Perspective

Diverse Viewpoints on Computational Aspects of Molecular Diversity

Yvonne C. Martin

J. Comb. Chem., **2001**, 3 (3), 231-250• DOI: 10.1021/cc000073e • Publication Date (Web): 08 March 2001

JOURNAL OF combinatorial CHEMISTRY

Volume 3, Number 3 May/June 2001

Perspective

Diverse Viewpoints on Computational Aspects of Molecular Diversity

Yvonne C. Martin

Computer Assisted Molecular Design, Abbott Laboratories, Abbott Park, Illinois 60064-3500

Received August 25, 2000

The following reminiscences are as diverse as the molecules designed from diversity principles. The scientists invited to participate had contributed to the molecular diversity literature by 1996—not all invited accepted the invitation. Sources of inspiration for the calculation of diversity vary from techniques for document searching, classical and 3D QSAR, graph theory, and chemical reaction planning. The contributors each weave a fascinating story that reveals the sources of their ideas. However, together all testimonies present an even more fascinating story of how modern science borrows ideas from many disciplines and emerges with a slightly different flavor from each group.

The challenge of molecular diversity is the realization that there are more potentially useful molecules than there are atoms in the universe.[1] Molecular diversity can be used to design "random" or targeted combinatorial libraries, to select "representative" subsets of a compound collection for screening, or to focus and explode a traditional one-by-one synthetic program. In all cases the question is the same: How does one select the molecules on which to concentrate? How will these molecules be represented to the computer? How many molecules are needed to represent all potential molecules—or is this even a realistic question?

While most of the reminiscences that follow concentrate on the pharmaceutical uses of molecular diversity, that by Gasteiger provides a reminder that the concepts of diversity and similarity of molecules can be extended to diversity and similarity of reactions. This is a little-explored area, but one that could have a large impact on the experimental side of combinatorial chemistry.

The Emergence of Concepts of Molecular Similarity and Diversity

Peter Willett

Background. After first receiving a degree in Chemistry from Oxford University, I went to the University of Sheffield to take their 1 year MSc Information Science program. I stayed there to do a Ph.D. on methods for indexing databases of chemical reactions,[2,3] this requiring the development of ways of measuring the degree of structural similarity between pairs of molecules. As a result of this, I became interested in the concept of similarity more generally, and thus read widely about the similarity and clustering techniques that were already under active investigation by researchers in textual information retrieval (IR) but that were not then (the late 1970s) familiar to the chemoinformatics community. At the time, these techniques were limited to very small amounts of data, and I thus started a postdoctoral study that sought to develop methods that could be applied to large document databases. In fact, I did not get very far with the work before I was offered a faculty position and had to start thinking about preparing lecture courses and all the other responsibilities of a new lecturer; however, it did mean that I had by then gained an appreciation of the need for clustering methods that were both effective (i.e., could achieve the

10.1021/cc000073e CCC: $20.00
Published on Web 03/08/2001

desired results in the domain of interest) and efficient (i.e., achieved these results in a minimal amount of time).

In the early 1980s, I started again to look at methods for document clustering, focusing upon the hierarchic agglomerative methods that had already attracted some attention from the IR community (see ref 4 for an extended review of work in this area). Early IR research had focused upon the well-known single-linkage agglomerative method: our studies investigated the use of the other agglomerative methods (such as complete linkage and group average) for document clustering. The work demonstrated that these other methods were generally superior to the single linkage method but that none of them were noticeably more effective than simple, unclustered text searching, while being extremely demanding of computational resources.[5−7] Shortly after this work had started, I began an analogous program of research to investigate the use of clustering methods for chemical databases, using the experience gained in the text domain as a starting point. As we shall see, this proved to be a much more productive application of cluster analysis.

The rationale for this second research program was my realization that there are clear similarities in the ways that chemical and textual database records are characterized, and hence in the ways that those records can be processed. For instance, the documents in a text database are each typically indexed by some small number of keywords, in just the same way as 2D or 3D structures in a chemical database are each characterized by some small number of substructural features chosen from a much larger number of potential attributes. Moreover, both types of attribute follow a well-marked Zipfian distribution, with the skewed distributions that characterize the frequencies of occurrence of characters, character substrings, and words in text databases being mirrored by the comparable distributions for the frequencies of chemical moieties. Finally, in just the same way as a document either is or is not relevant to some particular user query, so a molecule is or is not active in some particular biological test, thus allowing comparable performance measures to be used to assess search effectiveness in the two types of retrieval system. I have discussed the close relationship that exists between these two types of database processing elsewhere[8] (although I must emphasize that it was my colleague, Michael Lynch, who first brought this home to me as he had already been studying both chemical and textual information retrieval for several years when I first joined the Department in Sheffield[9]).

Comparison of Similarity and Clustering Methods. We took as our basis for the chemical clustering work the, in my view, seminal paper by Adamson and Bush,[10] which was the first to suggest that the simple fragments that were already starting to be used in screening systems for chemical substructure searching could also be used to quantify the degree of structural similarity between pairs of 2D structures. Specifically, Adamson and Bush introduced the idea that one could measure the resemblance between a pair of molecules by using an association coefficient based on the number of fragments common to their bit-strings. We developed this idea in a Ph.D. study that was carried out in collaboration with Pfizer Central Research in Sandwich, U.K., with the aim of developing an automated way of selecting compounds for biological testing, as an alternative to the existing, and highly time-consuming, manual selection methods that were employed when these initial studies were carried out in the early 1980s. Specifically, the idea was to cluster a corporate compound file into a set of clusters and then to pick one compound, or some small number of compounds, from each cluster in turn for submission to the biological test of interest. If a selected compound proved to be active, then the other compounds in its cluster would also be submitted for testing: a simple idea now but one that was not at all common at the time. To achieve this end, we first needed to identify an effective similarity measure, i.e., one for which a high level of structural similarity would imply a high level of property or activity similarity; next, to identify a clustering method that could process the resulting intermolecular similarities to yield meaningful structural clusters; and then to ensure that these methods were also sufficiently rapid in operation to enable them to be applied to data sets of nontrivial size. The last requirement was of particular importance since most of the important clustering methods were first developed for applications in numerical taxonomy where only a few tens of objects needed to be clustered.[11]

A problem that has to be faced in any comparative study of similarity (and/or clustering) methods is the need for an evaluation criterion that is external to the data used for the generation of the similarities (or clusters) in the first place. Adamson and Bush suggested that if the compounds under study had associated properties or activities then this information could be used to evaluate a similarity or clustering method by means of a simulated property prediction experiment. Thus, one could take each compound in turn and assume its property is unknown, then calculate its similarity to each of the other molecules in the data set using the similarity measure under investigation, and obtain a predicted property value by taking the observed value for its nearest neighbor (or neighbors). If this is repeated for each member of the data set in turn, then one can correlate the predicted values with the original experimental data and assess the overall effectiveness of the similarity measure by the extent of this correlation. A similar procedure can be used to evaluate clustering methods, with the predictions in this case being based on the compounds occupying the same cluster as the compound for which a prediction is required. The basis for such ideas is, of course, the similar property principle, which was subsequently popularized by Johnson and Maggiora in their influential monograph on molecular similarity.[12]

Our detailed comparative tests[13] showed that the use of the Tanimoto coefficient with simple fragment incidence data provided an effective measure of structural similarity, and we then considered how this finding might be applied to database searching. Our original intention had been to provide a simple way of ranking the outputs of recall-oriented substructure searches, but we were able to demonstrate that such rankings could also be generated for searches of an entire database, thus resulting in one of the first operational systems for chemical similarity searching;[14] an analogous system had been described just a few months previously by

workers from Lederle Laboratories.[15] Having identified the Tanimoto coefficient as an appropriate similarity measure, we then moved on to consider the various types of clustering method that could be used for the grouping of chemical data sets. Over a period of some years, we evaluated some 30 different clustering methods (including hierarchic agglomerative and divisive methods, relocation methods, and single-pass and near neighbor methods) by means of simulated property prediction as discussed previously. We concluded that the method due to Jarvis and Patrick was the most suitable of those tested, and we subsequently demonstrated its use on subsets of the Pfizer corporate file.[16]

How Does the Work Stand Up Now? The experiments summarized above comprised much of the book that I wrote in 1987 on chemical similarity and clustering,[17] and the Tanimoto coefficient and the Jarvis−Patrick clustering method have subsequently become the tools of choice in many applications of molecular similarity analysis. Given the rate at which chemoinformatics has developed over the past few years, it is perhaps surprising that these methods are still widely discussed.

The Tanimoto coefficient continues to be used extensively for the analysis of fragment bit-string data, despite some inherent limitations that can produce counterintuitive similarities in some cases;[18−20] that said, many of the other association coefficients that might be used for this purpose are similar in form and would probably exhibit comparable behavior. The situation with regard to the Jarvis−Patrick method is rather different. Our comparative property-prediction experiments identified Ward's hierarchic agglomerative method as the most effective, but with Jarvis−Patrick performing almost as well. At the time that these comparisons were carried out, computer limitations (in terms of both raw CPU speeds and the clustering algorithms available) meant that Ward's method could not be applied to chemical databases of substantial size and we hence chose Jarvis−Patrick for such applications. This was rapidly adopted as the clustering method of choice in commercial chemical database software, not only to select compounds for random screening but also for applications such as clustering the outputs of substructure searches that retrieve very large numbers of molecules. However, the method does have limitations,[20] and subsequent comparisons, both by us[21] and by others,[22] have reaffirmed the general superiority of Ward's method. The availability of improved computer hardware and of the efficient reciprocal nearest neighbors algorithm[23] means that this method can now be applied to databases containing some hundreds of thousands of molecules in an acceptable amount of time, and the method is thus becoming available in commercial chemical database software.

With the need to select diverse sets of compounds that has arisen over the past few years, it is hardly surprising that clustering has enjoyed a renaissance of interest, although there are, of course, several other approaches that have now been developed for this purpose, including dissimilarity-based, cell-based, and optimization-based methods.[24] Even so, the grouping of objects is a very natural human procedure, and I would thus expect there to be continuing interest in cluster-based approaches, not just for diversity analysis but also for database processing more generally. In particular, there is considerable interest in clustering in the rapidly growing area of data mining, and it is possible that research here will provide new algorithms and data structures of potential benefit to the chemoinformatics community.

More Recent Work. The work described above was done quite some time ago and long before developments in combinatorial chemistry resulted in the current interest in methods for diversity analysis. We became involved in this as an indirect result of the development of 3D pharmacophore-searching systems. These often yield very large outputs in response to user pharmacophoric patterns, and this led us to reconsider our previous work on grouping 2D substructure search outputs, where the output is clustered and where a representative subset is chosen for review by selecting one molecule from each cluster in turn. Thus, one obtains a representative set of dissimilar molecules by first identifying all of the similar molecules (i.e., those in the same clusters), which seems a rather roundabout way of doing things. We were hence attracted to dissimilarity-based approaches to compound selection, where one tries to identify a maximally dissimilar subset of the molecules in the database directly rather than indirectly via an initial clustering. The general algorithm for this approach, as first described by Bawden[25] and Lajiness,[26] involves selecting a compound at random and then iteratively choosing that previously unselected compound that is most dissimilar to those that have already been selected.

The Bawden−Lajiness algorithm for dissimilarity selection is very simple in concept but has an expected time complexity of order $O(n^2N)$ for selecting an n-compound subset from an N-compound data set, and it might hence be too time-consuming for online subset generation from large 3D search outputs. It was at this point that we turned again to work on hierarchic document clustering. The various hierarchic methods differ only in the precise criterion that is used to measure the similarity between two clusters of documents at each stage in the creation of the hierarchy. The appropriate criterion for the intercluster similarity in the group-average method is the average of all of the pairwise interdocument similarities, where one document is in one of the two selected clusters and the other document is in the other cluster. Voorhees[27] demonstrated that precisely the same average similarity could be obtained from a procedure that involved just a single similarity calculation using the weighted centroids of the two clusters. This elegant equivalence provides a highly efficient way of implementing the group-average method; however, we realized that it can also be applied much more generally to any situation where sums of similarities, rather than individual similarities, are required; where the cosine coefficient is used to measure the similarity between pairs of objects; and where the objects that are being compared are characterized by some form of vector-like representation (such as fragment bit-strings). Specifically, Voorhees' result can be used to choose the most dissimilar molecule in the selection step of the Bawden−Lajiness algorithm; if by "most dissimilar" we mean that molecule with the largest sum of dissimilarities to the molecules that have already been chosen.[28] This results in an algorithm with

an expected time complexity of $O(nN)$, and the procedure can hence be used to select subsets not just from 3D search outputs but also from entire databases, a requirement that was first becoming apparent at around this time.[29]

Although other selection algorithms have since been shown to be superior,[30−32] the sum-based algorithm provided the starting point for our work on comparing reagent-based and product-based approaches to compound selection. Specifically, the Voorhees' equivalence can be used to provide an extremely rapid way of computing the diversity of a set of compounds,[33] a fact that we were able to use subsequently in the fitness function of a genetic algorithm (GA) that we developed for dissimilarity-based compound selection.[34] This is sufficiently fast to enable the selection of combinatorial libraries not only from the sets of reagents in a combinatorial synthesis but also from the fully enumerated set of products for that synthesis, thus permitting a direct comparison of the diversities of the two types of library. Our initial results demonstrated the superiority of product-based approaches, in terms of yielding more diverse libraries, and this has since been confirmed in subsequent, more detailed comparisons.[35,36]

I'll finish this personal account as I started, by noting another link between IR and chemoinformatics, specifically a GA-based approach we developed to optimize the weights associated with a set of descriptors. The basic idea was first published in a paper describing the calculation of weights that represent the ability of keywords to discriminate between relevant and nonrelevant documents in searches of text databases.[37] While this study was in progress, we developed an analogous GA to calculate weights that represent the ability of physicochemical properties to discriminate between drug and nondrug compounds.[38] Such drug/nondrug studies are now common,[39,40] but it proved quite difficult to persuade referees that this was an appropriate thing to try to do when our work was first submitted for publication. Most recently, we have incorporated these properties in our product-based selection algorithm so as to design libraries that are both druglike and diverse,[41] and current work is considering ways in which we can still further increase the effectiveness of such integrated selection procedures.

The Emergence of the Concept of Molecular Diversity at Pharmacia

Michael Lajiness, Mark Johnson, and Gerry Maggiora

The Roots in Molecular Similarity. The motivation to investigate and develop the concept of molecular diversity grew out of an even earlier and broader motivation to explore the concept of molecular similarity. Although the latter concept has many diverse seeds, the seed at Upjohn[†] was planted when, in the mid 1980s, Mark began seeking a QSAR formalism free of the congeneric restrictions associated with Hansch[42] and Free-Wilson analysis[43] and the high dimensional modeling operations of pattern recognition.[44] Mark's work led to the development of a graph-theoretic metric for measuring the similarity between two compounds based on their maximum common substructure.[45] These results, however, turned out to be of more theoretical than practical interest. Two reasons relevant to the present discussion were the difficulty of developing suitable software for implementing the ideas and the focus on analyzing the small data sets associated with QSAR studies at that time. Still, the idea grew that a similarity-based formalism would find a place in drug discovery.

When Gerry joined Upjohn as Director of what was then called the Computational Chemistry Unit in 1985, he took an immediate interest in molecular similarity as a concept that merited further exploration in terms of possible applications in drug discovery. The paper by Willett and Winterman[13] had a big impact on our discussions. It introduced us to the work being carried out in chemoinformatics, in general, and Peter's group, in particular. This study clearly corroborated the intuitive principle that similar structures tend to have similar properties. It also demonstrated that some plausible molecular similarity measures work better than others do.

Shortly thereafter, Mic joined the CADD unit. He was responsible for working with scientists in getting data in to and out of Cousin, our "home grown" chemical-informatics database management system, and in bringing in and developing tools for doing so. Unlike the SAR focus that characterized many of our earlier discussions, Mic's perspective derived from his experiences with handling the large amount of chemical and screening data that resided in Cousin. When Subhash Basak[46] presented a seminar on molecular similarity at Upjohn, Mic recognized a similarity tool immediately relevant to his interests. With Mic on board, Gerry organized a molecular similarity team with Tom Hagadone, who had programming responsibilities for Cousin, as an ad hoc advisor. An important occurrence, which helped to further crystallize much of our thinking on molecular similarity, came from our participation in the 1987 International Chemical Structures Conference held at the Leeuwenhorst Congress Center in The Netherlands. Molecular similarity was now an idea ready to happen at Upjohn, but it was still not readily accepted within the Pharmaceutical Research Division of the company.

Development of Molecular Similarity at Upjohn. Initially, most of our effort was directed at implementing similarity searching of the Cousin database. The program POLY, developed by Subash, provided a quick solution. POLY computed more than 90 topological indices that could be reduced to 10 principal components without losing significant intermolecular distance information. These 10 variables were easily stored in the Cousin system and provided our first similarity searching capability. At that time there was considerable debate concerning the "meaning" of these topological indices. For us they provided a new tool that could be used by medicinal chemists to find sets of similar compounds in large chemical databases—practically speaking, that was all that mattered.

During the early days, before the field of molecular similarity had gained full status as a legitimate area of chemical research, the terms "molecular similarity" and,

[†] The Upjohn Company is now incorporated into the Pharmacia Corporation as a result of two mergers since 1995, but it was the name of the company under which the work discussed here was carried out.

consequently, "molecular dissimmilarity" or "molecular diversity" did not appear in titles, keyword lists, or abstracts. Rather, work for which such terms might now serve as an umbrella term was disparate and dispersed in unrelated journals. Gerry broached the idea of organizing an ACS minisymposium to bring the relevant investigators together to present their work as it related to molecular similarity as a concept of interest in its own right. The minisymposium was held in 1988 and resulted in the first volume of collected writings on molecular similarity.[12] Interestingly, the word "similarity" that heretofore had been largely absent in scientific reviews sessions at Upjohn, now began to be heard with great regularity—the concept had become part of the scientific culture at Upjohn!

Dissimilarity Selection Slowly Takes Root. Another line of development was directed toward increasing the efficiency of compound screening at Upjohn. At the time it was quite popular to screen every compound in the proprietary collection. As high throughput screening (HTS), let alone ultra-HTS, methods were yet to be developed, it took considerable time to screen the thousands of compounds in our collection—sometimes from 6 months to a year! Thus, the idea that a relatively small subset of our entire compound collection could be obtained that had a higher probability of containing active compounds than would be the case with a random sample of comparable size was very appealing. Willett, Winterman, and Bawden[16] proposed a nonhierarchical clustering approach based on molecular fragments as a means for selecting a representative subset of compounds for screening. As our compounds were represented by only 10 principal component scores, we could accomplish a nonhierarchical clustering using PROC FASTCLUS.[47]

Although we found the rationale for screening diverse collections of compounds quite compelling, Mic recalls having to cajole project teams to use "dissimilarity thinking" when selecting compounds for screening. Nor was dissimilarity thinking as a means of selecting compounds for screening quickly accepted by the QSAR community. When we presented our work at the 1988 European QSAR Conference,[48,49] there were few positive reactions; most of the excitement at the conference was directed toward emerging developments in 3D QSAR. Mic also recalls that at the QSAR Gordon Research Conference the following year only a handful of individuals attended his poster on dissimilarity-based compound selection, even though an oral session on molecular similarity was well received.

Emergence of Diversity as a Concept in Its Own Right. Things changed dramatically with the advent of combinatorial chemistry and HTS technologies. As time went on, more and more people became accustomed to the ideas behind dissimilarity thinking. Soon it became accepted, even expected, at Upjohn to first examine dissimilar or diverse collections of compounds in a new assay. In the early 1990s dissimilarity-based selection tools were used to enhance the diversity of our proprietary compound collection by identifying commercially available compounds for purchase that were structurally dissimilar to compounds already in our collection.[50,51] It is our recollection that the term "dissimilarity" gave way to "diversity" primarily in response to the arrival of combinatorial chemistry and library design.

Dissimilarity became mainstream at Upjohn in the mid 1990s when it was integrated into our Cousin chemical-information system. Scientists now had access to an easy-to-use tool that could be used to select structurally diverse subsets of compounds from a much larger set. This proved useful for browsing substructure search results, selecting reagents for combinatorial chemistry, and for selecting active compounds for secondary screening. Currently, dissimilarity or diversity tools remain very much in use within Pharmacia, and we continue to evaluate new ways for characterizing the diversity of compound collections and combinatorial chemistry libraries.[52,53]

It is interesting how far our industry has come in adopting dissimilarity thinking and the associated tools from a point where almost no one considered diversity or dissimilarity advantageous to a point where virtually everyone uses diversity tools to purchase compounds, design libraries, or evaluate collections—even whole conferences have been devoted to the subject of diversity!

Diversity of Combinatorial Libraries: The Challenge at Chiron

Eric Martin

Early Design of Combinatorial Libraries. Recalling my initial work in combinatorial library design, I am struck by how the solution to nearly every difficulty was fortuitously supplied by someone else. First came the problem of combinatorial library design itself. When I presented the first poster on "Combinatorial Library Design" at the 1993 QSAR Gordon Conference and published the corresponding papers,[54,55] I believe I was simply the first computational chemist to be asked to design a combinatorial library from a very large set of potential building blocks. Since combinatorial chemistry was originally applied to biopolymers, the number of possible building blocks had always been small. Since there are only 8000 coded tripeptides, one could easily make them all by parallel synthesis. Even using all commercially available protected amino acids gave just $60 \times 60 \times 60 = 216\,000$ tripeptides, which could all be made as 3600 pools of 60 by split resin synthesis. But in 1992, following a suggestion by Steve Kent, Ron Zuckerman began robotically producing peptoids by submonomer synthesis.[56] Suddenly, at least 1000 readily obtainable low molecular weight amines were available as building blocks. It was impossible to make 10^9 tripeptoids, and he could no longer defer to nature for a small privileged subset. Thus, an advance in library synthesis produced the corresponding practical problem of combinatorial library design at Chiron.

The notion of molecular diversity was already commonplace. However, I needed rigorous tools to measure and maximize molecular diversity, and I wanted something more continuous and quantitative than cluster analysis. Over lunch at a previous Gordon Conference, Nouna Kettaneh-Wold had acquainted me with D-optimal design for selecting substituents for QSAR regression modeling. While my goal was not to fit a regression model, I did want a subset of points

that covered the full dimensionality of a property space as nonredundantly and orthogonally as possible. Since D-optimal design maximizes the variances within properties and minimizes the covariance between properties, it was ideal for this purpose. Thus, Nouna had provided me with a molecular diversity measure, which commercial programs could optimize, that was perfect for library design.

Now that every available amine was a potential building block, I could no longer use precompiled Hammett-type parameters to create the property space.[57] I hoped, instead, to combine topological indices[58] as overall molecular shape descriptors with Daylight fingerprints[59] as local chemical functionality descriptors. Thus, I needed somehow to combine a 2048-D binary fingerprint space with a continuous topological index space to produce a single, small, coherent, Euclidean property space. I knew it was not valid to simply perform principal components analysis (PCA) on the binary space, but I knew of no statistical tool to convert the binary space into an Euclidean space of minimum dimension. Coincidentally, I was at the very same time proofreading a chapter on distance geometry for Jeff Blaney and Scott Dixon.[60] It was several days before it occurred to me that these were the very same problem! The solution was to use the Daylight fingerprints to construct a Tanimoto similarity matrix; then apply multidimensional scaling (MDS), the statistical analogue of distance geometry, to create the Euclidean property space that best reproduced the pairwise similarities. This space could then be compared to and combined with the topological shape space by PCA.

The third problem was incorporating isosterism into the property space to represent, for instance, that a tetrazole ring can substitute for carboxylate, because they put negative charge at similar locations. With the Daylight toolkit[61] it was easy to write SMARTS rules to identify atomic properties such as ionization, radius, H-bonding, or aromaticity. For each building block, I could then make an "atom layer table" that summed each property, for all atoms, at each bond-count-distance from the site of template attachment. I wanted to create a similarity matrix and property space as I had for the Daylight fingerprints, but Tanimoto similarity assumes strings of binary data rather than tables of continuous values. I had seen several "generalizations" of Tanimoto distance, but none seemed to quite capture its spirit. On the basis of rather clumsy arguments, I devised "Smin/Smax" similarity: the sum of the minima of the corresponding table cells divided by the sum of the maxima. I liked the way the coefficient behaved, but was hard pressed to justify it. Soon afterward, at an American Chemical Society meeting, Gerry Maggiora spoke on fuzzy logic.[62] He described how the minimum and maximum are fuzzy logic analogues of the intersection and union. Hence, my Smin/Smax similarity was precisely the fuzzy logic extension of Tanimoto similarity, and I was now justified in using it.

Thus, every difficult piece of the library design puzzle was supplied, as if by providence, at just the right time. The problem itself was offered to me when Ron Zuckerman began to employ submonomer synthesis. Nouna Kettaneh-Wold introduced me to D-optimal design, which provided both a measure of diversity and the engine to optimize it. Blaney and Dixon's distance geometry chapter suggested MDS as a rigorous way to incorporate binary data into an Euclidean property space. Finally, Gerry Maggiora's lecture on fuzzy logic justified the use of Smin/Smax similarity to compare atom layer tables. I have never properly thanked all these contributors, and I am pleased to have the opportunity to publicly thank them now.

How Do I Feel about the Work Today? By initially focusing on maximum diversity, I solved only half of the problem. After finding many potent ligands, which were hard to convert to effective drugs, we soon noticed that maximizing diversity inadvertently biased our libraries toward heavy, hydrophobic, flexible compounds. Adding a mechanism to constrain the D-optimal diversity selection to a desirable distribution of physical properties, QSARs, and docking results completed our approach to library design.[63,64] We have since added improvements, such as sensitivity analysis[65] and 3D similarity measures.[66,67] Nevertheless, the basic formula remains the same: devising molecular similarity measures that can be computed for any potential substituent, converting the similarities to a single, coherent, Euclidean property space with MDS, and selecting a D-optimal subset to maximize diversity (now subject to additional pharmaceutical constraints) are still the bases of our library design efforts today.

Abbott Perspectives on the Calculation of Molecular Diversity

Yvonne C. Martin and Mark G. Bures

This reminiscence reflects the work and intense discussions of Yvonne Martin, Mark Bures, and Rob Brown (now at Pharmacopoeia).

Yvonne's early work involved traditional QSAR of small sets of molecules. Work published in 1973 by Hansch and colleagues[68] reported using cluster analysis to group together aromatic substituents that are similar in hydrophobic, steric, and electronic properties. A good series design would involve the synthesis of one analogue from each cluster; this would provide the maximum uncorrelated spread in physical properties, a diverse subset in today's terms. Yvonne extended the cluster analysis to substituents at an aliphatic site.[69] She and Helen Panas[70] also showed that cluster selection gave a series with better separation of physical properties (more diversity!) than traditional manual series design or than a computer program[71] that helped medicinal chemists choose diverse substituents, but that it also allowed the user to reject hard-to-make molecules. This enthusiasm for cluster analysis was further expanded to include 3D CoMFA steric descriptors of molecules aligned by a common pharmacophore.[72] Cluster analysis was preferred to competing strategies[71,73] mainly because in practical situations it is important to know for a design which molecules can substitute for one another, that is, are similar to each other.

Mark became involved in molecular diversity in the early 1990s when Jim McAlpine, who was setting up high throughput screening at Abbott, started to purchase compounds to augment the Abbott collection. We decided to describe molecules by their Daylight fingerprints[61] and use

Jarvis−Patrick clustering[74] to group them. Precedents for this were Peter Willett's analysis clustering algorithms[17] and Jeff Blaney's work clustering the whole DuPont compound collection.[75] We also realized that diversity was not the only criterion and put into place exclusion rules to avoid purchasing reactive, unstable, or otherwise unattractive compounds—those that would not make good leads. Our first strategy was to cluster the offered compounds with those at Abbott for which there was enough sample for HTS, identify the clusters that contained no Abbott compounds, and suggest for purchase one compound from each of these clusters.

Mark soon found that this strategy produces some very large clusters of very diverse structures and also many singletons. When he examined the singletons he discovered that often two or more of these were similar, but the Jarvis−Patrick algorithm did not form them into a cluster. Accordingly, Mark would postprocess the suggested list to winnow out one of a pair of similar singletons. From our experience with 3D molecular modeling, particularly substructure searching,[76] and pharmacophore detection,[77] we also expected that intermolecular property-based 3D descriptors would provide a more realistic description of the molecules considered. However, at this point we (1) did not have a metric for the quality of a clustering run; (2) did not know which type of molecular descriptor would perform best; and (3) did not know which clustering algorithm would perform best.

To investigate these problems, we were given permission to hire Rob Brown, who had just obtained his Ph.D. with Peter Willett. Rob phrased the question as "Which combination of descriptors and clustering methods best groups active compounds together and away from inactive compounds?" Following the precedent of Willett, Rob used a number of easily calculated molecular descriptors and available algorithms to cluster these data sets.[22] He also implemented a program to generate a single-conformation 3D pharmacophore fingerprint.[78]

Rob initially investigated high throughput screening data, but the proportions of active compounds and the noise in the data made it hard to see the signal in the data. Yvonne then remembered the compounds screened in the late 1950s for inhibition of monoamine oxidase. These 1650 compounds, of which 390 have some activity, were mainly hand-selected by a medicinal chemist who was searching for a novel type of inhibitor of monoamine oxidase. There are two series of synthesized molecules in the set, however: acylhydrazines based on the compounds reported earlier by Hoffmann-LaRoche; and propargylamine derivatives, a series designed from an early weak screening hit. We also selected two more recent data sets for which most of the molecules tested were designed and synthesized for the project.

Because the goal of this work was to select descriptors and clustering methods most appropriate for selecting compounds for biological testing, to measure clustering effectiveness Rob tabulated how well a method placed only active compounds in clusters with other active compounds. The higher the proportion of actives in a cluster, the more likely one is to discover that "active" cluster by testing only one member.

Before it made sense to perform the clustering, we needed to reassure ourselves that "similar" compounds have similar biological properties and that as molecular similarity decreases so does the biological similarity. Rob measured the pairwise similarity of all compounds to each of the actives and produced a histogram of similarity vs the fraction of compounds within that similarity window that are active. He found that if compounds are approximately 0.85 chemically similar, then if one is active the other has an 80% chance to also be active.

We concluded that structural diversity in a data set for biological screening does not depend on some measure of maximum diversity, only that the compounds have no neighbors that are more than 0.85 similar. This insight led to a simplification of the compound selection strategy: we now compare the vendor database with the Abbott compounds available for screening and reject those that are 0.85 similar to an Abbott compound. We then cluster the remaining vendor compounds and select one from each cluster.

Rob discovered that the major problem with our previous strategy was with the Jarvis−Patrick clustering method. Of the clustering methods tested, Ward's performed the best and is fast enough to cluster databases of reasonable size. Different descriptors did make some difference. Although we had expected our 3D pharmacophoric feature descriptor set to perform the best, descriptors based on the structure diagram of the molecule performed better than any derived from atom-based properties of the Concord 3D structure. The best performer was a fragment-based fingerprint, which marginally outperformed more sophisticated 2D fingerprints.

On the basis of this work, a measure of "diversity density" of a data set is the average number of compounds per "biohomogeneous" cluster and a measure of raw diversity is the number of biohomogeneous clusters in the data set. The monoamine oxidase has an average of 1.9 compounds per cluster; the Abbott library in 1992 (before diversity additions), 2.6; and one particular combinatorial library, 16.2.[79]

We also observed that the Abbott collection contains >90% of the possible pharmacophoric triangles with 0.5 Å tolerance based on hydrogen bonding acceptor and donor properties, positive and negatively charged groups, and aromatic ring centers.

Because we believe that physical properties are somehow correlated with biological properties, at least in smaller sets of molecules, Yvonne convinced Rob to perform a similar analysis using physical rather than biological properties.[80] He found the same order of accuracy of descriptors and clustering methods for predicting octanol−water and cyclohexane−water log P; acid−base pK_a; Kier-Hall Φ and $\kappa\alpha1$, $\kappa\alpha2$, and $\kappa\alpha3$; molar refractivity; surface area and volume; and number of hydrogen bond donors and acceptors. We found that the descriptors do contain information about each of these properties and that there is no need to augment the substructural descriptors with physicochemical ones.

To enhance diversity in combinatorial libraries we use a genetic algorithm that simultaneously considers molecular weight redundancy in one position of the molecule, CLOGP

distribution of the library, and diversity measured as the number of clusters included.[81]

Even Ward's clustering is not perfect, however. In one case our selection method had led to the purchase of a large set of analogues. This was discovered when >200 "diverse" compounds hit in a screen. Rather than being dismayed by this, the medicinal chemist was delighted to get a large volume of SAR very quickly. Probing the issue further, we discovered that hits from HTS are much more likely to be followed up with synthesis if there are a number of analogues available for testing. Our current compromise is to purchase compounds only from large clusters. We expect that if one of the compounds hits, then we will be able to purchase a number of analogues to fill out the SAR and so guide the decision as to whether to start a synthetic program around this hit.

In summary, what started out as a simple exercise in selecting compounds to purchase evolved into chest-beating claims that "my database is more diverse than yours", then returned to a more measured view of diversity that emphasizes the relevance of the diversity measure for designing libraries for biological testing.

From Synthesis Design to Combinatorial Chemistry − and Back Again

Johann Gasteiger

I want to use this invitation to contribute a section to this special perspective on molecular diversity in the *Journal of Combinatorial Chemistry* to tell a story about the interplay of scientific interests and outside funding possibilities—and of being stubborn!

I guess I have to start with my Ph.D. work that dealt with the elucidation of reaction mechanisms and left forever in me the desire to understand—and model—chemical reactions. Thus, for a long time my group has worked on the development of EROS (elaboration of reactions for organic synthesis) that was to be applied both to reaction prediction and synthesis design.[82,83] However, more and more we had to realize that we have to segregate these two types of applications.

In came a very able co-worker, Wolf-Dietrich Ihlenfeldt, who almost single-handedly developed a new system, WODCA (workbench for the organization of data for chemical applications), for synthesis design.[84] The system was a drastic departure from our previous attempts, as it did not bother with explicitly modeling chemical reactions. Rather, the system concentrated on the comparison of chemical structures.

One of the first and most powerful tools for synthesis design incorporated into the WODCA system were similarity searches.[85] These were designed to rapidly find relationships between a synthesis target and available starting materials. Quite a few criteria were developed for the definition of relationships between the structure of a target compound and an available starting material. In fact, we explored about 50 different similarity criteria and still have more than 40 contained in WODCA.

Many of these criteria are based on substructure perception, e.g., a target and a starting material are considered as similar if they contain the same ring system or if they correspond in their carbon skeleton. However, of even more importance are those criteria that are based on generalized reactions: two structures are considered similar if they can be interconverted by certain reaction types. For example, a structure is considered similar to another one if it can be converted into this molecule by oxidation or by a substitution reaction. Thus, these methods do not only perceive similarity but they also tell which reaction types can utilize these similarities in synthesis design.

The important point is that these similarity criteria can be translated into structure transformations. An entire database of available starting materials can be converted according to these similarity criteria into transformed structures once and for all. Then, only the query structure has to be transformed to submit it to a similarity search in the transformed structure database. This, combined with a hashcoding scheme,[86] allows for very rapid similarity searches.

At the same time as these developments on the perception of structural similarity useful for the design of syntheses were going on in my research group, I was project leader for building the ChemInform reaction database for the Fachinformationszentrum (FIZ) in Berlin.[87] Quite a few of the members of the scientific board of FIZ Chemie became rather frightened at the prospect of having in a few years 200 000 and more reactions in a database. Who is going to be able to scan through such a large database? Will the user be swamped with such a high number of hits of reactions that he will not be willing to analyze the results of a query? The suggestion was to keep only a few representative reactions (whatever that meant) of previous years in the database and put emphasis on new reactions. I had to quite vehemently argue for the value of information imbedded in every single reaction instance. Each variation in the structure of starting materials or products carries important information on the scope of a reaction type. The point can therefore only be to keep each individual reaction but to organize them into reaction types in order to make reaction searching more efficient, more rapidly leading the user to those reactions he/she is mostly interested in.

The question was then, how could reaction instances be grouped into reaction types? Or, to put it into other words: how can the similarity of reactions be perceived?

The chemist's concept of a reaction type clearly rests on mechanistic considerations: Reactions are conceived of belonging to the same type if they have the same reaction mechanism, if they are governed by the same physicochemical effects. Over the years we had developed a variety of empirical methods to quantify all-important chemical effects such as charge distribution,[88] inductive effect,[89] resonance effect,[90] polarizability effect,[91] and bond dissociation energies.[92]

We had put great effort into making these methods quite rapid, as we had to evaluate a sizable number of reasonably large structures in reaction prediction and synthesis design. These procedures had been collected into the package

PETRA (parameter estimation for the treatment of reactivity applications).[93]

Thus, it suddenly became quite clear what to do: use the values of these physicochemical effects calculated for the atoms and bonds at the reaction center to perceive the similarity of reactions. Reactions that have similar values for these physicochemical effects at the reaction site must belong to the same reaction type and should therefore be grouped together.

We first used machine learning techniques such as conceptual clustering to perceive the similarity of reactions exploring both these physicochemical effects and the substructures that are bound to the reaction site.[94] Later on we switched to self-organizing neural networks for similarity perception solely relying on the quantitative values for physicochemical effects at the reaction site.[95]

Thus, we had ended up using **structural similarity** for the design of syntheses, and **similarity of reactions** as a basis for making searching in reaction databases more efficient.

In the beginning of the 1990s we had to realize that although we had put nearly two decades of work into the development of synthesis design tools, organic chemists were not prepared to use them. (We were not in a singular situation: None of the other synthesis design systems had a broad user community!)

Thus, our attention shifted to the relationships between chemical structure and biological activity. The idea was that similar factors such as those responsible for chemical reactivity—charge distribution, polarizability, and steric effects—are also operating in the binding of a ligand to its receptor. Clearly, the three-dimensional structure of a ligand also plays a dominant role. Here we were in the lucky position of having already developed the 3D-structure generator CORINA that is able to automatically produce a 3D model for any organic molecule from its constitution as embodied in a connection table.[96]

Now the things that we had worked on for many years had to be put together: 3D structures, physicochemical effects, and similarity perception by neural networks.[97]

The 3D structure allowed us the calculation of molecular surfaces, and on the basis of the partial charges on the atoms as obtained from the PEOE method,[88] the electrostatic potential on the molecular surface could be calculated. As the PEOE method is quite rapid, large data sets of molecules such as those obtained from combinatorial chemistry experiments can be processed with a reasonable amount of computation time. However, we had not yet solved the problem. Methods that establish relationships between objects and their properties by learning from data such as statistical or pattern recognition methods or, as in our case, neural networks need a uniform representation of the objects and have to describe the objects with the same number of descriptors. In our case, the electrostatic potential on the molecular surface had to be transformed into a fixed set of descriptors, irrespective of the size of the molecule or the molecular surface. For this purpose the mathematical procedure of autocorrelation was used to transform the molecular electrostatic potential into a set of 12 descriptors.[98] We could show that such an autocorrelation of the molecular electrostatic potential was able to establish the similarity and dissimilarity of three different combinatorial libraries.[99]

Where have we gone since then? We have developed a hierarchy of molecular descriptors based either on the constitution, the 3D structure, or on molecular surfaces that consider a variety of chemical effects such as charge distribution, polarizability, as well as molecular electrostatic, hydrogen bonding, or hydrophobicity potential.[100]

We could show that these descriptions of the structure of molecules are quite useful for finding correlations with biological data. It is also quite clear that for a given problem different structure descriptions have to be explored in order to find one that is best suited for modeling a relationship between chemical structures and their biological properties, for the interaction of a ligand with its receptor can be governed by quite different effects depending on the receptor being studied.

In these efforts, neural networks have been of invaluable help.[97] It is our strong conviction that any study of the relationships between chemical structure and biological activity should first use an unsupervised learning method in order to establish whether the structure representation chosen is useful for reproducing the biological activity under investigation. Only when this has been established should a supervised learning method be employed to develop a quantitative model of this relationship. We are mostly using self-organizing (Kohonen) neural networks for unsupervised learning, and back-propagation or counterpropagation neural networks for supervised learning.

What next? Reactions! Over the years we have continued working on understanding reactions. Three lines of research and development were pursued: synthesis design (WODCA), reaction prediction (EROS), and acquisition of knowledge on reactions from databases (CORA).[101] Substructure searches have been incorporated into WODCA, allowing it to be applied to the design of combinatorial libraries. EROS has been developed to handle a variety of reaction techniques, from laboratory reactions to combinatorial syntheses.[83] The system CORA has been used to define the diversity of chemical reactions.[102]

In the community it becomes increasingly clear that a better understanding of the reactions employed in parallel synthesis and combinatorial chemistry is absolutely necessary. A large part of work has to be devoted to the optimization of a reaction type. If computer methods can be used for establishing the scope and limitations of a reaction type, valuable time and money could be saved. The challenge given by reactions and by synthesis design is still existing!

Tripos Perspectives on Molecular Diversity: Looking toward General Methods of Library Design

Richard D. Cramer

My own perspective on diversity in combinatorial chemistry owes much to the converging contributions of my colleagues at Tripos, particularly those of David Patterson and Bob Clark. For example, it was Dave who became most

convinced at the 1993 Gordon Conference on QSAR that molecular diversity and combinatorial chemistry were becoming a new and critical concern for Tripos's software customers. He and I soon responded by turning our attention from improvement of Leapfrog and DISCO to development of two of the earliest and still popular "diversity software" products, Legion and Selector, for the construction and design, respectively, of combinatorial libraries. Of course we then had only the crudest of ideas about how combinatorial libraries should be designed, and I bought several dinners for Eric Martin while trying to understand how the pioneering Chiron group[55] thought about the new "diversity" buzzword. The resulting first release of Selector (end 1994) helped the user to select one or more "representative" building blocks from each cluster among a set of reagents. These clusters were automatically generated hierarchically once the user provided the candidate reagents, the number of these to select, and the physical properties and weights to use in the clustering (hence in the selecting).

But "which of the physical properties provided by Selector should we use for clustering/selecting?" was a logical next question, asked of me with particularly memorable emphasis by computational chemists from Schering AG (Berlin). Although I readily agreed to the importance of this question, I had no idea then how to answer it. Motivation to start thinking about it came quickly, though, from my boss, Tripos's president John McAlister. In early 1995 John found a way for Tripos to start becoming "more than a software company", by allying Tripos with PanLabs to produce a "general screening library", trade-named Optiverse.[103] Tripos's contribution was to be design, marketing, and sales, and Dave and I were the ones charged with deriving principles and methods for screening library design.

Since by definition there can be no advance information about the screening targets for such a library, we could imagine only three design principles, the first two almost self-evident:

- Avoid structures whose physical properties are less compatible with good oral bioavailability (e.g., heed the Lipinski[104] rules).
- Avoid substructures that are likely to confer unwanted toxicities or other nonspecific biological responses.
- Assuming a "neighborhood behavior" (meaning that "similar" structures usually have similar biological properties), the design should space the structures to be synthesized far enough apart to avoid the presumed waste of repeatedly "looking in the same place" for bioactivity.

Ever since, my own scientific activities have mostly addressed just three words of the third principle, "far enough apart". How far is "enough"? In what "property space" is this distance best measured? Can a "better" property space be devised (where "better" means a higher frequency/intensity of neighborhood behavior)? Is there any objective way to compare property spaces? And most importantly, what other applications exist for a "better" property space? (Indeed, and ironically, today the third design principle itself seems questionable. To rapidly distinguish the "true hits" from the equally numerous "false positives" generated by even the most accurate HTS, it is instead an increasingly common practice to rely on "neighborhood behavior" by requiring that any apparently active structure be confirmed by observing similar activity from at least one "similar" structure. Or, put differently, optimal HTS design practice today is to test numerous small sets of "not quite duplicate" structures. Of course, the "other applications" that subsequently emerged justify a continuing interest in "better property spaces".)

Encouraging evidence that objective comparison of the neighborhood behaviors of different property spaces was possible came at the spring 1995 American Chemical Society meeting, where Rob Brown and Yvonne Martin described how the Tanimoto coefficients of nonaliassed "2D fingerprints" (MACSS keys) segregated actives from nonactives significantly better than did other similar descriptors.[22,105] Meanwhile, in response to a general feeling that a "proprietary 3D descriptor" would be an asset to the Optiverse "design story", I had begun experimenting with "topomers". The starting point was a conviction Bob Clark had previously brought to Tripos, that is, that the investigation of structural alignments for CoMFA should begin with "inertial alignment" of complete structures (as implemented within the above-mentioned Selector program). In combinatorial library design, where the basic task is to select a set of side chains, the arguments in favor of such a "standardized conformer" approach to 3D alignment seemed attractive to me. In particular, superposition of attachment bonds, which all of a set of side chains share by definition, was a compelling new approach to the long-standing difficulty in shape comparison of mutually orienting 3D molecular shapes.

The first topomer experiments, hierarchically clustering the roughly 700 thiol-containing reagents offered in ACD according to the differences in a standard CoMFA steric field enclosing their topomer conformations, yielded results much more convincing than I had expected.[106] (Thereby initiating an extraordinary and most encouraging trend—it seems to me that topomers have always performed better than I expect them to.) But there was an evident challenge in communicating these results. Few colleagues would have the time or interest to evaluate hundreds of sets of clustered 3D structures on a computer screen, as I had done. Was there some way, without any access to large screening data sets such as Rob and Yvonne had used, and considering only side chain descriptors since combinatorial design was our purpose, to objectively compare topomer steric shapes with other descriptors? Plainly the literature does provide many examples of biological measurements for a set of structures differing only in their side chains, even if those sets are not very large. I began assembling such sets and regressing differences in various candidate side chain properties against the corresponding differences in bioactivities. But the resulting r^2 values were statistically marginal, and the residual plots were funny looking.

Dave Patterson was sure he could do better than these validation efforts, and after a few days of tinkering and only a few more days of argument he convinced me he had done so. His "Patterson plots" test the existence of any neighborhood behavior for a given descriptor and data set very directly, by providing empirical rules for enumerating any

counter-examples (appearing as data points within a forbidden upper left triangle, whenever any small structural change produces a large biological change).[107] Not only did Patterson plots seem logically incontestable, they then ranked descriptors in accord with my intuitive expectations. (Twenty years earlier I had discovered a very high "BC(DEF)" intercorrelation among properties such as aqueous solubility, octanol/water logP, molar refractivity, and boiling point.[108] This finding left me convinced that such scalar properties, no matter how usefully related to "3D-orientation-averaged" biological transport, would have little value in describing "3D-orientation-specific" phenomena such as receptor binding.)

By the time all these concepts were all worked out, the supporting experiments programmed and executed, patents filed, and the first pair of publications drawn up, say early in 1996, we were all thinking about a much more important use for "neighborhood searching" than screening library design. The virtual libraries which were readily accessible in principle, by applying two or three simple synthetic steps to commercially offered starting materials, reference many orders of magnitude more "drug-like" compounds than the total number of compounds ever characterized in the literature. Could we think of ways to rapidly and effectively identify the biologically most promising structures in such vast virtual libraries, by performing $m + n + o$ calculations on its $m + n + o$ individual "pieces" or "synthons" rather than $m \times n \times o$ calculations on its fully enumerated "product" structures?

As an illustration of the methods we sought, consider how rapid a virtual library search can be for all product structures whose molecular weight is less than some upper bound. A huge advantage compared with conventional structural searching is that molecular weight comparison involves only a single numerical operation. Furthermore, any synthons whose individual molecular weights are greater than the upper bound cannot yield acceptable products and are immediately excluded. If the acceptably small *m*'s, *n*'s, and *o*'s remaining are then sorted and processed by descending molecular weight, it will similarly not be necessary to consider most of the remaining products, either. The more restrictive the molecular weight criterion, the more restrictive are these criteria, and the faster the search.

Of course, similarity in molecular weight does not predict similarity in biological properties to any useful degree (as confirmed by our descriptor validation studies). But, fortunately, we did succeed in devising related methodological short cuts with the two molecular descriptors for which we had found the most consistent neighborhood behavior, Tanimoto 2D fingerprints and topomers. It was also self-evident by 1996 that the most accessible and versatile packaging for such a new computational methodology would be as a "Web application", and thus ChemSpace was born.[109]

Since then I have been fully occupied by extending ChemSpace in various directions, Fred Soltanshahi particularly helping me with its programming, Kathe Andrews-Cramer with its testing and use, and many colleagues at Tripos Receptor Research (Bude, Cornwall, U.K.) and BMS with its applications. Consistent success in these applications so far, both published[110,111] and unpublished, has increased my belief that topomers represent an unprecedentedly useful, as well as novel and sometimes controversial, paradigm for comparing molecules. Perhaps this story is only at its beginning.

A Reminiscence about Chemical Diversity

Robert S. Pearlman

They say that necessity is the mother of invention. That was certainly true of Concord! Early in 1986, despite a clear indication that my logP model was based on the 3D structures of a small training set of pollutants, the EPA sent me 2000 Smiles strings as a validation set for renewal of my largest grant. The original version of Concord was hurriedly written so that I could convert the 2000 Smiles to 3D and meet the renewal deadline.

Very shortly after recovering from the panic of the grant renewal, we realized that Concord could be used to convert large databases to 3D and, thereby, to enable 3D searching as a new approach to lead discovery. But not long after MDL and Tripos introduced MACCS-3D and Unity, it occurred to me that 3D searching within huge databases of "virtual" compounds might be far more exciting and fruitful than 3D searching within the typical corporate database of ~250 000 compounds (in the late 1980s). Thus, before I learned that pioneers at Glaxo and elsewhere were starting to develop methods for real combinatorial chemistry, our group began working on a program for making databases of virtual combinatorial compounds for 3D searching. *Naturally*, we named the program CombinDBMaker. Once "library" emerged as the standard term for a collection of combinatorially generated products, we changed the name to CombiLibMaker[112]—more descriptive and *so* much easier to pronounce...

The 3D searching software of the early 1990s remained too slow to search combinatorial libraries as large as we could make. Also, I questioned whether it was really worth searching all of the "redundant" products produced from the trivial "R-groups" I was using in our early combinatorial libraries. For these reasons, I began thinking about methods for diverse subset selection. Shortly after that, I learned that some industrial scientists were thinking about selecting diverse, "representative" subsets for wet-lab screening while others were thinking about selecting diverse subsets for actual synthesis. The notion of diverse subset selection motivated not only our group's interest but everyone else's interest in "chemical diversity", and software developers around the world immediately seized upon the challenges and opportunities which that interest presented.

Unfortunately, opportunity frequently spawns competition, and in my opinion, the flurry of articles presenting and reviewing competitive claims for one diverse subset selection algorithm over another or, even more ethereal, one measure of "diversity" over another, comprise a rather sad chapter in the annals of computer-assisted drug discovery. In retrospect, these articles might be regarded as "much ado about nothing." Noncomputational industrial scientists delighted in reporting that the results of screening diverse subsets were

often no better (and sometimes slightly worse) than the results of screening randomly chosen subsets.

Somewhat ironically, diverse subset selection—the task which motivated our original interest in chemical diversity—is now (or should be) generally regarded as the least useful of all diversity-related tasks. In retrospect, the notion of screening a diverse, "representative" subset of a screening library seems somewhat naïve. Chemistry-space metrics (descriptors or fingerprints) must be applicable not only in the context of today's discovery target but in the context of tomorrow's as well. Indeed, chemistry-space metrics must also be applicable in contexts wherein there is no reference to a particular receptor. Imagine three compounds—A, B, and C—which differ from each other by just a single methylene group. Chemists regard such compounds as highly similar and expect that "meaningful metrics" will place the three compounds very near each other in chemistry-space. Indeed, we reject metrics which do not meet this expectation. Diverse subset selection might select either A or B or C or some other compound to "represent" that region of chemistry-space, and chemists are initially pleased with the notion that resources are not wasted on screening the "redundant", similar compounds in the same region. However, the literature abounds with examples demonstrating that apparently trivial structural differences between homologues can result in either trivial or drastic differences in the bioactivity of the compounds. If compounds A and B are active and C is not—a *very* real possibility despite their similarity—the diverse subset selection algorithm is no more likely to select A or B than C or some other compound to "represent" that region of chemistry-space. Random subset selection could possibly do just as well or even better since, unlike diverse subset selection, random selection might choose both A *and* B.

We realized the potential folly of diverse subset selection very early in the "diversity game", but we also realized that there were other, more useful "diversity-related tasks" which could be addressed with valid chemistry-space metrics and novel algorithms. Our diversity software was originally named DiverseSelector, but we quickly changed the name to DiverseSolutions[113] to reflect that realization. The same BCUT metrics[2] we originally developed to enable cell-based diverse subset selection in a low-dimensional chemistry-space have proven extremely useful for both diverse and focused combinatorial library design, hit/lead follow-up based on iterative near-neighbor searching, rational compound acquisition strategies, meaningful methods for comparing two or more populations of compounds, computer-graphic visualization of actual coordinates of compounds (actives or entire libraries) in chemistry-space, etc.[114−116]

BCUT metrics can be computed based on either a 3D or 2D representation of chemical structure. While 3D-BCUTs appear to offer some advantage for applications within a particular combinatorial library, it appears that 2D-BCUTs work just as well when dealing with a more diverse population of compounds. Thus, recalling that our lab's contributions to the field of "chemical diversity" were originally motivated by our work on Concord, it seems entirely fitting to close this reminiscence by noting an amusing similarity. Just as drug discovery often benefits by serendipity, so too does the "discovery" of new methods for computer-assisted drug discovery. At least that is how things go in our lab.

Molecular Diversity: Molecular Properties to 3D Pharmacophore Fingerprints New

Jonathan S. Mason

Our interests in "diversity" began at the Rhone-Poulenc, Dagenham, U.K., CADD group from a need to be able partition the corporate database for screening, and they continued with force with the arrival of combinatorial chemistry design as one of our major tasks. Partitioning, or cell-based methods, appealed to the group, as it is possible to see both what is there and what is missing, in absolute terms and between databases without any need for reanalysis. Three-dimensional database searching for lead generation became an early success for the CADD group, providing a clear deliverable, i.e., one that could not have occurred anyway within medicinal chemistry without the use of CADD methods. Multiple 3D pharmacophoric shapes or "pharmacophore fingerprints"[117−125] from systematic analyses have long been the focus of research interest of ours for the early pharmacophore-derived query (PDQ) method[117,118] based on automated 3D database searching, Stephen Pickett and Iain McLay, and later Dan Cheney for site-derived pharmacophores. The fingerprints provided us with a powerful approach in which the "descriptor" can be used in a partitioning, cell-based way, a compound generally occupying many cells instead of just one. Initial efforts were directed to profiling ligands, but this was later extended to protein sites, a natural extension to site-based 3D database searching: complementary site-points with associated pharmacophoric features are first generated, and the fingerprint is either generated directly from these[121] or by using the protein site as a steric constraint during the fitting of matching conformations.[124,126] A significant increase in the amount of shape information and resolution was found using four-point pharmacophores, including the ability to distinguish chirality, a fundamental requirement for many ligand−receptor interactions.[122,127]

Despite our passion for 3D database searching and pharmacophores, when our first diversity-related project came along 10 years ago we had to make a more pragmatic choice. The challenge given to us was to come up with a "rational" screening set of 1000 compounds culled from the entire corporate database (>150 000) that "represented" the drug-like diversity of the entire set. This would then be used for the many lower throughput biological screens. Although we were concerned whether this could be effectively achieved, the priority of the need at the time meant that a major part of the resources could be put into a single project, a rarer opportunity now. Iain McLay (now at Glaxo-Wellcome) and myself had been looking into molecular/physicochemical properties and pharmacophores, and when Richard Lewis (now at Eli Lilly) joined the group, he was able to focus a dedicated effort, leading to the DPD ("diverse property derived")[128] partitioning method. We decided for pragmatic

reasons to look for six reasonably noncorrelated molecular/physicochemical properties/descriptors that could be relevant to ligand−receptor interactions and use these to select the set. This turned out to be harder than we thought, as so many potentially relevant indices had too much pairwise correlation with descriptors already chosen. We ended up creating two "new" descriptor combinations, something that was not at all in our initial plans! An intuitive/subjective medicinal chemistry approach was used to select and evaluate descriptors, which then had to pass the pairwise correlation filter. The polarity descriptor (the normalized sum of the squares of the atomic electrotopological indices) appeared to be able to score in a single number the polar nature of compounds better than descriptors from just partial charges. The "aromatic density" descriptor (the number of aromatic rings divided by the molar volume) was the result of a lengthy search to find a sixth descriptor with acceptable pairwise correlation that used "relevant" parameters. These were used together with counts of hydrogen bond acceptors and donors, a flexibility index, and calculated octanol−water logP. We considered logP a key physicochemical property, but its use in fact caused much divided discussion among the medicinal chemists as to its relevance for ligand−receptor interactions, the goal of the screening set. We had chosen it knowing that cell-based assays would also be used and it could be directly relevant. Together with the hydrogen bond counts and molecular weight (MW), it is now part of Lipinski's[104] "rule of 5" absorption guidelines.

The DPD method was designed to select "diverse" rather than "representative" compounds. Well-represented chemotypes possibly only had a few compounds in the set. Unfortunately this was too diverse for many, exploring too much the extremes of the chemistry space expressed by all the compounds. By design, only a very few hits would be expected for a particular screen, which could be followed up with "similar" compounds. The DPD descriptors together with MW were, however, found to be useful for profiling combinatorial libraries, to moderate large deviations from "drug-like" databases. This extended set includes the Lipinski "rule of 5" descriptors.

When the follow-up request for a larger set of 10 000 compounds came, issues such as the diversity of the compounds and the low hit rate led us to abandon such a "diversity" approach for a more "representative" set. Richard got a RPR global project going, involving Isabelle Morize and Claude Luttman in France and Paul Menard and myself in the United States, using Jarvis−Patrick clustering of 2D structural fingerprints (Daylight). "Cascaded" clustering was devised to handle the large number of singletons.[129] Although we had been working on automating a systematic 3D pharmacophoric search with the new PDQ method,[117,118] we opted against using this approach because of the large calculation times involved and the lack of a validated method to use such descriptors effectively to pick a representative subset.

Three-dimensional pharmacophoric approaches had been a focus of research interest of Iain and myself, and extensive customized atom-typing parametrization and conformational sampling were being developed. We had been inspired by the pioneering work of John Van Drie and Yvonne Martin with the Aladdin 3D database searching system,[130] and we are grateful to Yvonne for supplying code to enable us to try the software. The Chem-X software[119] became our standard small molecule molecular modeling software, and we focused our development efforts around the Chem-DBS-3D module that provided 3D database searching capabilities, with conformational sampling performed "on-the-fly" at search time for each search. We were trying to develop a "diversity" method involving pharmacophores by determining "all" the potential pharmacophores a molecule could exhibit through 3D searching using multiple queries ("all" being pragmatically limited to three points, six distance ranges, and six feature possibilities: hydrogen bond acceptor, hydrogen bond donor, acid, base, aromatic ring, and hydrophobe). Again the arrival of the next member of the group enabled a focused internal effort to a finalized new method, with Stephen Pickett (now at Roche) then leading efforts on the pharmacophore-derived query (PDQ) method.[117,118] This was based on automated systematic 3D database searching, using the Chem-X/ChemDBS-3D software, with extensive scripts. A "fingerprint" of potential pharmacophores matched for each molecule was thus calculated, with conformational flexibility being taken into account. The conformational analysis stage was rate limiting, and it forced us to limit ourselves to about 6000 3D searches per molecule to provide a bearable throughput on the CPUs of that time. Considerable effort was put into the atom typing parametrization to ensure that pharmacophoric features were correctly assigned (to avoid both false negatives and false positives) and to provide customized conformational sampling. As the queries were actual 3D database queries, considerable flexibility was possible in the actual definitions, with additional constraints possible. The final output per molecule was a bit string indicating which pharmacophore queries could be matched (i.e., have the molecule as 3D database search hit). A key issue was hydrophobic features, which are key to so many ligand−receptor interactions, and we tried many different methods to assign them. A systematic addition of "dummy atom" hydrophobe centroids to any potential groups of atoms was followed by an evaluation of a simplified local "electrostatic potential" to eliminate unsuitable dummy atoms. This gave us reasonably intuitive results, achieving the objective of being able to automatically place hydrophobic centers in positions, which usually would be selected by an experienced medicinal chemist. It suffered from the disadvantage of a relatively time-consuming preprocessing of the molecules to assign the hydrophobe features, and it was later dropped in favor of the very rapid method added to Chem-X that only analyzes bond polarities and gave reasonable results after some fine-tuning of the atom electronegativities. Hydrophobic feature assignment has remained a problem generally with 3D database systems, as assignment at search time based on substructure/Markush queries can be difficult due to the large number of possibilities, and/or generate too many centers, such as when just atom types are used.

We were excited by the potential power of the pharmacophore fingerprinting approach, using parameters for similarity and diversity that we all felt were more relevant for

ligand−receptor interactions. We decided that the investment of computational time needed to systematically calculate the 3D pharmacophoric shapes with conformational flexibility was a worthwhile goal, and as machine power increased we were able to use this diversity/similarity method for practical applications. The proposal of Chemical Design to add such a method to Chem-X,[119] as the ChemDiverse module, enabled us to publish our work, which was otherwise considered a proprietary idea. Nowadays, companies, including large pharmaceutical ones, may opt to patent such methods, and despite 5 years of prior art, such things are happening! We pursued development of the method with Chemical Design from thereon, as the PDQ method was slow from having to repeat the conformational sampling for each pharmacophore query, whereas with ChemDiverse only a single conformational sampling per molecule was needed. This enabled around a million three-point pharmacophores to be resolved, something not possible with individual searches! Collaborative funding led first to the development of four-point pharmacophore fingerprints[121,124] and then to "design in receptor" (DiR), [117,125] a type of protein site defined diversity method discussed later.

For the development of the four-point pharmacophore fingerprints, we thus decided that machine memory and power were going to both increase and become cheaper, and persuading Chemical Design to collaborate on developing a fairly brute-force extension of the three-point ChemDiverse method to four-point pharmacophores was considered very important. Driving forces for the four-point pharmacophores included the need to distinguish chirality (that a planar three-pont pharmacophore never can) and to include more shape information (ligand−receptor docking studies showing shape to be a major contributor). Three-point pharmacophore fingerprints do, of course, contain a lot of useful information, particularly when a count of the occurrence is also used with sets of molecules. Encouraged by early results, which had rather slow calculations, speed improvements were soon incorporated, to make the method usable for practical applications. On Silicon Graphics R1200 machines, only a second or two can be used per molecule, including conformational sampling with a bump check to exclude inaccessible conformations. Chemical Design had envisaged the outputting of the pharmacophore fingerprints ("keys") more for databases than individual molecules, but as we wished to use individual molecule fingerprints, more compact formats than the Chem-X binary key were developed which only store pharmacophores that are matched (a: Pickett, S. D. Unpublished results. b: Cho, S. J.; Mason J. S. Unpublished results).

These pharmacophore fingerprints, an extension to the original PDQ method, are now used at BMS routinely for many virtual screening and combinatorial library design applications, and they have been even more useful than I could have hoped for, given their limitations, such as the mixing of information for all the conformations and the limited shape description. The fingerprints (∼10 and 2.3 million four-point possibilities with 10 and 7 distance ranges, respectively, per feature−feature distance) give a common frame of reference for comparing different ligands and for comparing ligands to protein structures using the complementary potential pharmacophores. Other software packages are now able to calculate three-/four- point pharmacophore fingerprints. The development of methods for "relative" diversity was another useful extension to the method, focusing the diversity/similarity measure around pharmacophoric shapes that contain a substructure or feature of interest. We used this extensively at the RPR United States site for the design of combinatorial libraries using G-protein coupled receptor "privileged" substructures.[122,123] More than 100 000 compounds were synthesized using condensation reactions such as the Ugi reaction. The quantification provided by the fingerprints was useful to indicate when no new relevant "diversity" could obtained from a particular reaction, as was the ability to interpret "missing diversity" in terms of pharmacophoric features and shapes, superimposed if required on the reference molecules giving rise to them. We have also found them very useful for 3D similarity searching for lead identification, enabling multiple potential pharmacophore hypotheses to be readily handled. This is particularly useful where small peptides are the input, an input that many other methods cannot readily handle if the desired result is not another peptide. One or more flexible molecules have been successfully used as input for new lead generation, without the need to derive a refined model. Using precalculated fingerprints, stored efficiently, it is possible to do 3D pharmacophoric searching for 500 000 compounds in minutes.

We found for combinatorial library design that it could be more effective to combine an optimization of pharmacophore diversity (or similarity) with a simultaneous optimization of other properties. Simple profiling properties (e.g., DPD flexibility, "shape" estimates, pharmacophoric promiscuity) were initially used, and through excellent collaboration with Professor Bob Pearlman we became involved with other atomic/molecular properties, the BCUT descriptors (DiverseSolutions, DVS).[114,131] We found that chemistry spaces derived with these were very useful for many diversity-related applications including library design. Complementary information to the pharmacophore fingerprints is obtained, and both methods were used to create a representative subset of combinatorial libraries synthesized at RPR.[122,123] The development of the use of both partitioning methods, the pharmacophore fingerprints and the BCUT cell-based chemistry spaces, has continued at BMS, and a paper illustrating the simultaneous optimization of both pharmacophoric and BCUT diversity using a simulated annealing method developed by Brett Beno at our Wallingford site is in press.[132]

Fingerprint calculation time is still rate-limiting for large (>1 million) virtual libraries, and initial analysis by much faster diversity methods, such as the DVS BCUT metrics, or at the reagent level is generally needed. Our bias has been to use whole molecule descriptors for diversity-related applications such as combinatorial library design that are optimally measured on the products of the reaction, rather than the reagents. This is more computationally intensive but to us had many advantages, including the ability to compare different libraries, databases, etc. Reagent profiling has been useful to select diverse subsets (reagents with

similar pharmacophore fingerprints are likely to produce products that have similar fingerprints), and to make these fingerprints more representative of the products, we have used a minimally substituted scaffold. Alternative strategies we have used to avoid having to calculate fingerprints for too large number of molecules include using small representative sets of reagents for all but one position and then iteratively optimizing each reagent possibility.[122] We have often used the fingerprints to fill "diversity voids", a useful application of partition (cell)-based methods,[120,133] in which it is necessary to use the whole molecule fingerprints.

Dan and myself had been pharmacophorically profiling protein binding sites by generating complementary site-points (e.g., using the GRID program) and connecting these to generate a hypothetical molecule. Fingerprints were then generated as for a normal molecule (but with no conformational analysis) and used for screening and selectivity studies.[121] We collaborated again with Chemical Design to extend the pharmacophore fingerprint method from ligands to protein binding sites to produce the "design in receptor" (DiR)[117,125] module. A single conformational analysis of the ligand evaluates the fit to all the pharmacophore hypotheses in the site, using the shape of the site as a constraint. This project was jointly funded by RPR and Bristol-Myers Squibb, my new group. It provided for us a new type of protein site defined diversity, the "diversity space" being the total possible potential complementary site pharmacophores (three- or four-point), and outputting for each molecule how many of these can be matched (together with an optional database of docked conformations). The method lacks any traditional scoring method, but unlike an energy-based score (where all the top scoring molecules could be binding in the same way) it enables a choice of molecules (from a database or for a combinatorial library design) that explores the maximum number of different binding orientations (removing possible force-field bias, and leaving it to the biological screening to find the more active compounds). The method could be improved by using external software such as DOCK to obtain a score to filter out molecules with low predicted binding.

For the future, with the current decision by Oxford Molecular to no longer develop Chem-X, and then the acquisition of Oxford Molecular by Molecular Simulations Inc., the further development of the 3D pharmacophore fingerprint methods therein is less certain, but there is clearly interest and activity from many parties in developing software methods to calculate the fingerprints. The best balance in the future will need to be determined for a particular application of the choice of pharmacophore type (two-, three-, four-points or more, the distance ranges, the features), the use of a pharmacophore count (per molecule, per conformation), the depth of conformational sampling, and the simultaneous optimization of other properties (2D, molecular and physicochemical, 3D shape...), and more validation studies need to be done to determine the "best" method for different applications. With the ever-increasing number of protein target structures, methods for structure-based or "biological" diversity will become more important. It will be exciting to see how our ligand-based hypotheses map onto the target structures.

Concluding Remarks

This set of perspectives reveals that the definition, calculation, and uses of molecular diversity remain as diverse as the people who discuss and use them. It highlights differing opinions on several areas of interest. It is up to the user to decide if these differences are significant or not. The true test of these ideas is whether they improve the success rate of the experimental collaborators.

One consensus is apparent from the articles. Several of the contributors point out the folly of assuming that only a few thousand molecules can represent a diverse collection of molecules and that biological screening has moved from the idea of an informative subset to the implementation of screening every compound available.

However, diversity measures continue to be used as part of the design of combinatorial libraries and to select vendor compounds to augment a compound collection. Although not emphasized in these reminiscences, most companies use filters to eliminate compounds that, even if active, would not be good leads for further chemistry. Usually this includes removing reactive or unstable compounds and considering if the molecules are drug-like[134,135] and if they possess features consistent with permeability.[104] For combinatorial libraries, the combinatorial constraint may preclude strict criteria for removal, but these and other considerations are part of the design process.

Diversity and its complement, similarity, are also considered when designing a targeted library: One would like the designed compounds to be similar to the known actives but different enough that the structure−activity relationships are well explored. Eric Martin, Cramer, and Pearlman provide different strategies for this use of diversity measures.

Even though molecular diversity is considered to be an important criterion for library design, the very nature of an optimally diverse data set is open to discussion. Is maximum diversity the best solution, or is "just diverse enough" a better approach? Should diversity be based on clustering, maximal diverse subset selection, or cell-based approaches? How does diversity in a library jibe with the universal chemistry experience that often enough very close analogues have very different biological or other properties?

This review emphasizes that computational chemists are faced with the challenging problem of how to convert molecular structures into the numbers recognizable by computer programs but relevant to the design goal: again this is an area of diversity of opinion. Although no one would argue that molecules are not three-dimensional, how to represent this three-dimensionality remains a challenge. Should it be by pharmacophores, by 3D properties embedded into a lower-dimensional space, or implicitly by the types of 2D substructures contained in the molecules? If a true three-dimensional representation is chosen, is one conformation per molecule sufficient or are "all" low-energy conformations needed? How does one ensure that the selected

molecules are also diverse in traditional properties such as lipophilicity, size, and charge distribution? Clearly, for uses of molecular diversity other than searching for new drugs, different types of molecular descriptors will be needed.

There are certain things that all contributors agree upon: Most importantly, the limitations in molecular diversity are no longer computer power but rather human intellectual insight. While challenges remain for data sets of greater than a million molecules, even larger data sets (26 million) have been studied[111] and sensible short-cuts are available.

Contributors' Addresses

Mark G. Bures
Eli Lilly
Lilly Corporate Center
Drop Code 1533
Indianapolis, IN 46285
mark.bures@lilly.com
Richard D. Cramer
cramer@tripos.com

Prof. Dr. Johann Gasteiger
Computer-Chemie-Centrum, Institut fuer Organische Chemie
Universitaet Erlangen-Nuernberg
Naegelsbachstr. 25
D-91052 Erlangen, Germany
gasteiger@ccc.chemie.uni-erlangen.de

Mark Johnson
Pharmacia Corp.
301 Henrietta St.
Kalamazoo, MI 49007-4940

Michael Lajiness
Pharmacia Corp.
301 Henrietta St.
Kalamazoo, MI 49007-4940

Gerald Maggiora
Pharmacia Corp.
301 Henrietta St.
Kalamazoo, MI 49007-4940

Eric Martin
Chiron Corporation, LSC 4.250
4560 Horton St.
Emeryville, CA 94608
martine@chiron.com

Yvonne C. Martin
Abbott Laboratories
100 Abbott Park Rd.
Abbott Park, IL 60064
yvonne.c.martin@abbott.com

Jonathan S. Mason
Bristol-Myers Squibb
PO Box 4000
Princeton, NJ 08543
masonj@bms.com

Professor Robert S. Pearlman
College of Pharmacy
University of Texas at Austin
Austin, TX 78712-1074
pearlman@vax.phr.utexas.edu

Professor Peter Willett
Department of Information Studies
University of Sheffield
Sheffield S10 2TN, U.K.
p.willett@sheffield.ac.uk
http://www.shef.ac.uk/~is/people/willett.html

References and Notes

(1) Martin, Y. C. Challenges and Prospects for Computational Aids to Molecular Diversity. *Perspect. Drug Discovery Des.* **1997**, *7/8*, 159−172.
(2) Lynch, M. F.; Willett, P. The Automatic Detection of Chemical Reaction Sites. *J. Chem. Inf. Comput. Sci.* **1978**, *18*, 154−159.
(3) Willett, P. The Evaluation of an Automatically Indexed, Machine-Readable Chemical Reactions File. *J. Chem. Inf. Comput. Sci.* **1980**, *20*, 93−96.
(4) Willett, P. Recent Trends in Hierarchic Document Clustering: A Critical Review. *Inf. Process. Manage.* **1988**, *24*, 577−597.
(5) Griffiths, A.; Robinson, L. A.; Willett, P. Hierarchic Agglomerative Clustering Methods for Automatic Document Classification. *J. Doc.* **1984**, *40*, 175−205.
(6) Griffiths, A.; Luckhurst, H. D.; Willett, P. Using Inter-Document Similarity Information in Document Retrieval Systems. *J. Am. Soc. Inf. Sci.* **1986**, *37*, 3−11.
(7) El-Hamdouchi, A.; Willett, P. Comparison of Hierarchic Agglomerative Clustering Methods for Document Retrieval. *Comput. J.* **1989**, *32*, 220−227.
(8) Willett, P. Textual and Chemical Information Retrieval: Different Applications but Similar Algorithms. *Inf. Res.* **1999**, *5*.
(9) Lynch, M. F.; Willett, P. Information Retrieval Research in the Department of Information Studies, University of Sheffield: 1965−1985. *J. Inf. Sci.* **1987**, *13*, 221−234.
(10) Adamson, G. W.; Bush, J. A. A Method for the Automatic Classification of Chemical Structures. *Inf. Storage Retr.* **1973**, *9*, 561−568.
(11) Sneath, P. H. A.; Sokal, R. R. *Numerical Taxonomy*; W. H. Freeman: San Francisco, 1973.
(12) Johnson, M. A.; Maggiora, G. M. *Concepts and Applications of Molecular Similarity*; Wiley: New York, 1990.
(13) Willett, P.; Winterman, V. A Comparison of Some Measures of Intermolecular Structural Similarity. *Quant. Struct.-Act. Relat.* **1986**, *5*, 18−25.
(14) Willett, P.; Winterman, V.; Bawden, D. Implementation of Nearest Neighbour Searching in an Online Chemical Structure Search System. *J. Chem. Inf. Comput. Sci.* **1986**, *26*, 36−41.
(15) Carhart, R. E.; Smith, D. H.; Venkataraghavan, R. Atom Pairs as Molecular Features in Structure−Activity Studies: Definition and Applications. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 64−73.
(16) Willett, P.; Winterman, V.; Bawden, D. Implementation of Nonhierarchic Cluster Analysis Methods in Chemical Information Systems: Selection of Compounds for Biological Testing and Clustering of Substructure Search Output. *J. Chem. Inf. Comput. Sci.* **1986**, *26*, 109−118.

(17) Willett, P. *Similarity and Clustering Techniques in Chemical Information Systems*; Research Studies Press: Letchworth, 1987.
(18) Flower, D. R. On the Properties of Bit String-Based Measures of Chemical Similarity. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 379−386.
(19) Dixon, S. L.; Koehler, R. T. The Hidden Component of Size in Two-Dimensional Fragment Descriptors: Side-Effects On Sampling in Bioactive Libraries. *J. Med. Chem.* **1999**, *42*, 2887−2900.
(20) Godden, J. W.; Xue, L.; Bajorath, J. Combinatorial Preferences Affect Molecular Similarity/Diversity Calculations Using Binary Fingerprints and Tanimoto Coefficients. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 163−166.
(21) Downs, G. M.; Willett, P.; Fisanick, W. Similarity Searching and Clustering of Chemical Structure Databases Using Molecular Property Data. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1094−1102.
(22) 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.* **1996**, *36*, 572−584.
(23) Murtagh, F. *Multidimensional Clustering Algorithms*; Physica Verlag: Vienna, 1985.
(24) Dean, P. M.; Lewis, R. A. *Molecular Diversity in Drug Design*; Kluwer: Amsterdam, 1999.
(25) Bawden, D. Molecular Dissimilarity in Chemical Information Systems. In *Chemical Structures 2*; Warr, W. A., Ed.; Springer-Verlag: Heidelberg, 1993; pp 383−388.
(26) Lajiness, M. Molecular Similarity-Based Methods for Selecting Compounds for Screening. In *Computational Chemical Graph Theory*; Rouvray, D. H., Ed.; New York: Nova Science Publishers: New York, 1990; pp 299−316.
(27) Voorhees, E. M. Implementing Agglomerative Hierarchic Clustering Algorithms for Use in Document Retrieval. *Inf. Process. Manage.* **1986**, *22*, 465−476.
(28) Holliday, J. D.; Ranade, S. S.; Willett, P. A Fast Algorithm for Selecting Sets of Dissimilar Molecules From Large Chemical Databases. *Quant. Struct.-Act. Relat.* **1995**, *14*, 501−506.
(29) Shemetulskis, N. E.; Dunbar, J. B.; Dunbar, B. W.; Moreland, D. W.; Humblet, C. Enhancing the Diversity of A Corporate Database Using Chemical Database Clustering and Analysis. *J. Comput.-Aided Mol. Des.* **1995**, *9*, 407−416.
(30) Snarey, M.; Terret, N. K.; Willett, P.; Wilton, D. J. Comparison of Algorithms for Dissimilarity-Based Compound Selection. *J. Mol. Graphics Model.* **1998**, *15*, 372−385.
(31) Agrafiotis, D.; Lobanov, V. S. an Efficient Implementation of Distance-Based Diversity Measures Based On K-D Trees. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 51−58.
(32) Mount, J.; Ruppert, J.; Welch, W.; Jain, A. N. Icepick: A Flexible Surface-Based System for Molecular Diversity. *J. Med. Chem.* **1999**, *42*, 60−66.
(33) Turner, D. B.; Tyrrell, S. M.; Willett, P. Rapid Quantification of Molecular Diversity for Selective Database Acquisition. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 18−22.
(34) Gillet, V. J.; Willett, P.; Bradshaw, J. The Effectiveness of Reactant Pools for Generating Structurally-Diverse Combinatorial Libraries. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 731−740.
(35) Jamois, E. A.; Hassan, M.; Waldman, M. Evaluation of Reagent-Based and Product-Based Strategies in the Design of Combinatorial Libraries. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 63−70.
(36) Gillet, V. J.; Nicolotti, O. New Algorithms for Compound Selection and Library Design. *Perspect. Drug Discovery Des.* In press.
(37) Robertson, A. M.; Willett, P. An Upperbound to the Performance of Ranked-Output Searching − Optimal Weighting of Query Terms Using A Genetic Algorithm. *J. Doc.* **1996**, *52*, 405−420.
(38) Gillet, V. J.; Willett, P.; Bradshaw, J. Identification of Biological Activity Profiles Using Substructural Analysis and Genetic Algorithms. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 165−179.
(39) Wagener, M.; Van Geerestein, V. J. Potential Drugs and Nondrugs: Prediction and Identification of Important Structural Features. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 280−292.
(40) Oprea, T. I. Property Distribution of Drug-Related Chemical Databases. *J. Comput.-Aided Mol. Des.* **2000**, *14*, 251−264.
(41) Gillet, V. J.; Willett, P.; Bradshaw, J.; Green, D. V. S. Selecting Combinatorial Libraries to Optimize Diversity and Physical Properties. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 169−177.
(42) Hansch, C.; Muir, R. M.; Fujita, T.; Maloney, P. P.; Geiger, F.; Streich, M. The Correlation of Biological Activity of Plant Growth Regulators and Chloromycetin Derivatives With Hammett Constants and Partition Coefficients. *J. Am. Chem. Soc.* **1963**, *85*, 2817−2824.
(43) Free, S. M.; Wilson, J. A Mathematical Contribution to Structure−Activity Studies. *J. Med. Chem.* **1964**, *7*, 395−399.
(44) Kowalski, B. R.; Bender, C. F. Pattern Recognition. A Powerful Approach to Interpreting Chemical Data. *J. Am. Chem. Soc.* **1972**, *94*, 5632−5639.
(45) Johnson, M. Relating Metrics, Lines and Variables Defined On Graphs to Problems in Medicinal Chemistry. In *Graph Theory and Its Applications to Algorithms and Computer Science, Proc. 5th Int. Conf. On The Theory of Graphs*; Alavi, Y., Chartrand, G., Lesniak, L., Lick, D. R., Wall, C. E., Eds.; Wiley: New York, 1985; p 457.
(46) Basak, S. C.; Magnuson, V. R.; Niemi, G. J.; Regal, R. R. Determining Structural Similarity of Chemicals Using Graph-Theoretic Indices. *Discrete Appl. Math.* **1988**, *19*, 17.
(47) *SAS*; SAS Institute Inc.: Cary, NC.
(48) Johnson, M.; Lajiness, M.; Maggiora, G. M. Molecular Similarity: A Basis for Designing Drug Screening Programs. In *QSAR: Quantitative Structure−Activity Relationships in Drug Design*; Fauchere, J. L., Ed.; Alan R. Liss: New York, 1989; pp 167−171.
(49) Lajiness, M. S.; Johnson, M. A.; Maggoria, G. M. Implementing Drug Screening Programs Using Molecular Similarity Methods. In *QSAR: Quantitative Structure−Activity Relationships in Drug Design*; Fauchere, J. L., Ed.; Alan R Liss Inc: New York, 1989; pp 173−176.
(50) Lajiness, M. S. Applications of Molecular Similarity/Dissimilarity in Drug Research. In *Structure−Property Correlations in Drug Research*; Waterbeemd, H. V. D., Ed.; R. G. Landes Company: Austin, TX, 1996; p 179.
(51) Lajiness, M. S. Dissimilarity-Based Compound Selection Techniques. *Perspect. Drug Discovery Des.* **1997**, *7/8*, 65−84.
(52) Maggiora, G. M.; Shanmugasundaram, V. *Similarity-Based Shannon-Like Diversity Measure. Theory*; National Meeting of the American Chemical Society, San Francisco, CA, 2000.
(53) Shanmugasundaram, V.; Maggiora, G. M. *A Similarity-Based Shannon-Like Diversity Measure. Application to Cell-Based Chemistry Spaces*; National Meeting of the American Chemical Society, San Francisco, CA, 2000.
(54) Simon, R. J.; Martin, E. J.; Miller, S. M.; Zuckermann, R. N.; Blaney, J. M.; Moos, W. H. Using Peptoid Libraries [Oligo N-Substituted Glycines] for Drug Discovery. In *Techniques in Protein Chemistry*; Crabb, J. W., Ed.; Academic Press: New York, 1994; Vol. V.

(55) Martin, E. J.; Blaney, J. M.; Siani, M. A.; Spellmeyer, D. C.; Wong, A. K.; Moos, W. H. Measuring Diversity: Experimental Design of Combinatorial Libraries for Drug Discovery. *J. Med. Chem.* **1995**, *38*, 1431−1436.
(56) Zuckermann, R. N.; Kerr, J. M.; Kent, S. B. H.; Moos, W. H. Efficient Method for the Preparation of Peptoids [Oligo-(N-Substituted Glycines)] By Submonomer Solid-Phase Synthesis. *J. Am. Chem. Soc.* **1992**, *114*, 10646−10647.
(57) Hansch, C.; Leo, A.; Hoekman, D. *Exploring QSAR: Hydrophobic, Electronic, and Steric Constants*; American Chemical Society: Washington, DC, 1995.
(58) Hall, L. H. *Molconn-X*, 1.0; Hall Associates: Quincy, MA, 1991.
(59) James, C. A.; Weininger, D. *Daylight Theory Manual*; Daylight Chemical Information Systems, Inc.: Irvine, CA, 1994.
(60) Blaney, J. M.; Dixon, J. S. Distance Geometry in Molecular Modeling. In *Reviews in Computational Chemistry*; Lipkowitz, K. B., Boyd, D. B., Ed.; VCH: New York, 1994; Vol. 5, pp 299−335.
(61) James, C. A.; Weininger, D.; Scofield, J. *Daylight Toolkit Programmer's Guide*; Daylight Chemical Information Systems, Inc.: Irvine, CA, 1994.
(62) Maggiora, G. M. *Chemical Applications of Fuzzy Mathematical Methods*; National Meeting of the American Chemical Society, Chicago, IL, 1993.
(63) Martin, E. J.; Critchlow, R. E. Beyond Mere Diversity: Tailoring Combinatorial Libraries for Drug Discovery. *J. Comb. Chem* **1999**, *1*, 32−45.
(64) Martin, E. J.; Spellmeyer, D. C.; Critchlow, R. E.; Blaney, J. M. Does Combinatorial Chemistry Obviate Computer Aided Drug Design? In *Reviews in Computational Chemistry*; Lipkowitz, K. B., Boyd, D. B., Eds.; VCH Publishers: New York, 1997; Vol. 10, pp 75−100.
(65) Martin, E. J.; Wong, A. Sensitivity Analysis and Other Improvements to Tailored Combinatorial Library Design. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 215−220.
(66) Martin, E. J.; Hoeffel, T. J. Oriented Substituent Pharmacophore Property Space (OSPPREYS): A Substituent-Based Calculation That Describes theLibrary Products Better Than the Corresponding Product-Based Calculation. *J. Mol. Graphics Model.* In press.
(67) Martin, E. J.; Critchlow, R. E.; Spellmeyer, D. C.; Rosenberg, S.; Spear, K. L.; Blaney, J. M. Diverse Approaches to Combinatorial Library Design. In *Pharmacochemistry Library; Vol. 29 (Trends in Drug Research II)*; Timmerman, H., Ed.; Elsevier: Amsterdam, 1998; pp 133−146.
(68) Hansch, C.; Unger, S. H.; Forsythe, A. B. Strategy in Drug Design. Cluster Analysis as an Aid in the Selection of Substituents. *J. Med. Chem.* **1973**, *16*, 1212−1222.
(69) Martin, Y. C. *Quantitative Drug Design*; Dekker: New York, 1978.
(70) Martin, Y. C.; Panas, H. N. Mathematical Considerations in Series Design. *J. Med. Chem.* **1979**, *22*, 784−791.
(71) Wooton, R.; Cranfield, R.; Sheppey, G. C.; Goodford, P. J. Physicochemical-Activity Relationships in Practice. 2. Rational Selection of Benzenoid Substituents. *J. Med. Chem.* **1975**, *18*, 607−613.
(72) Lin, C. T.; Pavlik, P. A.; Martin, Y. C. Use of Molecular Fields to Compare Series of Potentially Bioactive Molecules Designed By Scientists Or By Computer. *Tetrahedron Comput. Methodol.* **1990**, *3*, 723−738.
(73) Austel, V. Experimental Design in Synthesisis Planning and Structure−Property Correlations. In *Chemometric Methods in Molecular Design*; Van De Waterbeemd, H., Ed.; VCH: Weinheim, 1995; pp 49−62.
(74) Jarvis, R. A.; Patrick, E. A. Clustering Using A Similarity Measure Based On Shared Nearest Neighbors. *IEEE Trans. Comput.* **1973**, *C-22*, 1025−1034.
(75) Blaney, J. M.; Capobianco, P.; Eyermann, C. J. Clustering the Dupont Compound Collection. Personal communication.
(76) Martin, Y. C.; Bures, M. G.; Willett, P. Searching Databases of Three-Dimensional Structures. In *Reviews in Computational Chemistry*; Lipkowitz, K. B., Boyd, D. B., Eds.; VCH: New York, 1990; Vol. 1, pp 213−263.
(77) Martin, Y. C.; Bures, M. G.; Danaher, E. A.; Delazzer, J.; Lico, I.; Pavlik, P. A. A Fast New Approach to Pharmacophore Mapping and Its Application to Dopaminergic and Benzodiazepine Agonists. *J. Comput.-Aided Mol. Des.* **1993**, *7*, 83−102.
(78) Brown, R.; Bures, M.; Martin, Y.; Pavlik, P. 3D Property Based Precursor Selection for Combinatorial Library Construction. *Abstracts of Papers*, 210th National Meeting of the American Chemical Society; American Chemical Society: Washington, DC, 1995; 24-CINF.
(79) Martin, Y. C.; Brown, R. D.; Bures, M. G. Quantifying Diversity. In *Combinatorial Chemistry and Molecular Diversity in Drug Discovery*; Kerwin, J. F., Gordon, E. M., Ed.; Wiley: New York, 1998; pp 369−385.
(80) 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.* **1997**, *37*, 1−9.
(81) Brown, R. D.; Martin, Y. C. Designing Combinatorial Library Mixtures Using A Genetic Algorithm. *J. Med. Chem.* **1997**, *40*, 2304−2313.
(82) Gasteiger, J.; Ihlenfeldt, W. D.; Röse, P.; Wanke, R. Computer-Assisted Reaction Prediction and Synthesis Design. *Anal. Chim. Acta* **1990**, *235*, 65−75.
(83) Höllering, R.; Gasteiger, J.; Steinhauer, L.; Schulz, K.-P.; Herwig, A. The Simulation of Organic Reactions: From the Degradation of Chemicals to Combinatorial Synthesis. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 482−492.
(84) Ihlenfeldt, W.-D.; Gasteiger, J. Computer-Assisted Planning of Organic Syntheses: The Second Generation of Programs. *Angew. Chem., Int. Ed. Engl.* **1995**, *34*, 2613−2633.
(85) Gasteiger, J.; Ihlenfeldt, W.-D.; Fick, R.; Rose, J. R. Similarity Concepts for the Planning of Organic Reactions and Syntheses. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 700−712.
(86) Ihlenfeldt, W.-D.; Gasteiger, J. Hash Codes for the Identification and Classification of Molecular Structure Elements. *J. Comput. Chem.* **1994**, *15*, 793−813.
(87) Parlow, A.; Weiske, C.; Gasteiger, J. Cheminform − an Integrated Information System On Chemical Reactions. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 400−402.
(88) Gasteiger, J.; Marsili, M. Interactive Partial Equalization of Orbital Electronegativity - - A Rapid Access to Atomic Charges. *Tetrahedron* **1980**, *36*, 3219−3288.
(89) Hutchings, M. G.; Gasteiger, J. Residual Electronegativity − an Empirical Quantification of Polar Influences and Its Application to the Proton Affinity of Amines. *Tetrahedron Lett.* **1983**, *24*, 2541−2544.
(90) Gasteiger, J.; Saller, H. Calculation of the Charge Distribution in Conjugated Systems By A Quantification of the Resonance Concept. *Angew. Chem., Intl. Ed. Engl.* **1985**, *24*, 687−689.
(91) Hutchings, M. G.; Gasteiger, J. A Quantitative Description of Fundamental Polar Reaction Types. Proton and Hydride Transfer Reactions Connecting Alcohols and Carbonyl Compounds in the Gas Phase. *J. Chem. Soc., Perkin Trans. 2* **1986**, 447−454.
(92) Gasteiger, J. Automatic Estimation of Heats of Atomization and Heats of Reaction. *Tetrahedron* **1979**, *35*, 1419−1426.
(93) Gasteiger, J. http://www2.ccc.uni-erlangen.de/software/petra/index.html.
(94) Rose, J. R.; Gasteiger, J. HORACE: An Automatic System for the Hierarchical Classification of Chemical Reactions. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 74−90.
(95) Chen, L.; Gasteiger, J. Knowledge Discovery in Reaction Databases: Landscaping Organic Reactions By A Self-Organizing Neural Network. *J. Am. Chem. Soc.* **1997**, *119*, 4033−4032.

(96) Sadowski, J.; Gasteiger, J.; Klebe, G. Comparison of Automatic Three-Dimensional Model Builders Using 639 X-Ray Structures. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1000−1008.
(97) Zupan, J.; Gasteiger, J. *Neural Networks in Chemistry and Drug Design*, 2nd ed.; Wiley-VCH: Weinheim, 1999.
(98) Wagener, M.; Sadowski, J.; Gasteiger, J. Autocorrelation of Molecular Surface Properties for Modeling Coricosteroid Binding Globulin and Cytosolic AH Receptor Activity By Neural Networks. *J. Am. Chem. Soc.* **1995**, *117*, 7769−7775.
(99) Sadowski, J.; Wagener, M.; Gasteiger, J. Assessing Similarity and Diversity of Combinatorial Libraries By Spatial Autocorrelation Functions and Neural Networks. *Angew. Chem., Intl. Ed. Engl.* **1995**, *34*, 2674−2677.
(100) Gasteiger, J.; Handschuh, S.; Hemmer, M. C.; Kleinöder, T.; Schwab, C. H.; Teckentrup, A.; Sadowski, J.; Wagener, M. 3D Structure Descriptors for Biological Activity. In *Molecular Modelling and Prediction of Bioactivity*; Gundertofte, K., Jorgensen, F. S., Eds.; Kluwer Academic Plenum Press: New York, 2000; pp 157−168.
(101) Satoh, H.; Sacher, O.; Nakata, T.; Chen, L.; Gasteiger, J.; Funatsu, K. Classification of Organic Reactions: Similarity of Reactions Based On Changes in the Electronic Features of Oxygen Atoms At the Reaction Sites. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 210−219.
(102) Gasteiger, J.; Pförtner, M.; Sitzmann, M.; Höllering, R.; Sacher, O.; Kostka, T.; Karg, N. Computer-Assisted Synthesis and Reaction Planning in Combinatorial Chemistry. *Perspect. Drug Discovery Des.* **2000**, *20*, 1−21.
(103) Ferguson, A. M.; Patterson, D. E.; Garr, C. D.; Underiner, T. L. Designing Chemical Libraries for Lead Discovery. *J. Biomol. Screening* **1996**, *1*, 65−73.
(104) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and Computational Approaches to Estimate Solubility and Permeability in Drug Discovery and Development Settings. *Adv. Drug Delivery Rev.* **1997**, *23*, 3−25.
(105) Brown, R. D.; Bures, M. G.; Martin, Y. C. Similarity and Cluster-Analysis Applied to Molecular Diversity. *Abstracts of Papers*, 209th National Meeting of the American Chemical Society; American Chemical Society: Washington, DC, 1995; 3-COMP.
(106) 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.* **1996**, *39*, 3060−9.
(107) 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.* **1996**, *39*, 3049−59.
(108) Cramer, R. D. BC(DEF) Parameters. 1. The Intrinsic Dimensionality of Intramolecular Interactions in the Liquid State. *J. Am. Chem. Soc.* **1980**, *102*, 1837−1849.
(109) Cramer, R. D.; Patterson, D. E.; Clark, R. D.; Soltanshahi, F.; Lawless, M. S. Virtual Libraries: A New Approach to Decision Making in Molecular Discovery Research. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 1010−1023.
(110) Cramer, R. D.; Poss, M. A.; Hermsmeier, M. A.; Caulfield, T. J.; Kowala, M. C.; Valentine, M. T. Prospective Identification of Biologically Active Structures By Topomer Shape Similarity Searching. *J. Med. Chem.* **1999**, *42*, 3919−3933.
(111) Andrews, K. M.; Cramer, R. D. Toward General Methods of Targetted Library Design: Topomer Shape Similarity With Diverse Structures as Queries. *J. Med. Chem.* **2000**, *43*, 1723−1740.
(112) Pearlman, R. S.; Zhu, H.; Stewart, E. L.; Brusniak, M. Y. K. *Combilibmaker: Software for Generating Combinatorial Libraries of 2D and 3D Molecular Structures (Initially Known as Combindbmaker)*; Tripos: St. Louis, MO, 1993. Copyright, University of Texas: Austin, TX.
(113) Pearlman, R. S.; Smith, K. M. *Diverse Solutions: Novel Software for Chemical Diversity, Library Design, and Accelerated Drug Discovery (Initially Known as Diverseselector)*; Tripos, Inc: St. Louis, MO, 1994.
(114) Pearlman, R. S.; Smith, K. M. Novel Software Tools for Chemical Diversity. *Perspect. Drug Discovery Des.* **1998**, *9*, 339−353.
(115) Pearlman, R. S.; Smith, K. M. Software for Chemical Diversity in the Context of Accelerated Drug Discovery. *Drugs Future* **1998**, *23*, 885−895.
(116) Pearlman, R. S.; Smith, K. M. Metric Validation and the Receptor-Relevant Subspace Concept. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 28−35.
(117) Pickett, S.; Mclay, I.; Lewis, R.; Mason, J. A Novel Method for Evaluating Molecular Diversity Within A Compound Library: 3D-Pharmacophore-Derived Queries. *Chem. Des. Autom. News* **1995**, *10* (4), 38−39.
(118) Pickett, S. D.; Mason, J. S.; Mclay, I. M. Diversity Profiling and Design Using 3D Pharmacophores − Pharmacophore-Derived Queries (PDQ). *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 1214−1223.
(119) Davies, E. K. In *Molecular Diversity and Combinatorial Chemistry: Libraries and Drug Discovery*; American Chemical Society: Washington, DC, 1996; pp 309−316.
(120) Mason, J. S.; Pickett, S. D. Partition-Based Selection. *Perspect. Drug Discovery Des.* **1998**, *7/8*, 85−114.
(121) Mason, J. S.; Cheney, D. L. Ligand−Receptor 3-D Similarity Studies Using Multiple 4-Point Pharmacophores. In *Biocomputing '99*, Proceedings of the Pacific Symposium; Altman, R. B., Lauderdale, K., Dunker, A. K., Hunter, L., Klein, T., Eds.; World Scientific Publishing: Singapore, pp 456−467. Available on-line at http://www.smi.stanford.edu/projects/helix/psb99/Mason.pdf.
(122) Mason, J. S.; Morize, I.; Menard, P. R.; Cheney, D. L.; Hulme, C.; Labaudiniere, R. F. A 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 Privileged Substructures. *J. Med. Chem.* **1999**, *42*, 3251−3264.
(123) Mason, J. S.; Absolute vs Relative Similarity and Diversity. In *Molecular Diversity*; Dean, P. M., Lewi, R. A., Eds.; Kluwer Academic Publishers: Dordrecht, 1999.
(124) Mason, J. S.; Cheney, D. L. *Library Design and Virtual Screening Using Multiple 4-Point Pharmacophore Fingerprints*; Proc. Pacific Symposium on Biocomputing; in press.
(125) Pickett, S. D.; Mclay, I. M.; Clark, D. E. Enhancing the Hit-to-Lead Properties of Lead Optimization Libraries. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 263−272.
(126) Murray, C. M.; Cato, S. J. Design of Libraries to Explore Receptor Sites. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 46−50.
(127) Mason, J. S.; Cheney, D. L. Recent Advances in Pharmacophore Similarity in Structure-Based Drug Design. *Abstracts of Papers*, National Meeting of the American Chemical Society, Dallas, TX, 1998; American Chemical Society: Washington, DC, 1998.
(128) Lewis, R. A.; Mason, J. S.; Mclay, I. M. Similarity Measures for Rational Set Selection and Analysis of Combinatorial Libraries − the Diverse Property-Derived (DPD) Approach. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 599−614.
(129) Menard, P. R.; Lewis, R. A.; Mason, J. S. Rational Screening Set Design and Compound Selection: Cascaded Clustering. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 497−505.
(130) Van Drie, J. H.; Weininger, D.; Martin, Y. C. ALADDIN: an Integrated Tool for Computer-Assisted Molecular Design and Pharmacophore Recognition From Geometric, Steric, and Substructure Searching of Three-Dimensional Molecular Structures. *J. Comput.-Aided Mol. Des.* **1989**, *3*, 225−251.

(131) Pearlman, R. *Combinlibmaker and DiverseSolutions*; Distributed By Tripos, Inc., St. Louis, MO; Copyright, University of Texas: Austin, TX, 1996.
(132) Mason, J. S.; Beno, B. R. Library Design Using BCUT Chemistry Descriptors and Multiple 4-Point Pharmacophore Fingerprints − Simultaneous Optimization and Structure-Based Diversity. *J. Mol. Graphics Model.* In press.
(133) Mason, J. S.; Hermsmeier, M. A. Diversity Assessment. *Curr. Opin. Chem. Biol.* **1999**, *3*, 342−349.
(134) Sadowski, J.; Kubinyi, H. A Scoring Scheme for Discriminating Between Drugs and Nondrugs. *J. Med. Chem.* **1998**, *41*, 3325−3329.
(135) Ajay; Walters, W. P.; Murcko, M. A. Can We Learn to Distinguish Between "Drug-Like" and "Nondrug-Like" Molecules? *J. Med. Chem.* **1998**, *41*, 3314−3324.

CC000073E