywords: chemical filters, property filters, scoring functions, property-based compound selection, combinatorial library design, drug-like, interactive visualization, diversity

Color Plates for this article are on page 541.

Corresponding author: Thomas C. Kühler, Department of Medicinal Chemistry, AstraZeneca R&D Mölndal, S-431 83 Mölndal, Sweden. Tel.: +46 31 776 1362; fax: +46 31 776 3724. *E-mail address*: thomas.kuhler@astrazeneca.com (T.C. Kühler).

INTRODUCTION

The design of combinatorial or parallel libraries is a complex process.[1-3] The loose term "maximum chemical diversity," often suggested as *the* solution in this context, rests on the underlying notion that serendipity (by a sheer numbers game) would rescue what the lack of intellectual input has failed to provide.[4] Diversity, however, is not independent of context[5-13] and may vary between different discovery strategies depending on whether the main focus is lead generation or lead optimization. To inhibit a particular target in one of the protease families, for instance, a certain structural bias would be required that most likely would differ significantly from that required for a ligase in bacterial cell wall biosynthesis. Both proteases and ligases, however, ultimately are involved in chemistry dealing with the same functionality, namely, the amide bond.[14,15] Moreover, structural diversity alone does not take into account fundamental aspects such as the presence of pharmacophore(s) or acceptable pharmacokinetic and pharmacodynamic behaviors, to name a few, and as such (diversity) does not necessarily embrace what is called drug-like properties.[4]

In this article, we discuss how unbiased reagent-for-synthesis searching and virtual library construction may suggest a set of compounds for synthesis. We then elaborate on how any such set can be refined by applying computational chemistry criteria.[16-22] The objective is to secure maximum output, that is, identified drug candidates, from minimum input, that is, the smallest number possible of compounds synthesized and tested.

REAGENTS FOR SYNTHESIS AND CHEMICAL FILTERS

Selecting reagents for synthesis can be a lengthy and cumbersome process.[23-33] Even when aiming for only a small parallel

Journal of Molecular Graphics and Modelling 18, 512–524, 2000

655 Avenue of the Americas, New York, NY 10010

1093-3263/00/$–see front matter
PII S1093-3263(00)00066-8

array or library, the process of selecting starting materials may become overwhelming. The simple introduction of an alkyl side chain on a primary amine employing reductive amination conditions is shown in Figure 1. A search in Molecular Design Ltd.'s (MDL's)[34] Available Chemicals Directory (ACD) in March 1999 provided 2,441 aldehydes. Various methods can be applied to assist in determining which of these should be selected for a particular synthesis. First, one can reduce the number of potential reagents by introducing *chemical filters*,[35,36] which will exclude entries with undesired chemical moieties. For the aldehyde set, we deselect compounds having more than one aldehyde functionality, structures containing stable or unstable isotopes, entries containing metals, and other undesired functionalities.

Practice has shown that the possibility to work interactively at this stage of the operation is critical, as a query should be allowed to evolve, and sometimes even change, during the filtering process. The dynamic nature of the selection process requires that compound lists be monitored immediately in order to promptly correct possible inaccuracies in query definitions.[27] The possibility to readily move forth and back between bifurcation points separating various sublists in the filtering process would allow for this and give the required flexibility. These and other features have only recently become available in commercially accessible products.

A screen-shot of a filtering process under way employing MDL's Reagent Selector, which accommodates many of our chemical filtering requirements, is shown in Figure 2.

ENUMERATION, PROPERTY FILTERS, AND SCORING FUNCTIONS

Once the lists of suitable starting materials have been prepared, the combination of them into products is made.[20] Prior to this enumeration, *property filters*[20,36-39] may be applied to further refine the lists with respect to physical measures such as pK_a, lipophilicity, size, etc. Currently, we rely on a two-step process in which we first apply physical property filters on reagents and then apply the same filters on enumerated products. This ensures that both unacceptable extremes and potential redundancies are removed at the levels of reagents as well as products. Applying property filters on products is important because the combination of reagents could result in products with extreme properties. For instance, combining R-groups that are extremely lipophilic could result in the corresponding products becoming too lipophilic, and hence these should be deselected. Despite this, algorithms aimed at property filtering on reagents only, hence eliminating the need for double-checking enumerated products, currently are being developed in our laboratories.[40]

A third set of filters that may be applied are based on *scoring functions*. These estimate the enumerated products' drug likeness, oral availability, metabolic stability, potential toxicity, etc. General and reliable computational tools allowing one to predict these features are, however, not available. Thus far, any measure of drug-like properties has hitherto without exception been addressed by empirical approaches.[41-43] One well-documented example, the "Pfizer rule of 5," proposes certain limitations on molecular properties for a compound (drug) to be orally available: namely, a molecular weight of ≤500 Dalton, a clogP ≤5, number of hydrogen bond donors ≤5, and number of hydrogen bond acceptors ≤10.[41] Violation of any two of these criteria decreases significantly the chance of a compound being orally available.

We developed a similar set of inclusion/exclusion criteria by interrogating MDL's Drug Data Report (MDDR, drug-like) and ACD (non-drug-like) databases.[36,44] In doing so, we further refined the descriptors used in the "Pfizer rule of 5" and added more descriptors. Briefly, we consider the following to be acceptable limits for a compound to qualify as drug-like (and thus match 75% of the MDDR property distribution): molecular weight between 200 and 450 Dalton, clogP between −2 and 4.5, number of rotatable bonds between 1 and 9, number of rings ≤5, number of hydrogen bond donors ≤5, and number of hydrogen bond acceptors between 1 and 8.[36]

Any of the two sets of criteria can be used for filtering purposes, and such simple rules of thumb have the advantage of being readily understood by medicinal chemists. Employing rules based on small numbers of correlated molecular descriptors, however, may not be the best way to address the complex question of what is and what is not drug-like, and alternative approaches using statistical methods have appeared recently in the literature. One approach was based on a comparison of the World Drug Index (WDI) with ACD,[42] and another was based on a comparison of MDDR and Comprehensive Medicinal Chemistry (CMC) with ACD.[43] Some 4,000 compounds from each database were described by various molecular descriptors related to their 2D structures (MACCS-keys,[45] Ghose-Crippen atom types,[46] or chemical properties such as molecular weight, clogP, the number of H-bond donors or acceptors, etc.). Each nondrug was assigned a score of "0" and each drug a score of "1." An artificial neural network then was trained to correlate the 2D descriptors with the scores in nonlinear models. Predictions were made by calculating the same properties for new compounds as those calculated for the training set and feeding the data into the net. Scores (ideally) distributed between "0" (nondrug-like) and "1" (drug-like) were obtained as output. The models rendered good discriminating capacities; 77% of the WDI, 80% of the MDDR, and 90% of the CMC were classified as drug-like as opposed to only 10–17% of the ACD.

We used PLSDA (PLS Discriminant Analysis) implemented in SIMCA[47] to derive property-based discriminating schemes as outlined below.[48] These are based on the observation that some properties have different distributions in MDDR and ACD. For instance, 61% of the compounds in ACD have ≤2

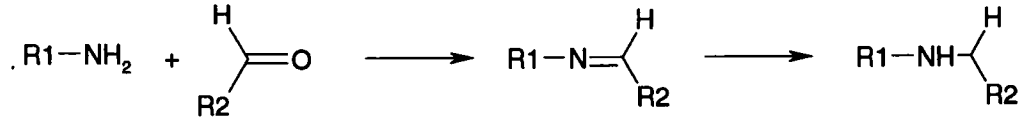

Figure 1. Introduction of an alkyl side chain on a primary amine via reductive alkylation.

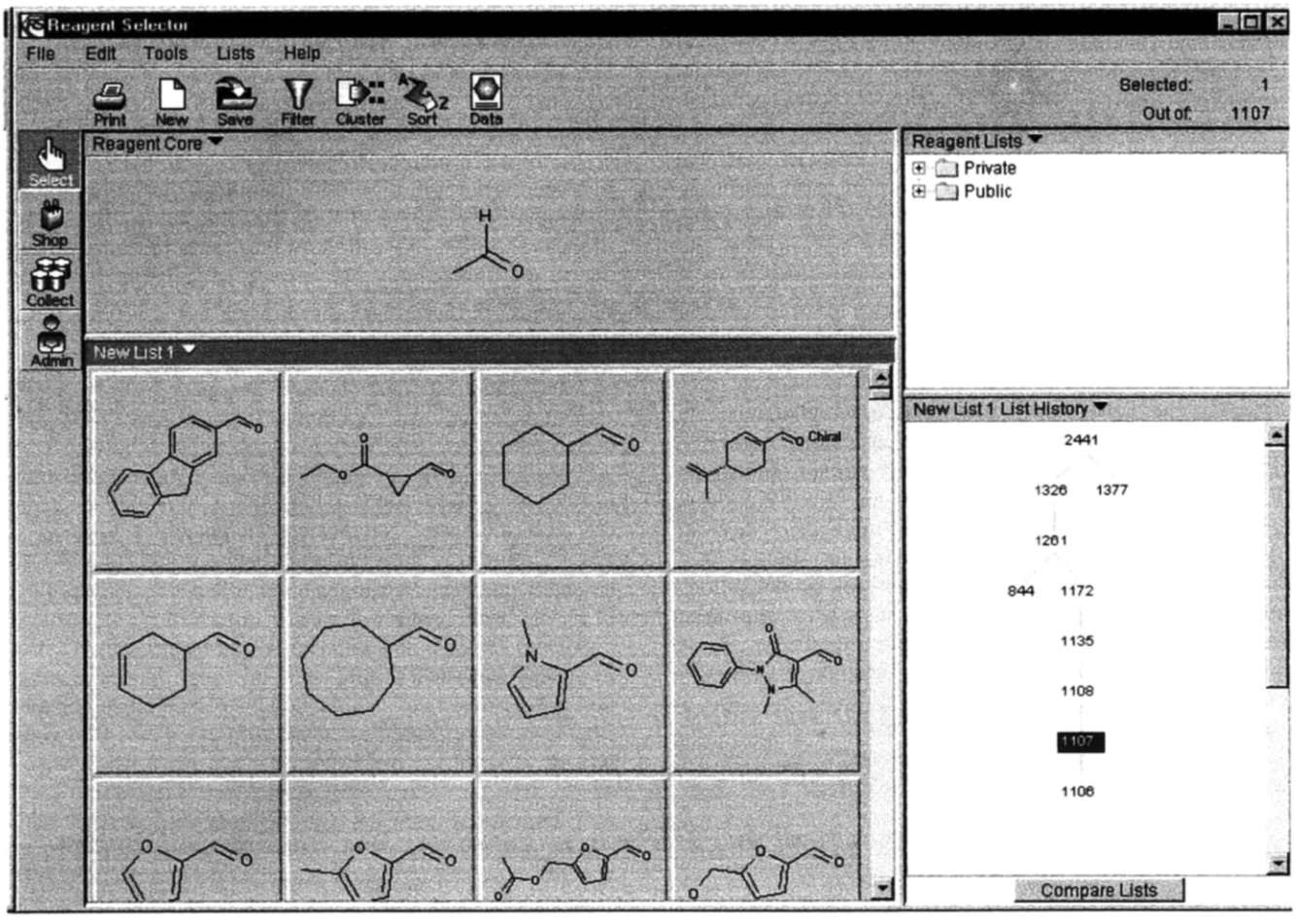

Figure 2. Screen-shot of a chemical filtering process under way using MDL's Reagent Selector. Note the List History Box on the right-hand side, which keeps track of the various sublists that have been created. The top entry states the number of reagents in the initial selection. Levels 2 through 6 from the top show the number of compounds left after different filters have been applied. Levels 7 and 8 appeared after entries had been manually deselected from the compound display box. The highlighted sublist denotes a new possible bifurcation point.

rings and ≤17 rigid bonds, whereas 66% of the compounds in MDDR have ≥3 rings and ≥18 rigid bonds. To improve the discriminating capacity we included several additional (informative) descriptors.[36]

A randomly selected training set was formed by combining 6,000 compounds from each of MDDR (drug-like) and ACD (nondrug-like). Molecular descriptors were related to 2D structures (the 4,096-bit Daylight fingerprints, DFPs),[49] or eight chemical properties and pharmacophore features (PPFs). The PPFs were molecular weight, clogP, number of rings, number of H-bond donors or acceptors, number of positively or negatively ionizable groups, and number of rotatable bonds. Each of the eight PPFs were mapped into 30-bit binary arrays in which each bit corresponded to a certain range of a certain PPF, resulting in a total of $8 \times 30 = 240$ bit binary array (Figure 3). Each bit was then treated as a separate variable in a linear model with positive or negative coefficients. The positive and negative attributes reflect a specific bit's contribution to the feature drug-like or nondrug-like, respectively.

Two different PLSDA models were built, one based on the DFPs and the other based on the eight PPFs. Their discriminating capacities for a number of databases are summarized in Table 1. Specifically, the DFP scoring function correctly classified 88% of the MDDR as drug-like and 78% of the ACD as nondrug-like. Although this scoring function accurately described MDDR, we were initially puzzled by its lower prediction rate for the ACD. This can be rationalized by the occurrence of drug-like compounds in ACD (Figure 4). The PPF scoring function demonstrated a slightly lower discriminating capacity for both MDDR (82%, drug-like) and ACD (66%, nondrug-like). Again, the lower number for the ACD could be due to drug-like compounds in this database.

Both of our scoring functions have inherent limitations related to compound description. For instance, the DFP based model does sometime give large and lipophilic compounds higher scores because individual fragments are drug-like (Figure 5). Similarly, the PPF-based model sometimes assigns compounds with reactive or intrinsicly toxic groups higher scores because properties fall in the drug-like range (Figure 6).

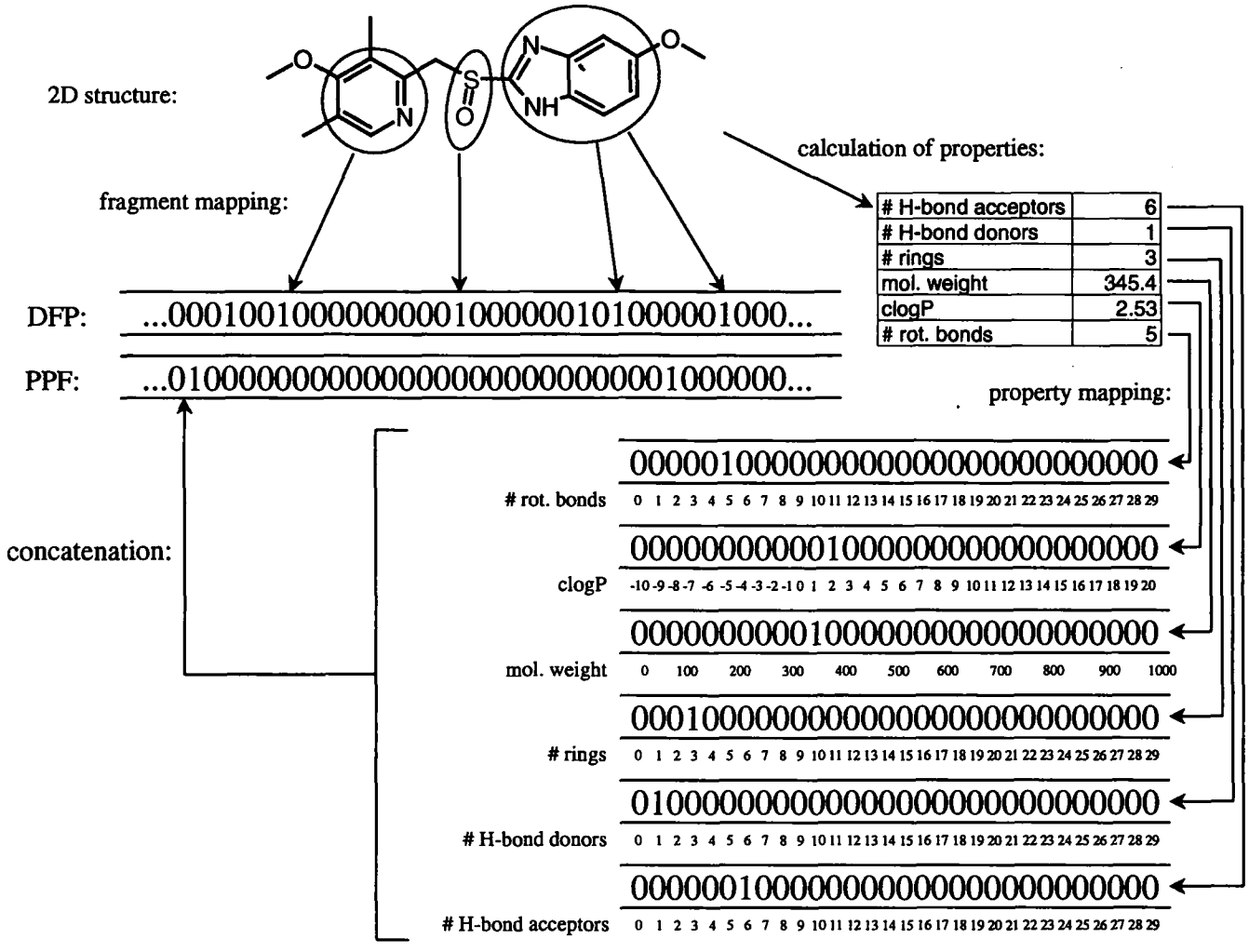

Figure 3. DFP and PPF fingerprints. The DFP algorithm examines the molecule and generates structural fragments that consist of a number of atoms and bonds connecting them. Each fragment is mapped into a 4,096-bit binary array where each bit, or a collection of bits, corresponds to the given fragment. Significant molecular properties (molecular weight, clogP, etc.) are calculated, and each of them is mapped into a 30-bit array, where each bit corresponds to the certain range of the property (for instance, molecular weight between 330 and 360 Dalton). Concatenation of individual property fingerprints gives a 240-bit property fingerprint (PPF).

Table 1. Discriminating capacities of the DFP and PPF scoring functions

Source database	Number of compounds in the validation set	Scored as drug-like by DFP (%)	Scored as nondrug-like by DFP (%)	Scored as drug-like by PPF (%)	Scored as nondrug-like by PPF (%)
CMC	7107	78	22	73	27
NCE	420	88	12	81	19
PDR	412	83	17	72	28
FALERT	15557	83	17	81	19
MDDR	83405	88	12	88	12
ACD	209978	22	78	43	57

Despite these shortcomings, the two scoring functions still provide value. They measure the similarity of compounds to existing drugs either by structural features (the DFP model) or by chemical properties (the PPFs model) and thus give two different but complementary views. We use the scoring functions either to enrich compound-libraries-

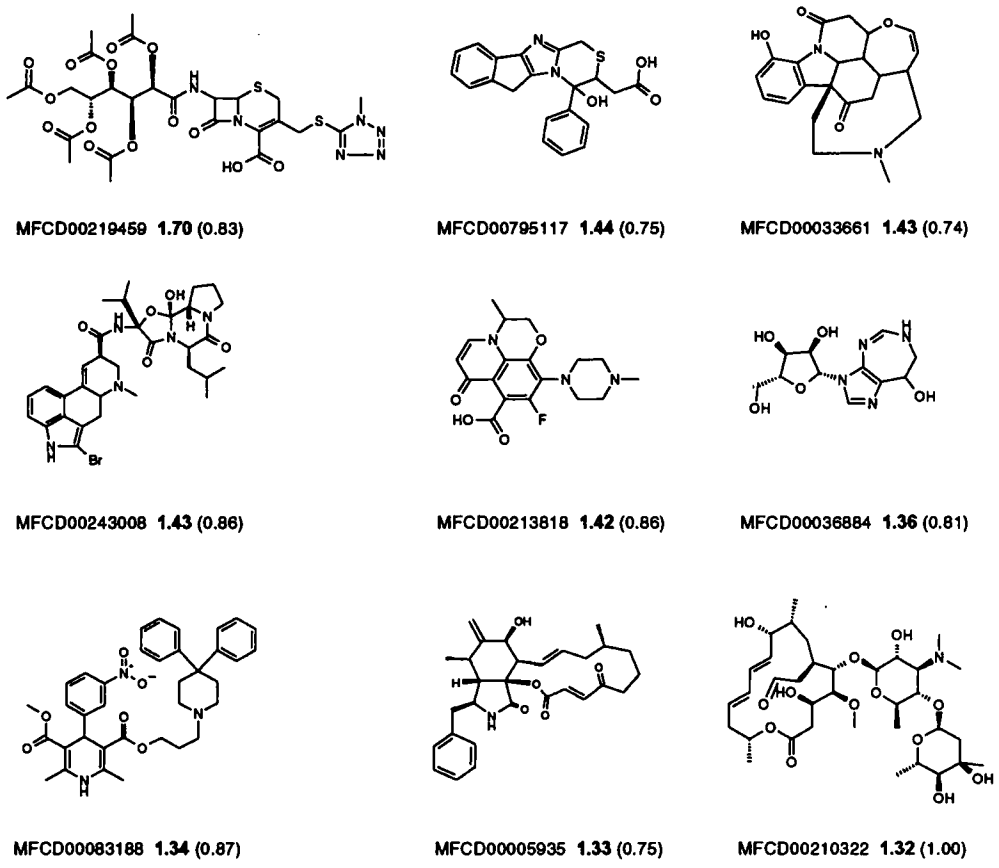

Figure 4. Drug-like structures in ACD as estimated by both the DFP and PPF scoring functions. The MFCD number, the score estimated by the DFP scoring function (bold), and the score estimated by the PPF scoring function (within parentheses) are given for each structure.

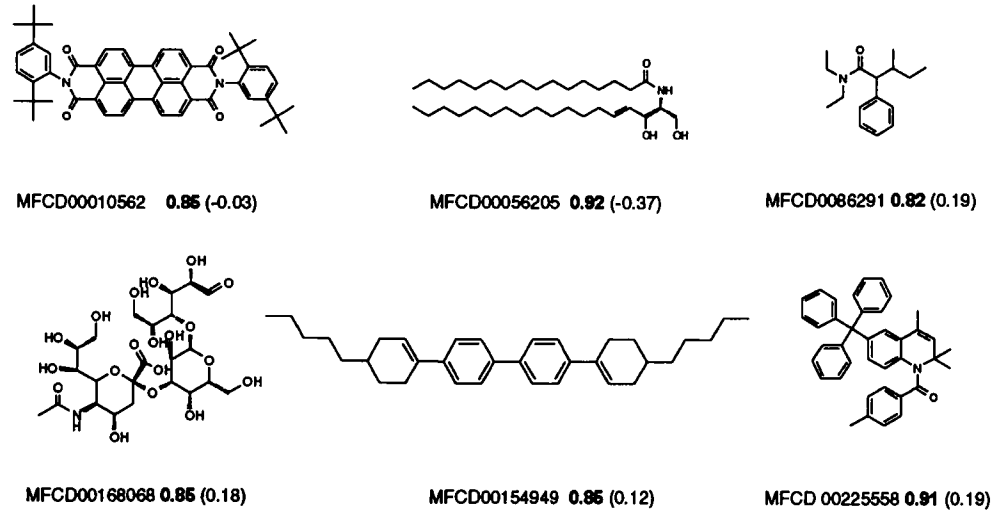

Figure 5. Structures with high DFP and low PPF scores. The MFCD number, the score estimated by the DFP scoring function (bold), and the score estimated by the PPF scoring function (within parentheses) are given for each structure.

to-be-synthesized in drug-like structures or to bias compound selections such that desired structural motif(s) become abundant in the end-library. Although these two applications are our primary interest, we also use the scoring functions for compound acquisition purposes, as they are ideally suited to assess to what degree commercial libraries contain compounds within the desired property space.[50–55]

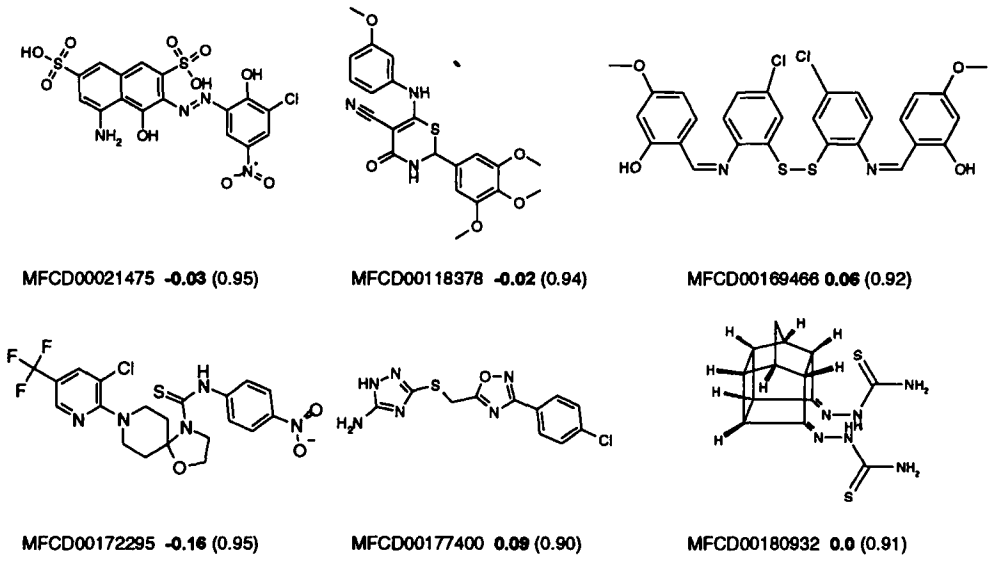

Figure 6. Structures with low DFP and high PPF scores. The MFCD number, the score estimated by the DFP scoring function (bold), and the score estimated by the PPF scoring function (within parentheses) are given for each structure.

ITERATIVE AND INTERACTIVE VISUALIZATION

Although chemical filters, property filters, and scoring functions have been discussed in sequential order, they are likely to be applied in an iterative *as well as an interactive* manner. This requires that each structure and its coordinates in the chemical space relative to the other compounds in the candidate set, as well as all possible data that may become useful in the selection process, be readily accessible and viewable. To access the compiled data, we use the application Spotfire[56] (which is based on both interactive and dynamic[57] query techniques and tight coupling between query devices[58]) together with a structure visualization plug-in connected to an MDL Isis structure database (Color Plate 1). This setup allows for quick and interpretable representations of compound selections in the multidimensional property space and allows for ready visualization of the structure associated with each data point. Another important feature is the possibility to interactively invoke both chemical filters and property filters.

The property calculations are made with a proprietary library design application that exports a principal component representation of the property space of the collection of reagents or products in question along with an sd-file that is converted into a local Isis database. The in-house application exports the original descriptors and information on which compounds were selected and optionally can export multidimensional Euclidean distances over all significant principal components to the selected compounds. The latter allows for distance filters to readily be applied between either a selected compound and its neighbors or a cluster of selected compounds and its neighbors. The compound set can be expanded further or biased with additional compounds identified by specific project criteria. The visualization application then is launched with the exported data table and the plug-in is connected to the local Isis database. Color Plates 1a and b illustrate many of the features and capabilities of this graphic interface.

This setting is well suited for interdisciplinary group sessions providing the opportunity to work iteratively on-line. Each expertise can input criteria for compound selection and incorporate alternative aspects such as synthetic feasibility, novelty, and cost, to name a few, and immediately watch the effect on, for example, chemical space coverage. We believe that in general this interaction provides an enhanced quality in our compound selection. Hence, this process may result in shortened time lines compared to the situation where each discipline works independently and communicates its selection criteria sequentially.

BIOLOGICALLY RELEVANT CHEMICAL DIVERSITY

Diversity often is perceived intuitively but is difficult to assess in an objective and consistent manner. One needs to distinguish chemical from biological information, both of which are relevant in drug discovery. We note that biologically relevant diversity is derived *a posteriori* once sufficient biological information has been collected. In contrast, chemical molecular diversity is an *a priori* analysis, because its utility is critical during compound-for-synthesis selection, that is, before the actual biological testing and information gathering process starts.

For example, chemical information can be used to categorize the two compounds in Figure 7 as catecholamines. Not knowing better, these molecules could be regarded as quite similar, the difference being only a methyl group on the amine functionality. Some of the measured properties are quite similar, too: logP = −1.06 and $pK_{a(amine)}$ = 8.60 for noradrenaline and logP = −1.37 and $pK_{a(amine)}$ = 8.87 for adrenaline. However, biological information differentiates noradrenaline from adrenaline. The former decreases heart rate (HR) and increases peripheral vascular resistance (PVR) whereas the latter does the opposite; it increases HR and decreases PVR.[59]

This biological and chemical information in combination is termed *biologically relevant chemical diversity*.[4,51,60] We de-

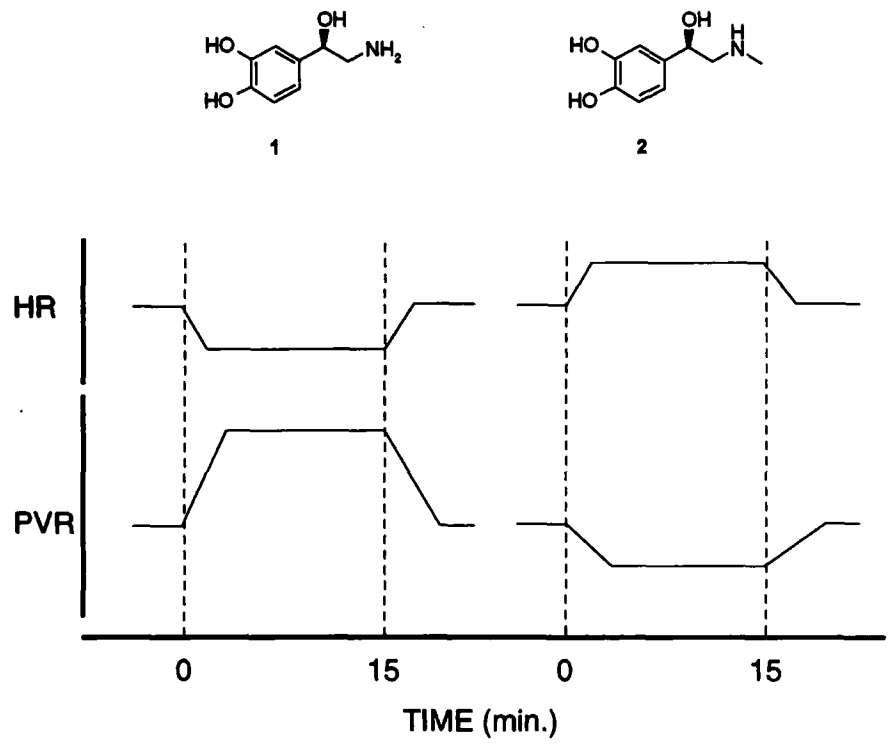

Figure 7. Structures of noradrenaline (1) and adrenaline (2). Note that the structures differ only by a methyl group. Effects of noradrenaline and adrenaline on heart rate (HR) and peripheral vascular resistance (PVR) in human beings after intravenous infusions of 10 μg/min. for 15 minutes are shown. (Modified from Allwood et al.[59])

fine this as measures reflecting similarity or dissimilarity of compound properties in a physiological environment when interacting with the biological target of interest. Although it is obvious that compounds in these regards are not adequately described by, for instance, their color in solution or their constituting atoms' ionization energies, it is equally obvious that there are no common or universal measures that could be used instead.

Thirty years of QSAR experience, however, have identified some chemical properties as more important than others in modeling biological response.[61] These include molecular and fragmental hydrophobicity, hydrogen bonding capacity, conformational flexibility, molecular shape, and solubility. The problem is that there are no reliable ways of describing, let alone computing, some of these properties with sufficient speed and accuracy. Furthermore, assessing the relative importance of, for instance, a π,π-stacking versus that of a hydrogen bond is not trivial and incurs weighting problems. Yet another complication is that solvent interactions, for example, solvation-desolvation energies, are hard to assess. New findings in this area have challenged the traditional ways that salt bridges and ion-pair interactions have been treated.[62] Hence, there is continued pressure on investigating and improving ways of describing molecules.

Furthermore, it sometimes is difficult to validate molecular descriptor improvements. To address this, we investigated whether principal components analysis (PCA) using different ways of describing biologically active compounds (from MDDR) would cluster in the property space, compounds with similar activities. Although this unpublished work shows promising results, it still needs further refinement before it can be used for general purpose validation of biologically relevant chemical diversity. We also intend to investigate if a set of experimentally measured responses, such as protein binding, water lipid-membrane partitioning, and capillary electrophoreses employing "biological" electrolytes, e.g., bile salts, glucoconjugates, and micelles, could render valid descriptors. Our set objective would be to obtain descriptions that furnish predictive models for more than one of these responses.

DIVERSITY ASSESSMENT AND EXPERIMENTAL DESIGN

One approach to rate chemical diversity is from the perspective of experimental design. For experimental design purposes, statistical methods that place essentially no constraints on experimental settings have been developed, the archetype method being factorial design. Applying the factorial design perspective to diversity assessment corresponds to selecting the compounds representing the high and low values of all the significant principal components. The compounds can be seen as occupying the corners of a multidimensional hypervolume in which each significant principal component is represented by one axis in the coordinate system.

Factorial designs, however, are not perfect from an information point of view. They do not take into account curvature to correlated responses, and linear response relationships often can be as well estimated with fewer experiments than full factorial designs.[63] Another limitation is that factorial designs cannot deal with constrained experimental hypervolumes, that is, when a certain corner or region of a hypervolume for some reason cannot be experimentally represented, for instance, an "impossible" structure that at the same time should be both small, lipophilic, and carry several hydrogen bond donors, or a reaction condition calling for simultaneous high pressure, high temperature, and an oxidizing reagent ratio that would render a mixture explosive.

Alternative approaches to compound selection are space-filling designs[3,64] and clustering methods.[65,66] Space filling is based on optimizing the coverage of the principal property space rather than maximizing the hypervolume. Clustering techniques are used widely in chemical space analysis but may not always be the method of choice for larger data sets.

To handle truncated hypervolumes, the D-optimal design algorithm is well suited.[67] It maximizes the determinant of a matrix corresponding to the constrained hypervolume.[68] In the case of which compounds should be selected for synthesis to best cover a certain experimental space, each structure in the candidate set represents a (data)point for putative selection. The D-optimal design algorithm reduces the candidate set and proposes a number of structures that represent the set with a good coverage of outskirt compounds. Thus, diversity in principal could be assessed by the numerical value of the determinant where a higher value indicates a better (more diverse) selection than a lower number that implies a worse (less diverse) selection. A larger number of structures allowed to represent the candidate set does not necessarily give a better representation but rather redundancy.

A regression based on a linear D-optimal selection would not capture curvatures to hypothetical regressors. This shortcoming may be handled by either combining the D-optimal selection with a space-filling addition of compounds or adding higher-order terms to the D-optimal model. Hence, the selection can be guided by the condition that some or all square and cross-terms of the factors should be added to the model.[69] Space-filling approaches often provide redundancy, and D-optimal algorithms often can select compounds with extreme properties further warranting the use of adequate filtering methods before any selection is started.

To improve the comparability of compound properties, we would not use the untreated candidate set matrix. It is made up of nonscaled columns and descriptors that might describe properties that are similar and hence may be correlated. Rather, we would arrive at useful property measures by subjecting the candidate set matrix to a PCA.[69] PCA provides property scores, that is, t-vectors, that are scaled, orthogonal, and often well distributed. When a property is described by multiple descriptors, for instance, when *size* is described by molecular volume, molecular weight, and molecular surface, whereas another property is described by one descriptor only, for instance, when *lipophilicity* is described by only clogP, we would use Pareto or block scaling to balance the contribution from each descriptor. In addition, we find it helpful that PCA allows ready visualization of a selection via the compounds' relative property maps, i.e., the scores (the t-vectors) and the variables' influence maps, i.e., loading plots (the p-vectors).[69]

Until now, principal property calculations have been limited to defined "localized" compound sets, such as amino acids[70] and aromatic heterocycles.[71] The current methods are well suited for occasional calculations on relatively small candidate sets, $n < 1{,}000$, but tend to become awkward for repeated work on larger selections.

Assuming that the property space of all available aldehydes had been delineated and that subsequently new members of the set had become accessible, redelineating the entire set in order to insert and accurately position a few additional structures represents a lot of work that may be hard to justify. A means of readily checking for where in the property space they would position themselves without having to redelineate the entire set would not only be helpful but also would conserve time.

A local model based on a limited compound set would allow accurate t-score predictions, that is, positionings, to be made as long as new compounds fall *within* the model's property space (interpolations) but would become increasingly unreliable as soon as new compounds fall outside that space (extrapolations). This problem is addressed by the *Chemical Global Positioning System* (ChemGPS), a proprietary tool, that we use as a standardized platform for chemical space navigation.[72] ChemGPS allows one to position compounds in the chemical property space in a consistent manner without risking the uncertainties associated with extrapolations.

CHEMGPS: THE ART OF CHEMICAL SPACE NAVIGATION

Working with large numbers of compounds (10^4–10^6) is a computational challenge, particularly when each structure is associated with one or more descriptor or property value(s). It is often the sheer size of the matrix (n rows of compounds × m columns of descriptors) that proves difficult to handle. Recursive partitioning has emerged as one of the methods designed to handle such extensive data sets.[73] Gaussian-computed molecular shape[74] is one of several methods[75] that handle chemical similarity[76] in a context that is relevant to combinatorial chemistry. Cell-based partitioning of the chemical space into multidimensional hypercubes has been used in conjunction with the χ-squared statistic and interactive cluster overlap (centered on active compounds) to yield an activity-seeded, structure-based metric for the chemical space.[50–53,77] Other methods rely on data reduction algorithms to reduce the complexity of the problem.[6,8,78] Two types of methodologies have been used to date: those that perform data compression (neural networks), and those that perform data reduction (multivariate analysis via PCA).[69] We prefer PCA because of its amenability to analysis and interpretation, and it is within this framework that we have developed ChemGPS.

In analogy to the global positioning system (GPS) for land, sea, and air navigation,[79] ChemGPS, too, requires a set of rules and a reference system. The rules define the boundaries within which calculated compound descriptors are allowed. This is the equivalent of stating that the GPS only covers land, sea, or air on planet Earth as opposed to, e.g., the solar system or the Milky Way. Which properties should be included and what their boundaries should be are open for the end-user to decide. To define drug-like space, we use the following descriptors: molecular weight < 1,500 Dalton, clogP between −6 and 15, nonterminal rotatable bonds (RTB) <30, and S, N, O, P, and X as the only elements allowed besides C and H.[80] In addition to these criteria, a set of chemically intuitive descriptors is included: size, lipophilicity, polarizability, charge, flexibility, number of hydrogen bond donors, and hydrogen bond acceptors. These parameters need not be orthogonal (independent variables), but are chosen in such a way that their meaning and interpretability remain clear to the chemist who will examine the analyzed compounds. However, the final analysis is performed using PCA on a predefined set of descriptors that is kept unchanged.[81] These are calculated once for each and every compound in the ChemGPS reference system and provide the basis for the predictive PCA model. Our experience so far has suggested that nine orthogonal dimensions are sufficient.

The ChemGPS reference system (Figure 8) comprises a set of "core structures" and a set of "satellite" structures. The core structures are representative sets of known drugs that keep the ChemGPS system focused on the drug-like chemical space. The satellite structures are intentionally chosen to be outside the drug-like chemical space and include molecules (that could be hypothetical) having extreme values in one or several of the analyzed properties while still retaining fragments present in drugs.

The positioning of novel compounds in the ChemGPS system is achieved by predicting the nine-dimensional property scores based on the previously defined set of calculated descriptors. These values are comparable across large numbers of chemicals, because they are achieved via external predictions based on the reference system. The nine-dimensional score does not change with new structures, which in turn provides a direct coordinate system for the drug-like chemical space.

ChemGPS allows one to circumvent the inherent limitations of any single local PCA model, which has the limitation of generating t-scores valid only within a certain selection of compounds. ChemGPS also provides a standardized way of describing compounds and allows one to position them in a consistent manner in the chemical property space. Although not intended for activity-seeded cluster analysis,[51] ChemGPS groups structurally and biologically related compounds into clusters and is well suited for comparing multiple compound selections (libraries), allowing one to keep track of previously explored regions of the chemical space.[72,82]

CONCLUSIONS AND FINAL REMARKS

If every conceivable compound in a drug discovery project could be synthesized and tested, then in principle all the necessary screening data (information) could be collected and

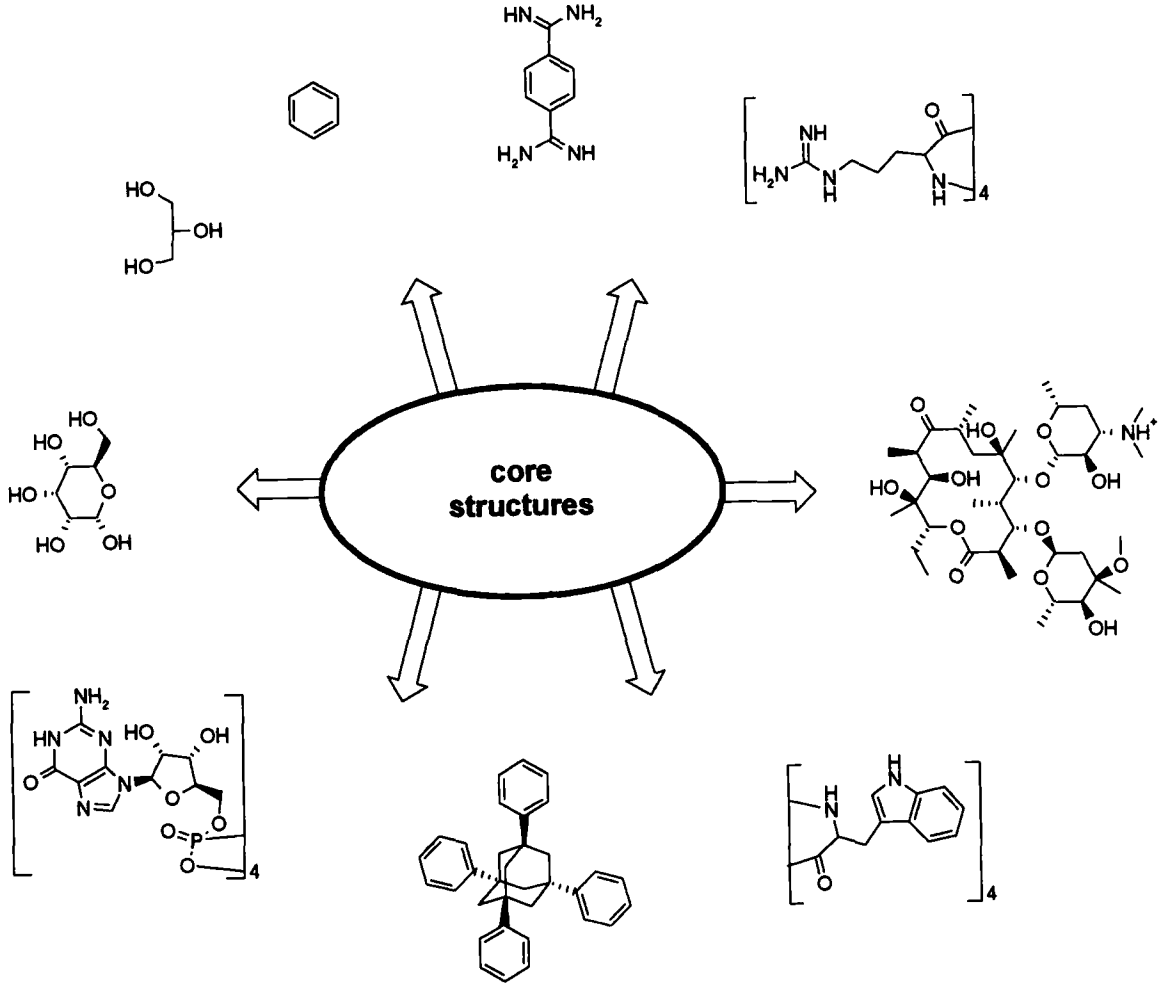

Figure 8. Schematic illustration of the ChemGPS system showing molecules that span several properties of interest. For example, polar, hydroxyl-rich molecules span the property space from allose to erythromycin, whereas hydrocarbons go from benzene to tetraphenyladamantane. In another dimension, charged species are covered from the guanosine tetramer to the arginine tetramer. Molecules with in-between properties are represented by glycerin and the tryptophan tetramer that cover an intermediate dimension between polarity and hydrophobicity. Similarly, p-amidino-benzamidine illustrates a molecule that is both hydrophobic and carries charges. The molecules in this example are deemed as "satellites," even though some are drugs, e.g., erythromycin. Note that oligomers are only schematically represented; terminal COOH, NH_2, and OH-groups are not shown.

refined into knowledge (SAR), which in turn could be used to identify a drug candidate. However, this is not practicable, cost effective, or time efficient. Therefore, we have devised a workflow by which an initial crude selection of starting materials can be distilled into a final selection of compounds to actually be synthesized, as detailed in Figure 9. Each step in the process is aimed at reducing the number of compounds to be synthesized and tested, without losing any of the initial selection's inherent and relevant information content (possible data points in a SAR).

The devised procedure comprises four major steps that allow the practicing medicinal chemist to maximize the output from any parallel or combinatorial synthesis effort. Many of the steps are, of course, used both in concert and in integrated as well as in iterative loops and are preferably applied to both reagents and enumerated products. Hence,

1. Chemical filters are applied, primarily to accommodate restrictions imposed by reaction conditions,
2. Physical property filters and scoring functions are applied, primarily to introduce any desired bias and to give the selection the right composition,
3. The selection is positioned in the "global" chemical space by ChemGPS to
   A. check that it occupies the desired "local" part of the property space, and
   B. to enable a straightforward comparison with previously positioned compounds, after which
4. The selection is subjected to experimental design algorithms, primarily to reduce further the number of compounds selected for synthesis.

Typically, the filtering and property calculations of reactants can take 1 to 2 hours. This may be followed by a multidisciplinary group session in front of a computer screen where the final reagent selections are made. This session can last 1 to 3

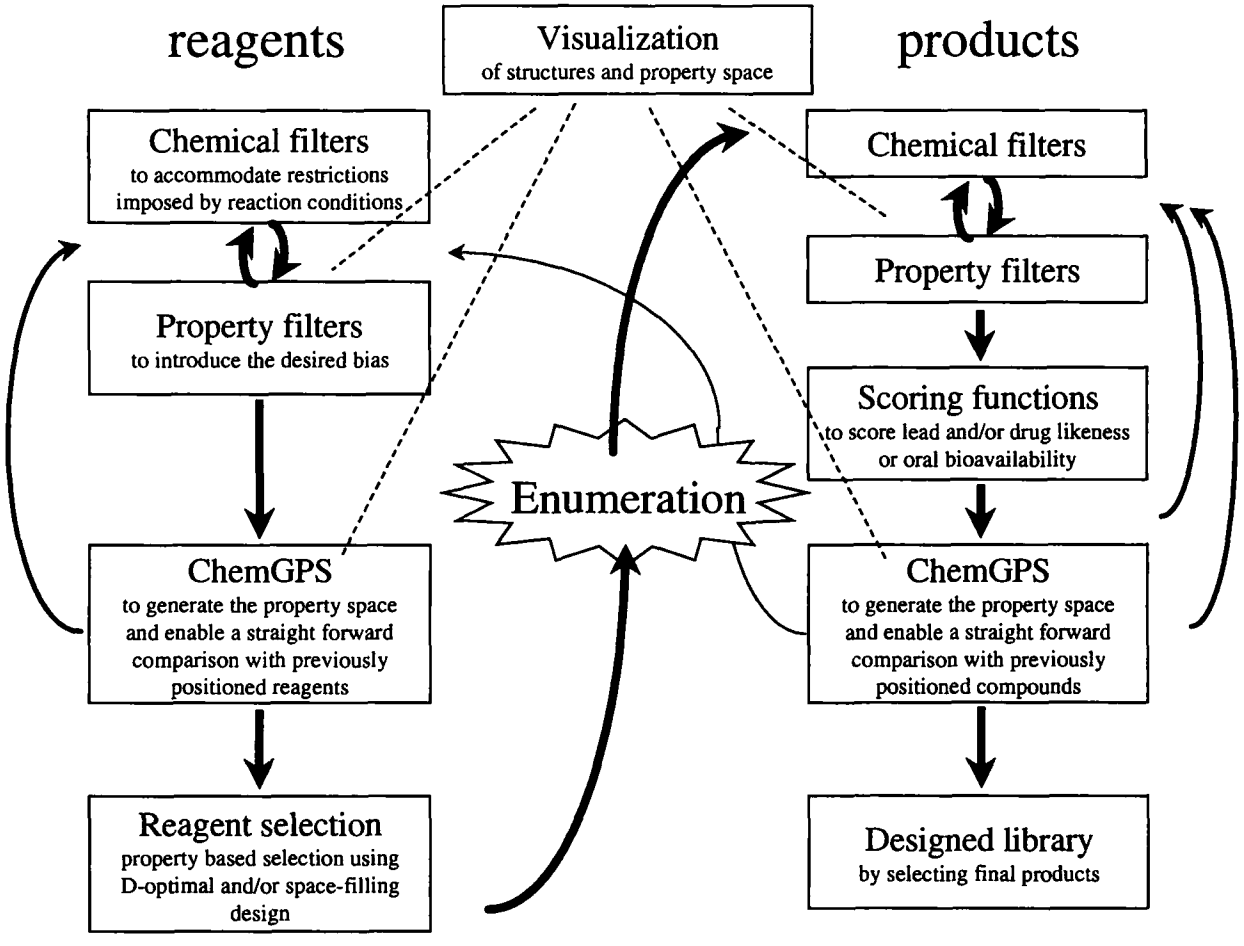

Figure 9. Proposed workflow by which an initial crude selection of staring materials can be reduced into a final selection of compounds to actually be synthesized. Note that most of the individual steps are used in iterative loops and applied on both reagents and enumerated products. Chemical filters are applied, primarily to accommodate restrictions imposed by reaction conditions, and physical property filters and scoring functions are applied, primarily to introduce the desired bias and to give the selection the right composition. The selection then is positioned in the "global" chemical space by ChemGPS to check that it occupies the desired "local" volume of the property space and to enable straightforward comparison with previously positioned compounds. Eventually the selection is subjected to experimental design algorithms, primarily to reduce further the number of compounds selected for synthesis.

hours depending on the complexity and size of the selection. After this, the enumeration is done, followed by an optional cycling of the enumerated product through the selection procedure.

ACKNOWLEDGMENTS

Special thanks are extended to our colleagues Drs. Magnus Björsne, Thomas Olsson, and Graeme Semple, AstraZeneca R&D, Mölndal, Sweden, for invaluable discussions.

REFERENCES

1 Martin, E.J., Blaney, J.M., Siani, M.A., Spellmeyer, D.C., Wong, A.K., and Moos, W.H. Measuring diversity: Experimental sedisn of cominatorial libraries for drug discovery. *J. Med. Chem.* 1995, **38,** 1431–1436

2 Gillet, V.J., Willett, P., and Bradshaw, J. Identification of biological activity profiles using substructural analysis and genetic algorithms. *J. Chem. Inf. Comput. Sci.* 1998, **38,** 165–179

3 Menard, P.R., Mason, J.S., Morize, I., and Bauerschmidt, S. Chemistry space metrics in diversity analysis, library design, and compound selection. *J. Chem. Inf. Comput. Sci.* 1998, **38,** 1204–1213

4 Reducing Time to Drug Discovery. Recent advances in solid-phase synthesis, informatics, and high-throughput screening suggest combinatorial chemistry is coming of age. *Chem. Eng. News* 1999, **March 8,** 33–48

5 Koehler, R.T., Dixon, S.L., and Villar, H.O. LASSOO: A generalized directed diversity approach to the design and enrichment of chemical libraries. *J. Med. Chem.* 1999, **42,** 4695–4704

6 Chabala, J.C. Historical overview of the developing field of molecular diversity. In: *Combinatorial Chemistry and Molecular Diversity in Drug Discovery,* Gordon, E.M., and Kerwin, J.F. Jr., Eds., Wiley-Liss, New York, 1998, pp. 3–15

7 Meyers, H.V. Chemical and biological approaches to molecular diversity: Applications to drug discovery. In: *The Biology-Chemistry Interface: A Tribute to Koji Nakanishi,* Cooper, R., and Snyder, J.K., Eds., Dekker, New York, 1999, pp. 271–287

8 Martin, Y.C., Brown, R.D., and Bures, M.G. Quantifying diversity. In: *Combinatorial Chemistry and Molecular Diversity in Drug Discovery,* Gordon, E.M., and Kerwin, J.F. Jr., Eds., Wiley-Liss, New York, 1998, pp. 369–385

9 Warr, W.A. Combinatorial chemistry and molecular diversity: An overview. *J. Chem. Inf. Comput. Sci.* 1997, **37,** 134–140

10 Hassan, M., and Waldman, M. Penalty-biased diversity: Design of diverse, druglike libraries. Book of Abstracts, 218th ACS National Meeting, New Orleans, LA, August 22–26 1999, American Chemical Society, Washington, DC

11 Jamois, E.A., Hassan, M., and Waldman, M. Evaluation of reagent-based and product-based strategies in the design of combinatorial library subsets. *J. Chem. Inf. Comput. Sci.* 2000, **40,** 63–70

12 Pavia, M.R., Hollinshead, S.P., Meyer, H.V., and Hall, S.P. Identifying novel leads using combinatorial libraries: Issues and successes. *Chimia* 1997, **51,** 826–831

13 Parks, C.A., Crippen, G.M., and Topliss, J.G. The measurement of molecular diversity by receptor site interaction simulation. *J. Comput.-Aided Mol. Design* 1998, **12,** 441–449

14 Babine, R.E., and Bender, S.L. Molecular recognition of protein-ligand complexes: Application to drug design. *Chem. Rev.* 1997, **97,** 1359–1472

15 Ward, J.B. Biosynthesis of peptidoglycan: Points of attack by wall inhibitors. *Pharmacol. Ther.* 1984, **25,** 327–369

16 Lobanov, V.S., Agrafiotis, D.K., and Rassokhin, D.N. Rational selections from virtual libraries. Book of Abstracts, 217th ACS National Meeting, Anaheim, CA, March 21–25, 1999, American Chemical Society, Washington, DC

17 Lehn, J.-M. Dynamic cominatorial chemistry and virtual combinatorial libraries. *Chem.-Eur. J.* 1999, **5,** 2455–2463

18 Gorse, D., Rees, A., Kaczorek, M., and Lahana, R. Molecular diversity and its analysis. *Drug Discovery Today* 1999, **4,** 257–264

19 Darvas, F., and Dorman, G. Early integration of ADME/Tox parameters into the design process of combinatorial libraries. *Chim. Oggi* 1999, **17,** 10–13

20 Leach, A.R., Bradshaw, J., Green, D.V.S., and Hann, M.M. Implementation of a system for reagent selection and library enumeration, profiling, and design. *J. Chem. Inf. Comput. Sci.* 1999, **39,** 1161–1172

21 Van Drie, J.H., and Lajiness, S. Approaches to virtual design. *Drug Discovery Today* 1998, **3,** 274–283

22 Barnard, J.M., and Downs, M. Computer representation and manipulation of combinatorial libraries. *Perspect. Drug Discovery Design* 1997, **7/8(Computational Methods for the Analysis of Molecular Diversity),** 13–30

23 Martin, Y.C., and Bures, M.G. Leveraging hts results to improve compound selection strategies. Book of Abstracts, 217th ACS National Meeting, Anaheim, CA, March 21–25 1999, American Chemical Society, Washington, DC

24 Snarey, M., Terrett, N.K., Willett, P., and Wilton, D.J. *J. Mol. Graphics Modell.* 1997, **15,** 372–385

25 Gillet, V., Willett, P., and Bradshaw, J. Development of bioactivity profiles for use in compound selection, Book of Abstracts, 211th ACS National Meeting, New Orleans, LA, March 24–28 1996, American Chemical Society, Washington, DC

26 Grethe, G., and Lawson, A. Reaction information requirements for the synthetic chemist. In: *Proceedings of the 1998 International Chemical Information Conference,* Collier, H., Ed., Infonortics Ltd., Tetbury, UK, 1998, pp. 154–165

27 Bronzetti, M., Gushurst, A.J., Henry, D.R., and Snyder, R.W. Reagent selector: A new tool for high througput synthesis. Book of Abstracts, 216th ACS National Meeting, Boston, MA, August 23–27 1998, American Chemical Society, Washington, DC

28 Grethe, G., and Moock, T.E. Similarity searching in REACCS: A new tool for the synthetic chemist. *J. Chem. Inf. Comput. Sci.* 1990, **30,** 511–520

29 Austel, V. Experimental design in synthesis planning and structure-property correlations: Experimental design. *Methods Princ. Med. Chem.* 1995, **2,** 49–62

30 Curran, D.P. Strategy-level separations in organic

synthesis: From planning to practice. *Angew. Chem., Int. Ed.* 1998, **37,** 1175–1196
31 Ugi, I.K. MCR.XXIII. The highly variable multidisciplinary preparative and theoretical possibilities of the Ugi multicomponent reactions in the past, now, and in the future. *Proc. Est. Acad. Sci., Chem.* 1998, **47,** 107–127
32 Sello, G., and Termini, M. Organic synthesis planning: Some hints from similarity. *Tetrahedron* 1997, **53,** 3729–3756
33 Ihlenfeldt, W.-D., and Gasteiger, J. Computer-assisted planning of organic synthases: The second generation of programs. *Angew. Chem. Int. Ed. Engl.* 1995, **34,** 2613–2633
34 Available from MDL Information Systems, http://www.mdli.com/dats/pharmdb.html. The ACD database is a compilation of over 250,000 commercially available substances from over 500 catalogs worldwide. Our ACD subset contains 194,511 structures
35 Rishton, GM. Reactive compounds and *in vitro* false positives in HTS. *Drug Discov. Today* 1997, **2,** 382–384
36 Oprea, T.I. Property distribution of drug-related chemical databases. *J. Comput.-Aided Mol. Design*, 2000, **14,** 251–264
37 Walters, W.P., Stahl, M.T., and Murcko, M.A. High-throughput "virtual" chemistry. In: *Encyclopedia of Computational Chemistry, Volume 2* von Ragué Schleyer, P., Ed., Wiley, New York, 1998, pp. 1225–1237
38 Pickett, S.D., McLay, I.M., and Clark, D.E. Enhancing the hit-to-lead properties of lead optimization libraries. *J. Chem. Inf. Comput. Sci.* 2000, **40,** 263–272
39 Teague, S.J., Davis, A.M., Leeson, P.D., and Oprea, T.I. The design of leadlike combinatorial libraries. *Angew. Chem. Ind. Ed.* 1999, **38,** 3743–3748; German version: *Agnew. Chem.*, 1999, **111,** 3962–3967
40 Linusson, A., Gottfries, J., Lindgren, F., and Wold, S. Statistical molecular design of building blocks for combinatorial chemistry. *J. Med. Chem.* 2000, **43,** 1320–1328
41 Lipinski, C.A., Lombardo, F., Dominy, B.W., and Feeney, P.J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings *Adv. Drug. Deliv. Rev.* 1997, **23,** 3–25
42 Sadowski, J., and Kubinyi, H. A scoring scheme for discriminating between drugs and nondrugs *J. Med. Chem.* 1998, **41,** 3325–3329
43 Ajay, Watlers, W.P., and Murcko, M.A. Can we learn to distinguish between "drug-like" and "nondrug-like" molecules? *J. Med. Chem.* 1998, **41,** 3314–3324
44 Shcherbukhin, V. Drug/non-drug discriminant analysis using daylight fingerprints and calculated properties. (in preparation)
45 SSKEYS, MDL Information Systems Inc., San Leandro, CA, http://www.mdli.com
46 Viswanadhan, V.N., Ghose A.K., Revankar, G.R., and Robins, R.K. Atomic physicochemical parameters for three dimensional structure directed quantitative structure-activity relationships. 4. Additional parameters for hydrophobic and dispersive interactions and their application for an automated superposition of certain naturally occurring nucleoside antibiotics. *J. Chem. Inf. Comput. Sci.* 1989, **29,** 163–172
47 SIMCA-P 8.0, Copyright 1993–1999 by Umetrics, Sweden, http://www.umetrics.com/
48 Shcherbukhin, V., Oprea, T.I., and Norinder, U. (in preparation)
49 Daylight Users Manual, Release 4.41, Copyright 1992–95 by Daylight Chemical Information Systems, Inc., Irvine, CA, http://www.daylight.com/
50 Pearlman, R.S., and Smith, K.M. Software for chemical diversity in the context of accelerated drug discovery. *Drugs Future* 1998, **23,** 885–895
51 Pearlman, R.S., and Smith, K.M. Metric validation and the receptor-relevant subspace concept. *J. Chem. Inf. Comput. Sci.* 1999, **39,** 28–35
52 Pearlman, R.S., and Smith, K.M. Novel metrics and validation of metrics for chemical diversity. Alfred Benzon Symposium. 1998, 42(Rational Molecular Design in Drug Research), 165–185
53 Pearlman, R.S., and Smith, K.M. Novel software tools for chemical diversity. In: *3D QSAR and drug design: Recent advances,* Kubinyi, H., Martin, Y., and Folkers, G., Eds. Kluwer Academic Publishers, Dordrecht, The Netherlands, 1998, pp. 339–353
54 Menard, P.R., Mason, J.S., Morize, I., and Bauerschmidt, S. Chemistry space metrics in diversity analysis, library design, and compound selection. *J. Chem. Inf. Comput. Sci.* 1998, **38,** 1204–1213
55 Holliday, J.D., and Willett, P. Definitions of "dissimilarity" for dissimilarity-based compound selection *J. Biomol. Screening* 1996, **1,** 145–151
56 Available from Spotfire, Inc., Göteborg, Sweden, http://www.spotfire.com
57 Ahlberg, C., Williamsson, C., and Shneiderman. B. Dynamic queries for information exploration: An implementation and evaluation. *Proc. ACM CHI'92: Human Factors in Comp. Syst.* 1992, 619–626. Also in Shneiderman, B. *Sparks of innovation in human-computer Interaction.* Ablex Publishing, Norwood, 1993
58 Ahlberg, C., and Shneiderman, B. Visual information seeking: Tight coupling of dynamic query filters with Starfield displays. *Proc. ACM CHI'94: Human Factors in Comp. Syst.* 1994, 313–317. Also in Baecker, R., Grudin, J., Buxton, W., and Greenberg, S. *Readings in human-computer interaction: Toward the year 2000, 2nd Edition.* Morgan Kaufmann Publishers, San Francisco, 1995
59 Allwood, M.J., Cobbold, A.F., and Ginsberg, J. Peripheral and vascular effects of noradrenaline, isopropylnoradrenaline, and dopamine. *Br. Med. Bull.* 1963, **19,** 132–136
60 Schnur, D. Designing large "smart" combinatorial libraries: Activity based validations of diversity hypotheses. Book of Abstracts, 217th ACS National Meeting, Anaheim, CA, March 21–25 1999, American Chemical Society, Washington, DC
61 Hanch, C., and Leo, A. *Exploring QSAR: Fundamentals and applications in chemistry and biology.* American Chemical Society, Washington, DC, 1995
62 Liljefors, T., and Norrby, P.-O. An *ab initio* study of the trimethylamine-formic acid and the trimethylammonium ion-formate anion complexes, their monohydrates, and continuum solvation *J. Am. Chem. Soc.* 1997, **119,** 1052–1058
63 Box, G.E.P., Hunter, W.G., and Hunter, J.S. *Statistics for experimenters.* Wiley, New York, 1978

64 Snarey, M., Terrett, N.K., Willett, P., and Wilton, D.J. Comparison of algorithms for dissimilarity-based compound selection. *J. Mol. Graphics Modell.* 1997, **15,** 372–385
65 Brown, R.D., and 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.* 1996, **36,** 572–584
66 Downs, G.M., and Willett, P. Clustering of chemical structure databases for compound selection. *Methods Princ. Med. Chem.* 1995, **3(Advanced Computer-Assisted Techniques in Drug Discovery),** 111–130
67 Johnson, M.E., and Nachtsheim C.J. Some guidelines for constructing exact *D*-optimal designs on convex design spaces. *Technometrics* 1983, **25,** 271–277
68 The SIMCA Users Manual, Umetri AB, Sweden, http://www.umetrics.com
69 Jackson, J.E. *A users guide to principal components.* Wiley, New York, 1991
70 Sandberg, M., Eriksson, L., Jonsson, J., Sjöström, M., and Wold, S. New chemical descriptors relevant for the design of biologically active peptides: A multivariate characterization of 87 amino acids. *J. Med. Chem.* 1998, **41,** 2481–2491
71 Clementi, S., Cruciani, G., Fifi, P., Riganelli, D., and Valigi, R. A new set of principal properties for heteroaromatics obtained by GRID. *Quant. Struct.-Act. Relat.* 1996, **15,** 108–120
72 Gottfries, J., and Oprea, T.I. N-Dimensional modeling of objects within a hypervolume. Patent application SE 9804127-0, 1998
73 Chen, X., Rusinko, A., and Young, S.S. Recursive partitioning analysis of a large structure-activity dataset using three-dimensional descriptors. *J. Chem. Inf. Comput. Sci.* 1998, **38,** 1054–1062
74 Nicholls, A. Method and apparatus for evaluating molecular similarity in pharmaceutical drug discovery and design. U.S. Patent WO 9944055, 1999
75 Bures, M.G., and Martin, Y.C. Computational methods in molecular diversity and combinatorial chemistry. *Curr. Opin. Chem. Biol.* 1998, **2,** 2376–2380
76 Willett, P., Barnard, J.M., and Downs, G.M. Chemical similarity searching. *J. Chem. Inf. Comput. Sci.* 1998, **38,** 983–996
77 Schnur, D. Design and diversity analysis of large combinatorial libraries using cell-based methods. *J. Chem. Inf. Comput. Sci.*, 1999, **39,** 36–45.
78 Li, J., Murray, C.W., Waszkowycz, B., and Young, S.C. Targeted molecular diversity in drug discovery: Integration of structure-based design and combinatorial chemistry. *Drug Discovery Today* 1998, **3,** 105–112
79 http://gps.laafb.af.mil
80 The cut-off criteria for the ChemGPS parameters were deliberately set outside the known drug-like limits for these parameters,[36,41] in order to ensure that the vast majority of compounds of interest would be encompassed by these values. By choosing molecules that have such extreme values, one can effectively place "satellite" molecules outside the drug-like space.
81 These descriptors include, for example, the heteroatoms count, the Kier and Hall topological descriptors, clogP and CMR (available from Daylight CIS), as well as simple Hückel-type of molecular orbital calculations.
82 Oprea, T.I., and Gottfries, J. Chemography: The art of navigating in chemical space. *J. Comb. Chem.* (submitted)