Articles

Chemography: The Art of Navigating in Chemical Space

Tudor I. Oprea*,† and Johan Gottfries‡

EST Lead Informatics and Medicinal Chemistry, AstraZeneca R&D Mölndal, S-43183 Mölndal, Sweden

Received May 10, 2000

Combinatorial chemistry needs focused molecular diversity applied to the druglike chemical space (drugspace). A drugspace map can be obtained by systematically applying the same conventions when examining the chemical space, in a manner similar to the Mercator convention in geography: Rules are equivalent to dimensions (e.g., longitude and latitude), while structures are equivalent to objects (e.g., cities and countries). Selected rules include size, lipophilicity, polarizability, charge, flexibility, rigidity, and hydrogen bond capacity. For these, extreme values were set, e.g., maximum molecular weight 1500, calculated negative logarithm of the octanol/water partition between −10 and 20, and up to 30 nonterminal rotatable bonds. Only S, N, O, P, and halogens were considered as elements besides C and H. Selected objects include a set of “satellite” structures and a set of representative drugs (“core” structures). Satellites, intentionally placed outside drugspace, have extreme values in one or several of the desired properties, while containing druglike chemical fragments. ChemGPS (chemical global positioning system) is a tool that combines these predefined rules and objects to provide a global drugspace map. The ChemGPS drugspace map coordinates are *t*-scores extracted via principal component analysis (PCA) from 72 descriptors that evaluate the above-mentioned rules on a total set of 423 satellite and core structures. Global ChemGPS scores describe well the latent structures extracted with PCA for a set of 8599 monocarboxylates, a set of 45 heteroaromatic compounds, and for 87 α-amino acids. ChemGPS positions novel structures in drugspace via PCA-score prediction, providing a unique mapping device for the druglike chemical space. ChemGPS scores are comparable across a large number of chemicals and do not change as new structures are predicted, making this tool a well-suited reference system for comparing multiple libraries and for keeping track of previously explored regions of the chemical space.

Introduction

Combinatorial chemistry is a rapidly evolving technology that has become a method of choice in drug discovery.[1] Using combinatorial chemistry, one can synthesize libraries of compounds on the order of 10^2−10^9 structures.[2,3] Initially, the numbers game was advocated, i.e., maximize the number of compounds synthesized in a given amount of time, then screen that library for everything or anything using high throughput screening (HTS). Today, most chemists agree that big numbers and serendipity are not enough.[4] As the range of synthetic possibilities, as well as the number of commercially available compounds, is increasing every day, the process of compound selection and prioritization has become crucial.[5,6] Selection involves the evaluation of molecular diversity, and a number of tools are available for this process.[7−10] The selection step is, so far, based on *local* models of the potential reactants and/or reagents, from which a subset is subsequently chosen.

The combinatorial problem of using amino acids as building blocks for combinatorial chemistry was investigated by Hellberg et al.[11] Even though the syntheses were constrained to using the 20 essential amino acids to yield di- to penta-peptides as the products, this still resulted in several thousands of possibilities. Rather than making vast numbers of (similar) products, selection tools were developed[11] via principal components analysis[12] (PCA), using appropriate descriptors. Following the pioneering work of Volkhaard Austel in the design of experiments,[13] this approach addressed the selection step in a rational manner, using D-optimal[14] or factorial design[15] for the subset of peptides to be synthesized. As soon as an expansion of the product space was demanded, i.e., when novel types of α-carbon side chains were required for peptide synthesis (55 compounds in total), a new set of calculations and selections had to be performed.[16] This was later expanded with an additional 32 α-amino acids set, which were selected to represent both intermediate and extreme physicochemical properties, compared to the 20 essential amino acids.[17]

When existing models become inappropriate, novel de-

* To whom correspondence should be addressed. E-mail: tudor.oprea@astrazeneca.com.

† EST Lead Informatics.

‡ Medicinal Chemistry.

10.1021/cc0000388 CCC: $20.00 © 2001 American Chemical Society
Published on Web 02/15/2001

scriptors and/or compounds must be included. These local models need to be rederived after each expansion. Furthermore, local models are not easily amenable to comparison, in particular when the descriptors used in the PCA models are not similar. For example, the principal properties for the amino acids[17] derived by PCA cannot be directly compared with the principal properties for heteroaromatic compounds,[18] also derived by PCA. This problem is of increasing importance when comparing various compound libraries, for example when one tries to fill in the diversity voids that exist in the in-house compound collection.[19] Such libraries, HTS libraries in particular, may have millions of compounds. PCA algorithms are relatively fast, but 10^5-10^6 rows by 100 column matrixes are rather difficult to handle, even for fast computers.

Local models have several shortcomings: *(i)* predictivity is limited by the distribution and chemical diversity of the training set;[20,21] *(ii)* models need to be recalculated whenever compounds are added or removed—in particular if the external predictivity of the local model is deficient in accuracy; *(iii)* by being local, such models are not easily amenable to comparison with other models, even when the same descriptors are used. There is a clear need for *global*, rather than local models. The first barrier to overcome is *extrapolation*, since it is the major source for recalculations in the field of chemically related developments.[22] One cannot a priori predict what structures to include in the next synthetic library. An ideal (global) model that would handle novel compounds via *interpolation*, not extrapolation, is rather unlikely using the current methods.

We therefore suggest the term *chemography*, by analogy with geography, as the art of navigating in chemical space. Similar to the conventional mapping systems used in geography, e.g., Mercator, chemography requires a standard convention for chemical space navigation. In geography, the existence of such conventional systems allows one to project on the same plane various objects located on a geosphere. Such conventions include a set of *rules* (e.g., meridians and parallels), and set of *objects* (e.g., mountains, cities, countries, etc.). In chemography, the rules are provided by the principal properties derived via PCA, whereas the objects are represented by molecular structures. In a manner similar to the satellites used in the Navstar global positioning system[23] (GPS), one can define a set of chemical structures intentionally located *outside* the chemical space of interest. For medicinal chemistry applications, this would be represented by the “druglike” chemical space (drugspace), i.e., the chemical space occupied by druglike molecules.[24,25] Therefore, “satellite molecules” would be chemical structures having at least one property value located outside the property range defined by the known drugspace. The ChemGPS procedure, introduced in the present paper, provides a standard tool for compound selection within the same PCA model using one training set, i.e., the ChemGPS set. The principal properties of novel compounds are predicted rather than recomputed, the same way as in local model forecasting, e.g., via PCA score predictions. The resulting ChemGPS scores provide a unique and potentially standard metric for the chemical space and, as such, are directly amenable to comparative analyses across chemistry and time.

Materials and Methods

Chemical Structure Construction and Selection. Molecules with extreme properties were selected from in-house and commercial databases, as described elsewhere.[26] Virtual molecules that represent objects with extreme properties, e.g., “hexazole”, a six-member nitrogen ring, were also included. These structures are referred to as “satellites” since they are located, in the principal property space, outside the drugspace. Examples of satellite structures are, as shown in Figure 1: sucrose (**1**), glycerin (**2**), benzene (**3**), *p*-amidinobenzamidine (**4**), the L-arginine tetramer (**5**), erythromycin (**6**), the L-tryptophan tetramer (**7**), tetra-phenyl adamantane (**8**), and the guanine nucleotide tetramer, GGG (**9**). A second class of compounds, “core structures”, was required to maintain the inner balance of the PCA model and to keep the model focused on the drugspace. These compounds, filling the core of the drugspace, were selected from a list of known registered drugs by taking into account their intestinal permeability properties (e.g., human intestinal absorption[27] above 10%), as well as other molecules patented in drug-related applications—see Figure 2. Some of the compounds intended for medical use were deemed as satellites rather than core structures, according to their properties (e.g., cyclosporine and methotrexate). For the sake of simplicity, only S, N, O, P, and halogens were considered as elements, besides carbon and hydrogen.

Molecular Descriptors. The following druglike properties were considered as intuitive and were represented in our molecular descriptor set: size, lipophilicity, polarizability, charge, flexibility, rigidity, and hydrogen bond capacity. Some of these have been chosen to match the QSAR paradigm for structure—permeability correlations[27] in an effort to capture properties relevant to oral drug absorption.[28] Size-related descriptors included molecular weight (MW), the number of heavy atoms, the number of carbons, and the calculated molecular refractivity, CMR.[29] Polarizability was estimated by CMR and by an atom-based polarizability scheme[30] implemented in SaSA.[31] Flexibility and rigidity were, in turn, estimated by counting the total number of bonds and rings (RNG), the number of rotatable bonds (RTB), and the number of rigid bonds (RGB)[32] and by several topological indices that estimate other properties[33] as well, e.g., size. The Wiener, Balaban, Randic, and Motoc indices, as well as the Kier and Hall suite of topological descriptors, were used.[34] Hydrogen-bonding capacity was estimated using four HYBOT[35] descriptors: HDOM, the maximum free energy H-bond donor factor (C_d); HDOS, the sum of C_d values; HACM, the maximum free energy H-bond acceptor factor (C_a); HACS, the sum of C_a values. All C_d values were given a positive sign, as previously suggested.[36] In addition, we have used the simple count of oxygens, nitrogens, H-bond donors (HDO), and H-bond acceptors (HAC), as implemented in SaSA.[31] Charge was estimated by counting the positive (N_POS) and negative (N_NEG) ionization centers, as well as the maximum positive and

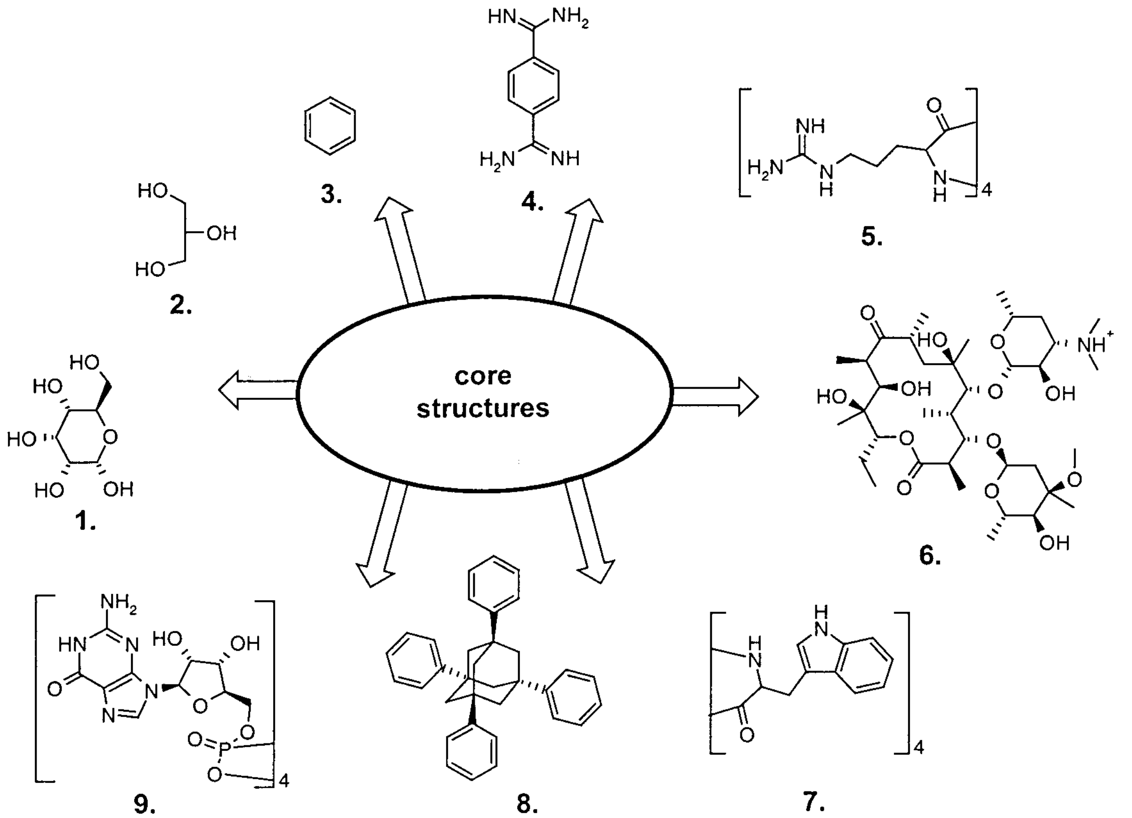

Figure 1. Typical satellites for drugspace.

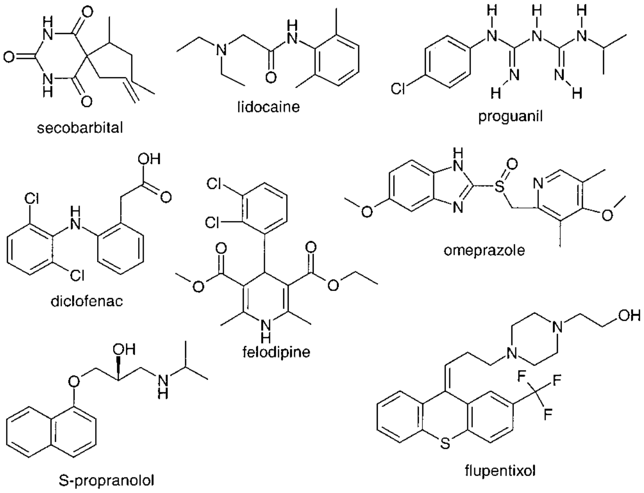

Figure 2. Examples of core structures.

negative charge, as calculated using the Gasteiger−Marsili method.[37] Lipophilicity was estimated using two methods that calculate the logarithm of the octanol/water partition coefficient:[38] CLOGP[39] and ACDLogP.[40] No missing values were encountered during the implementation of the ChemGPS training set (423 compounds). Overall, less than 1% of the entire data set (over 22 000 compounds) exhibited missing values, and in not more than two columns.

Statistics. All multivariate models were obtained using PCA,[12] as implemented in the SIMCA[41] package. The estimation of principal component (PC) significance was performed by the cross validation (CV) procedure[42] and provided as Q^2. The number of PCs in the ChemGPS model was decided by CV, with the additional criterion that any descriptor should load in at least one PC. The PCA modeling was interrupted whenever a PC included only loadings from descriptors that had already contributed high leverage in a previous PC, despite significant CV-based contributions for individual descriptors, which were not significant for the total explained variance.[42] The probability of the compounds to belong to the model, PModXPS, was calculated[41] in SIMCA. All compounds with a probability of belonging of less than 5% (PModXPS $<$ 0.05) were considered outliers, i.e., they were significantly different from the compounds used to build the model. Additional model validation, e.g., comparing local vs global *t*-scores, was performed using the projections to latent structures (PLS) method[43,44] implemented in SIMCA.[41] PLS component significance was estimated by the CV procedure[42] and provided as Q^2 (predictive R^2 according to CV using seven randomized groups). The significance of Q^2 and R^2 (fraction of explained variance) in QSAR models has been discussed elsewhere.[45]

Because several thousands of compounds were predicted as outliers in the initial stages, the task of selecting new satellites had to be rationalized. Rational selection of novel satellites and/or core structures by means of experimental design was performed in order to avoid choosing too-similar compounds as satellites (since these would appear as outliers in previous ChemGPS models). Therefore, random selections from large databases (e.g., ACD[46] and MDDR[46]) were made, and all descriptors were calculated for each selected compound. ChemGPS scores were predicted, and the probability of the compounds to belong to the ChemGPS model, PModXPS, was calculated.[41] All compounds with PModXPS $<$ 0.05 were submitted to a statistical molecular design[47] analysis, i.e., selecting a subset via the D-optimal algorithm. The outliers selected using the D-optimal criterion were

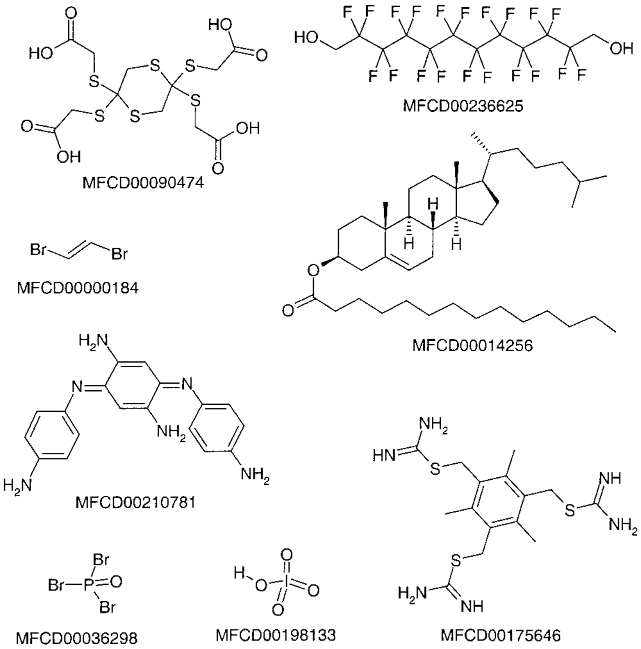

Figure 3. Selected tentative satellite compounds from those with largest distance to the ChemGPS model ($n = 23$).

individually scrutinized, then appropriate ones (see Figure 3) were included in a new version of the ChemGPS. This procedure was iterated three times. In choosing a lower, rather than larger, number of satellites, we have applied the parsimony principle ("less is better").

Results and Discussion

ChemGPS—A Heuristic Approach toward a Convergent Prediction System. The objective of chemography is to provide a consistent mapping device, namely the chemical global positioning system, ChemGPS,[48] that can avoid extrapolations when positioning the properties of a new arbitrary collection of leadlike or druglike organic molecules. This can be achieved when the principal property space of drugspace is well covered, on all parts, by relevant satellite structures (Figure 1). As this was not the case in the initial ChemGPS model, outliers, i.e., extrapolations, were detected and analyzed. Outliers that held interesting extreme property values in one or several directions were tested as new putative satellites (as described in the Methods section), with the aim to find appropriate molecules that would enhance the ChemGPS coverage of the drugspace and yield a convergent ChemGPS system. The appropriate outliers, selected using the D-optimal criterion (see Figure 3), were included in new versions of the ChemGPS, which heuristically approached convergence in three steps, as described in the previous section. Thus, the current ChemGPS model can predict property positioning in chemical space without extrapolation, as tested on over 22 000 compounds, of which only 23 were deemed outliers according to the PModXPS test (data not shown). We note that some of the outliers depicted in Figure 3 were not included in the current ChemGPS model (e.g., the polyfluorurated diol or the *n*-alkyl-substituted steroid), whereas other structures are currently included (e.g., the tetra-carboxylated thioether or the periodic acid).

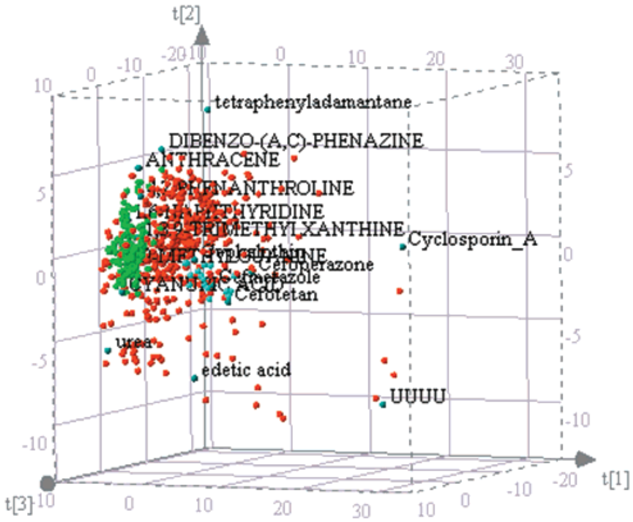

Figure 4. Projection on the three most significant dimensions (*t*1, *t*2, and *t*3, respectively) of the ChemGPS training set (satellites and core structures) in red, Cephalosporines in blue, and heteroaromatic compounds in green.

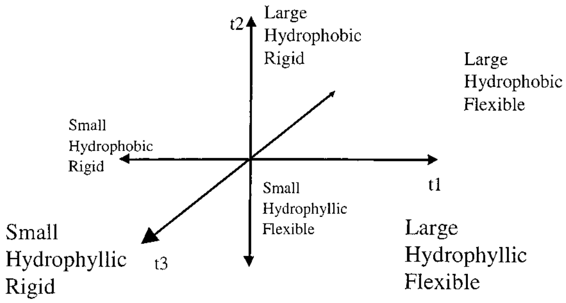
t2
Large Hydrophobic Rigid
Large Hydrophobic Flexible
Small Hydrophobic Rigid
t1
Small Hydrophyllic Flexible
Small Hydrophyllic Rigid
t3
Large Hydrophyllic Flexible

Figure 5. Principal property translation into interpretable chemical descriptors can be retrieved from the ChemGPS loadings, as exemplified for the three most significant dimensions (*t*1, *t*2, and *t*3, respectively).

The present ChemGPS data set consists of 423 virtual and existing structures (see Figure 4). The PCA modeling was terminated at nine PCs, after inspection of CV and loading vectors, as described in the Methods section. The present set of compounds and descriptors can be used for clustering overview (Figure 4) and property interpretation via the PCA loading vectors, as visualized in Figure 5. Because core compounds were deliberately selected to sample the chemical space occupied by orally available drugs (e.g., human intestinal permeability[27] above 10%), these structurally diverse compounds are not divided into major clusters but are rather homogeneously distributed in chemical space.

The advantage of ChemGPS is that, within one chemographic metric, diverse compounds can be compared to each other via simple prediction routines, in particular, in the area of drug discovery. The chemical space map is expected to evolve in time as more informative and complex descriptors are included, in particular, those that are CPU-intensive today but could be easily computed in the near future. As ChemGPS is used to predict more and more compounds, new outliers will be accumulated. By applying the statistical molecular design procedure, novel satellites will be included in future versions to improve the prediction accuracy. This indicates that ChemGPS is sometimes subject to the same shortcomings as local models, since PCA scores need to be

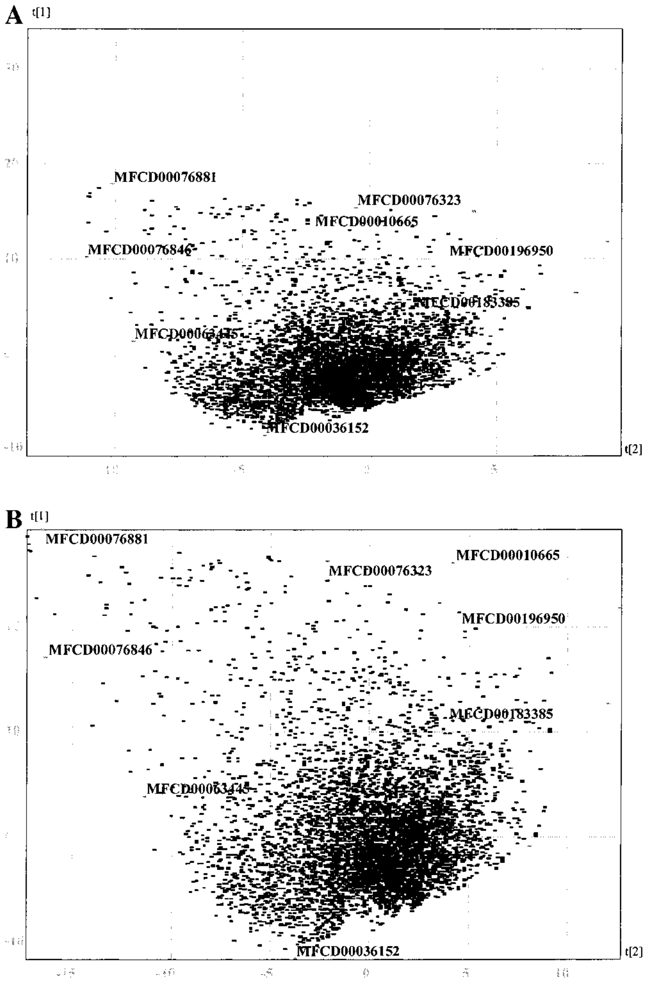

Figure 6. Principal properties for 8599 monocarboxylic acids, as obtained from the *t*-scores of a local PCA that used the carboxylates (A), and from the ChemGPS prediction (B).

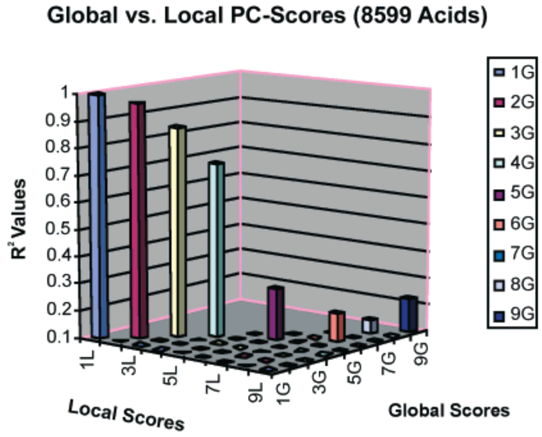

Figure 7. Fraction of explained variance (R^2) for the 1:1 correlation between the predicted *t*-scores obtained with ChemGPS (marked 1G−9G), and the *t*-scores derived from the local PCA model (marked 1L−9L), for the 8599 monocarboxylic acids. The first four dimensions yielded significant correlations: 1L−1G (0.998), 2L−2G (0.963), 3L−3G (0.873), and 4L−4G (0.74), where R^2 values are given in brackets.

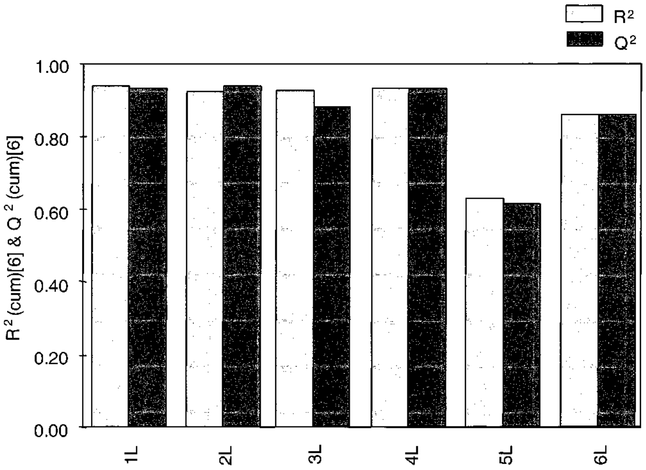

Figure 8. Summary of the six-components PLS model derived to explain the first six (1L−6L) local PCA scores, using the nine (1G−9G) scores from ChemGPS as *X* variables. For this model, $R^2X = 0.897$, $R^2Y = 0.894$, $Q^2 = 0.893$, and $N = 8599$.

rederived once the internal list of objects (the training set) has changed.

Combinatorial Chemistry Reagent Selection Using ChemGPS. Reagent selection can be performed using ChemGPS, as exemplified with a selection of monocarboxylic acids. Principal properties for a random set of 8599 structures extracted from the ACD database[46] were analyzed using both local PCA (Figure 6A) and ChemGPS predictions (Figure 6B). The shape of the cluster and the relative positions of individual acids within the cluster are preserved in both models (Figure 6A,B) for the first four dimensions, as can be observed from the Figure 7. In addition to the (expected) 1:1 correspondence between ChemGPS scores and local PCA scores, a significant PLS model that links the first six dimensions in the local model with the ChemGPS scores is summarized in Figure 8. The fact that the global and local dimensions are no longer correlated in the fifth (and higher) dimensions can be rationalized as follows: both local and global models capture relevant latent structures in the lower dimensions (e.g., size, lipophilicity, flexibility). Higher dimensions, however, are attributed different meaning (i.e., descriptors load differently) in global models, compared to local models. More specifically, whereas all carboxylates have the COOH moiety in common (not captured by the local model), this is not true for the global model. Thus one can expect that negative vs positive charges, or polarizability, will be treated differently.

In ChemGPS, one can examine monocarboxylic acids and, by extension, reagents in general by studying the ChemGPS score in a global map of the chemical space. The first four dimensions have a 1:1 relationship, when comparing local (PCA) and global (ChemGPS-predicted) *t*-score values. Local PCA models can be used for external prediction, in the same way as the ChemGPS model. However, local model predictions may be less trustworthy when switching from, e.g., carboxylic acids to amines, whereas ChemGPS scores are directly comparable across different classes of compounds. Thus, adding or removing compounds from the reagent list does not influence the ChemGPS score prediction. This opens up new possibilities for interactive reagent selection, e.g., pending on availability, price, purity, chirality, etc. By contrast, local PCA models may need to be recomputed when the composition of the compound set is changed.

Comparison to 3D-Based Principal Properties. A set of 40 heteroaromatic compounds was multivariately char-

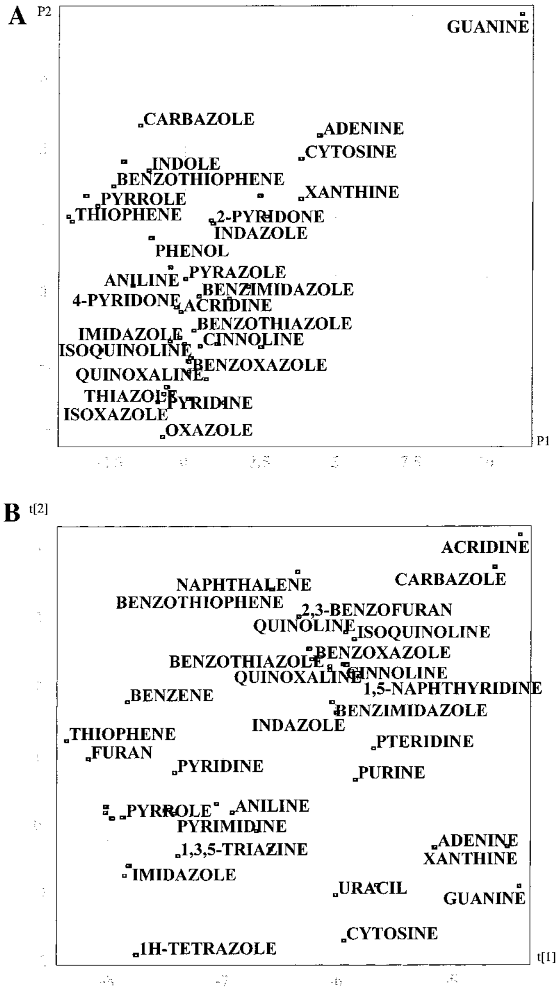

Figure 9. Principal properties for 45 heteroaromatics, as obtained from the *t*-scores of a local PCA model based on GRID properties (A) and from the ChemGPS prediction (B).

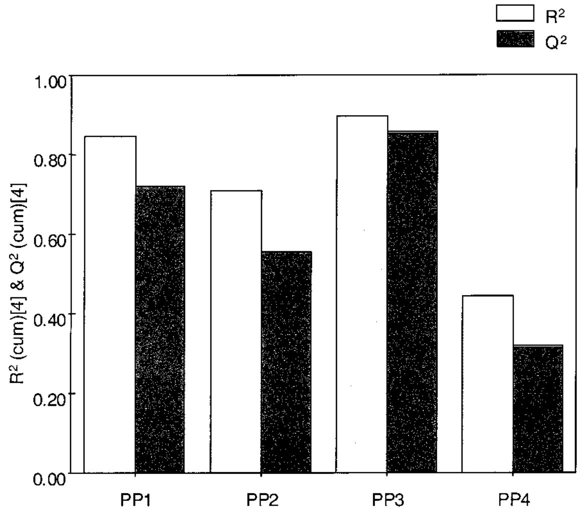

Figure 10. Summary of the four-component PLS model derived to explain the GRID principal property scores (PP1−PP4), using the nine-dimensional (1G−9G) scores from ChemGPS as *X* variables. For this model, $R^2X = 0.881$, $R^2Y = 0.725$, $Q^2 = 0.602$, and $N = 44$. Guanine was excluded from this correlation, since it was classified as an outlier by GRID (see also Figure 9A).

acterized[18] using 13 descriptors derived by GRID.[49] Four principal properties were extracted from the initial GRID data, the first two being graphically reconstructed in Figure 9A. Compared with the ChemGPS predictions (Figure 9B), a similar map is obtained—except for the apparent rotation of the PCA solution in the PC1 and PC2 plane. Guanine, an outlier in the first two dimensions, was excluded from the final analysis by Clementi and co-workers,[18] probably due to errors in the GRID calculations[50] (see also Figure 9A), but was well integrated into the xanthine cluster according to ChemGPS (Figure 9B). No significant 1:1 correspondence between ChemGPS scores and GRID principal properties was observed (data not shown). This is not surprising since, in contrast to the monocarboxylates example discussed above, a different set of molecular descriptors (i.e., GRID) was applied to characterize the compounds in the original study.[18] However, we found a significant PLS model that explains the four principal properties in GRID (here, used as *Y* variables), starting from the global (1G-9G) ChemGPS scores; see Figure 10 for details. This indicates that latent structures specific for these heteroaromatic compounds are well captured, albeit via different descriptors, i.e., 3D-based in GRID vs 2D-based in ChemGPS. The need for three-dimensional models appears to be less necessary for these molecules, since they are all rigid and flat (aromatic) ring systems.

Multivariate Characterization of α-Amino Acids. A set of 87 amino acids was multivariately characterized[17] using 26 descriptors. The five principal properties that were extracted from the initial data were later validated in quantitative sequence−activity models for elastase and neurotensin analogues.[17] The first two components of a local PCA model for the 87 amino acids, using the same descriptors as in ChemGPS, are shown in Figure 11A. The ChemGPS training set contains 10 α-amino acids: alanine, arginine, cysteine, glutamate, glycine, histidine, lysine, proline, serine, and tyrosine. However, these are included in the ChemGPS prediction of the 87 amino acids (Figure 11B). A comparison between the local PCA model and the ChemGPS predictions reveals a similar map, except for an inversion in the second principal component (PC2).

No significant 1:1 correspondence between ChemGPS scores and the amino acid *z*-scores (*z*1−*z*5) was observed (data not shown). This situation is similar to the GRID-based heteroaromatic compounds, as different descriptors were applied to characterize the amino acids in the original study.[17] However, we obtained a significant PLS model explaining the first four amino acid *z*-scores (*z*1−*z*4, used as *Y* variables) using predicted ChemGPS scores (1G−9G). The relationship between the *z*-scores and the ChemGPS scores for α-amino acids is shown in Figures 12 and 13. While the latent structures specific for the amino acids appear to be suitably well captured by ChemGPS, these global scores appear to project differently at higher dimensions: Both R^2 and Q^2 exhibit lower values going from *z*1 to *z*3, and from *z*1 to *z*4, respectively, whereas *z*5 could not be modeled (Figure 13).

Prediction of Novel α-Amino Acids. An external set of 20 amino acids was extracted from the ACD database, Figure 14. The cluster composition and relative positioning of these 20 amino acids are also reproduced in the PC2-inverted

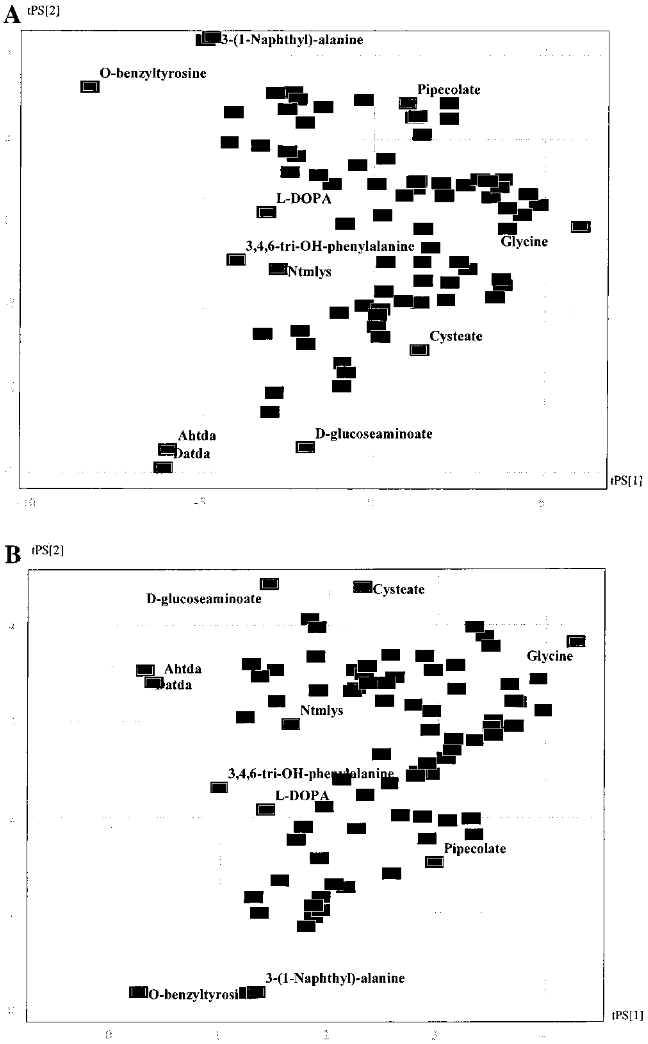

Figure 11. Principal properties for 87 α-amino acids, as obtained from the *t*-scores of a local PCA model (A) and from the ChemGPS prediction (B).

ChemGPS model, compared to the local PCA model predictions, Figure 15. Among these 20 amino acids, none had PModXPS below 5% in the ChemGPS prediction, whereas 17 compounds were deemed as outliers by the same test in the local PCA model. This indicates that ChemGPS is well suited for PCA score prediction for amino acids. A direct comparison between the predictivity of the PCA model proposed by Sandberg et al.[17] and the ChemGPS model was not possible, since the original PCA model contains experimentally measured descriptors from thin-layer chromatography and NMR shifts, in addition to the calculated variables.

General Comments. The present paper describes a procedure that can provide useful and coherent chemical compound property descriptions in a manner that is directly amenable to deriving a global similarity metric. This metric was validated by comparison to local models of (sub)grouped clusters, 3D-based PCA models, and property profiling of α-amino acids. The advantage given by chemographic tools is the long-term perspective, since time-related comparisons for individual compounds via ChemGPS predictions are facilitated, which can greatly simplify the process of compound collection enhancement. It may also provide a unique reference system wherein the characterization of molecules via different descriptors[51] from various research groups may become comparable.

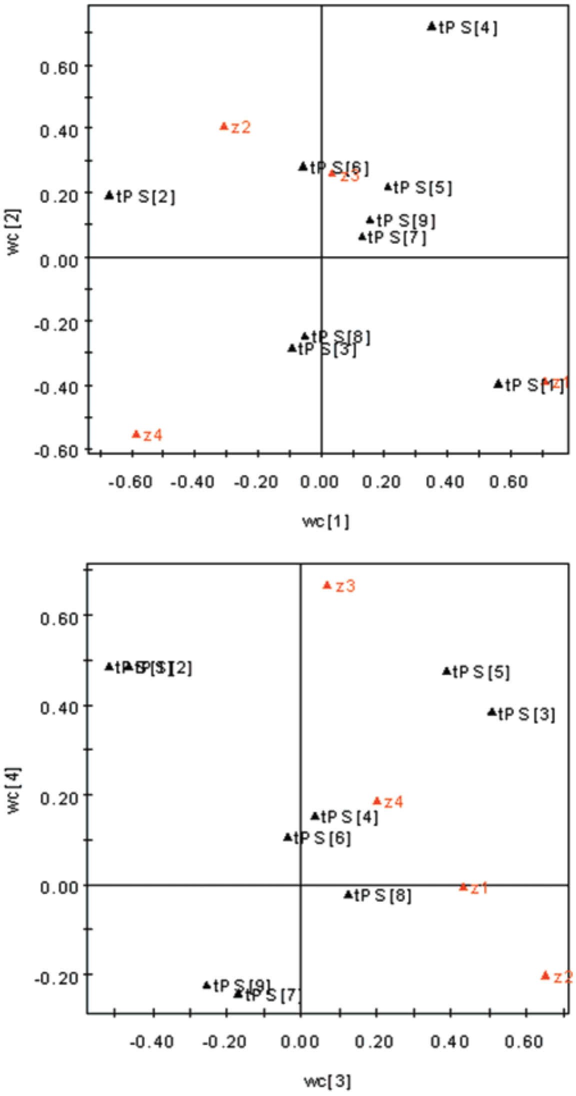

Figure 12. Loading plot of the four-component PLS model derived to explain the first four principal property scores of the amino acid dataset (*z*1−*z*4, marked as red triangles), showing the relationship between *z*1−*z*4 and the nine-dimensional scores from ChemGPS (black triangles).

One common feature with principal component analyses is their occasional rotation. For example, when a minor addition of objects is made, the loadings in two components may shift and the result can be perceived as a rotation in the scores.[12] This was illustrated by the comparison between the 3D-based description from Clementi et al. and our ChemGPS analysis for the same compounds (Figure 9A,B). In addition, the local PCA on the α-amino acid example revealed a shift in sign for the second PC with an apparent inversion along the second PC in both score and loadings, compared to the ChemGPS solution (Figure 11A,B). In the present study, we show how one can avoid such problems by making use of a reference training system (the ChemGPS data set) and extracting the relevant latent structures for different molecules in a systematic manner, be it heteroaromatic compounds (Figures 9 and 10) or amino acids (see Figures 11−13). Thus, a major advantage from the analysis of the

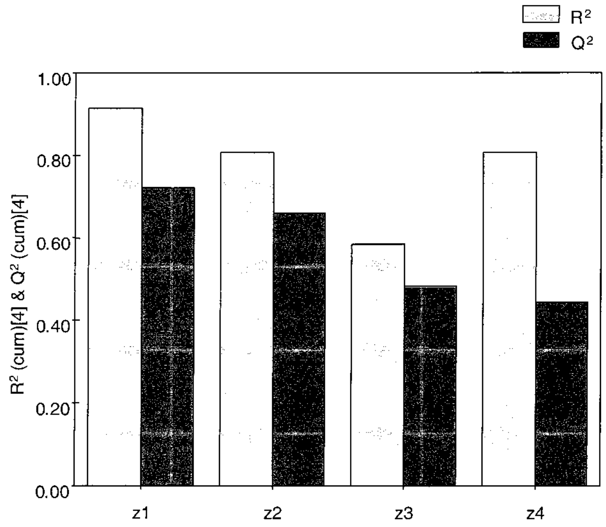

Figure 13. Summary of the PLS model of the $z1-z4$ amino acid scores, using the nine-dimensional (1G−9G) scores from ChemGPS as X variables. For this model, $R^2X = 0.492$, $R^2Y = 0.773$, $Q^2 = 0.576$, and $N = 87$.

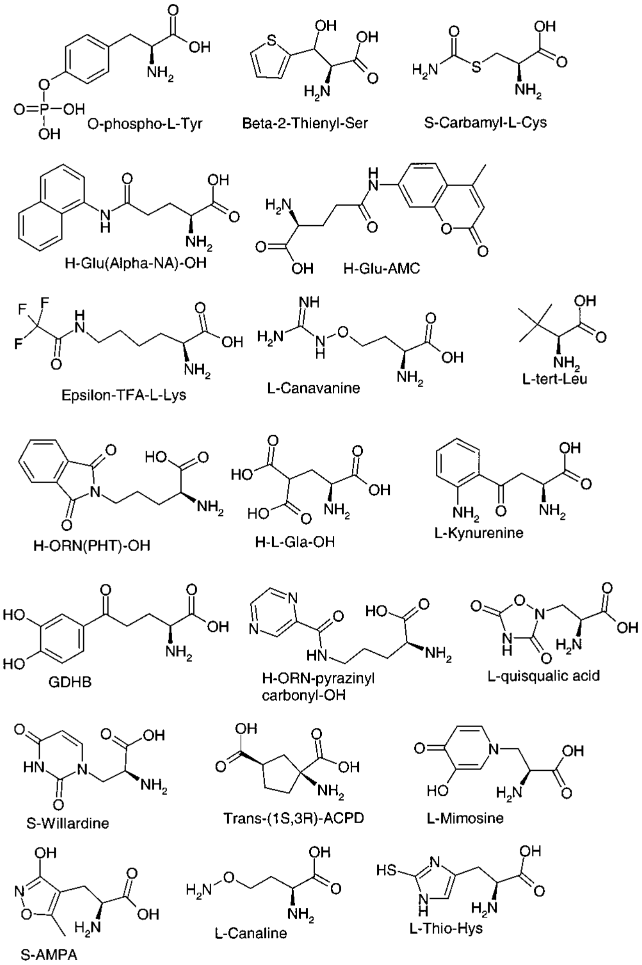

Figure 14. Chemical structures of the 20 α-amino acids used as a test set.

ChemGPS loadings, as provided by a standard ChemGPS molecular property estimation scheme, is the consistent contribution from different properties to a certain PCA dimension. In time, the familiarity with the ChemGPS scheme will provide a direct and robust interpretability for, e.g., compound library analysis and rational drug design.

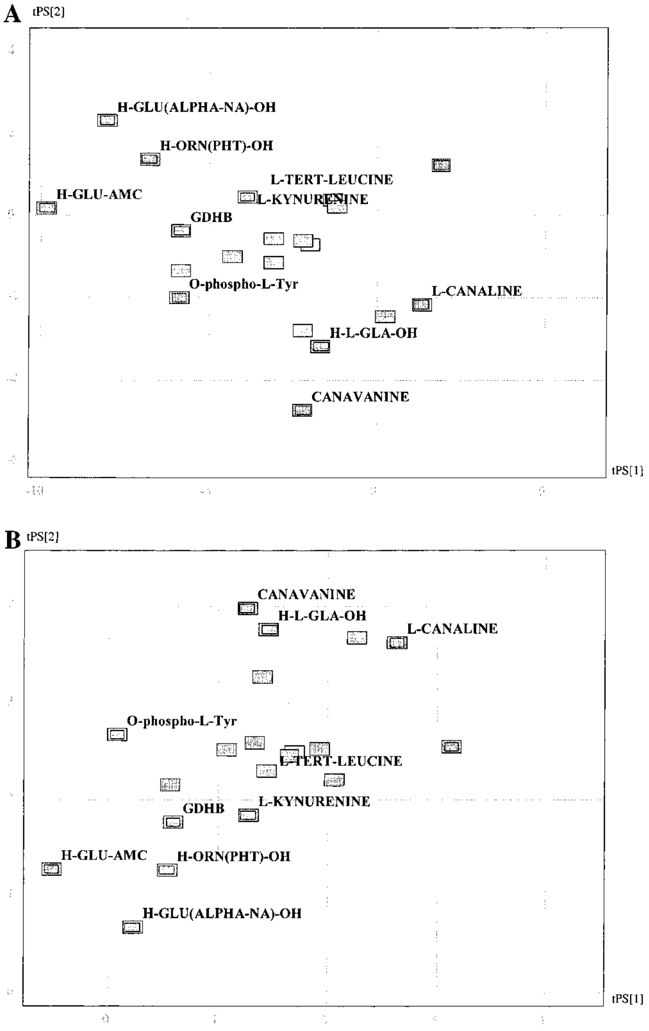

Figure 15. Principal property prediction for the 20 α-amino acids used in the test set, as obtained from the local PCA model (A) and from ChemGPS (B).

Conclusion

Chemography, the art of navigating in chemical space, attempts to address some of the shortcomings of local PCA models, e.g., predictivity via extrapolation and model rederivation upon data set alterations. Chemography provides a standard convention for chemical space navigation: A set of *rules* (principal properties derived by PCA instead of meridians and parallels), and *objects* (molecular structures instead of cities and countries). Similar to the Navstar[23] GPS satellite system, the ChemGPS method makes use of “satellite molecules” intentionally placed *outside* the druglike[22,24,25,32] chemical space. ChemGPS provides a standard tool for compound prediction within the same PCA model, using one training set. ChemGPS preserves cluster characteristics provided by local PCA models (as illustrated for a set of 8599 monocarboxylic acids, 45 heteroaromatics, and 87 α-amino acids), with the advantage of having a significantly reduced number of outliers, when compared to local models (as illustrated for a set of 20 α-amino acids). ChemGPS scores provide a standard metric for chemical space and are directly amenable to comparative analyses across chemistry and time.

Whenever latent structures are stable, they are reflected by projection methods regardless of the descriptor choice (if relevant for the system). When compared to GRID (for heteroaromatics) and to *z*-scores (for α-amino acids), ChemGPS was shown to capture relevant information, that is, amenable for direct comparison in a global manner. In retrospect it seems natural that compound properties comprise statistical self-similarity and perhaps even fractal dimensions.[52] This might be a significant reason for the apparent usefulness of molecular diversity tools in research and development activities that involve chemical structure optimization. Furthermore, ChemGPS is not limited to the choice of descriptors (metric): For example, we have successfully replaced the ChemGPS descriptors described in this paper with VolSurf descriptors[53] in order to obtain consistent maps of the druglike chemical space starting with pharmacokinetically relevant properties.[54] With appropriate training sets and descriptors, ChemGPS is likely to provide standard chemographic metrics in any area (e.g., agrochemicals, polymers, etc.) of chemical discovery.

Acknowledgment. We thank Drs. Thomas Olsson, Vladimir Sherbukhin, Ingemar Nilsson, Magnus Björsne, Ismael Zamora, and Anna-Lena Ungell (all from AstraZeneca R&D Mölndal) for invaluable comments. Dr. Maria Sandberg (Umetrics, Umeå) is gratefully acknowledged for providing an electronic version of the amino acid structures described in ref 17. We thank Dr. Anna Linusson (AstraZeneca R&D Mölndal) for proofreading this manuscript.

References and Notes

(1) Lebl, M. Parallel Personal Comments on “Classical” Papers in Combinatorial Chemistry. *J. Comb. Chem.* **1999**, *1*, 3−24.

(2) Geysen, H. M.; Meloen, R. H.; Barteling, S. J. Use of Peptide Synthesis to Probe Viral Antigens for Epitopes to a Resolution of a Single Amino Acid. *Proc. Natl. Acad. Sci. U.S.A.* **1984**, *81*, 3998−4002.

(3) Geysen, H. M.; Rodda, S. J.; Mason, T. J. A Priori Delineation of a Peptide which Mimics a Discontinuous Antigenic Determinant. *Mol. Immunol.* **1986**, *23*, 709−715.

(4) Kubinyi, H. Chance Favors the Prepared Mind: From Serendipity to Rational Drug Design. *J. Recept. Signal Transduction Res.* **1999**, *19*, 15−39.

(5) Warr, W. A. Combinatorial Chemistry and Molecular Diversity. An Overview. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 134−140.

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

(7) Pearlman, R. S.; Smith, K. M. Software for Chemical Diversity in the Context of Accelerated Drug Discovery. *Drugs Future* **1998**, *23*, 885−895.

(8) Jamois, E. A.; Hassan, M.; Waldman, M. Evaluation of Reagent-Based and Product-Based Strategies in the Design of Combinatorial Library Subsets. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 63−70.

(9) Van Drie, J. H.; Lajiness, S. Approaches to Virtual Library Design. *Drug Discovery Today* **1998**, *3*, 274−283.

(10) Walters, W. P.; Stahl, M. T.; Murcko, M. A. High-throughput “Virtual” Chemistry. In *Encyclopedia of Computational Chemistry*; Schleyer, P. v. R., Ed.; Wiley: New York, 1998; Vol. 2, pp 1225−1237.

(11) Hellberg, S.; Sjöström, M.; Skagerberg, B.; Wold, S. Peptide Quantitative Structure−Activity Relationships, a Multivariate Approach. *J. Med. Chem.* **1987**, *30*, 1126−1135.

(12) Jackson, J. E. *A users guide to principal components*; Wiley: New York, 1991.

(13) Austel, V. A manual method for systematic drug design. *Eur. J. Med. Chem.* **1982**, *17*, 9−16.

(14) Johnson M. E.; Nachtsheim C. J. Some guidelines for constructing exact D-optimal designs on convex design spaces. *Technometrics* **1983**, *25*, 271−277.

(15) Box, G. E. P.; Hunter, W. G.; Hunter, J. S. *Statistics for Experimenters*; Wiley: New York, 1978.

(16) Hellberg, S.; Eriksson, L.; Jonsson, J.; Lindgren, F.; Sjöström, M.; Skagerberg, B.; Wold, S.; Andrews, P. Minimum Analouge Peptide Sets (MAPS) for Quantitative Structure−Activity Relationships. *Int. J. Pept. Protein Res.* **1991**, *37*, 414−424.

(17) Sandberg, M.; Eriksson, L.; Jonsson, J.; Sjöström, M.; Wold, S. New Chemical Descriptors Relevant for the Design of Biologically Active Peptides. A Multivariate Characterization of 87 Amino Acids. *J. Med. Chem.* **1998**, *41*, 2481−2491.

(18) Clementi, S.; Cruciani, G.; Fifi, P.; Riganelli, D.; Valigi, R.; Musumarra, G. A New Set of Principal Properties for Heteroaromatics Obtained by GRID. *Quant. Struct.-Act. Relat.* **1996**, *15*, 108−120.

(19) Pearlman, R. S.; Smith, K. M. Novel Software Tools for Chemical Diversity. *Perspect. Drug Discovery Des.* **1998**, *9−11*, 339−353.

(20) Oprea, T. I.; García, A. E. Three-Dimensional Quantitative Structure Activity Relationships of Steroid Aromatase Inhibitors. *J. Comput.-Aided Mol. Des.* **1996**, *10*, 186−200.

(21) Oprea, T. I.; Waller, C. L.; Marshall, G. R. 3D-QSAR of Human Immunodeficiency Virus (I) Protease Inhibitors. II. Predictive Power Using Limited Exploration of Alternate Binding Modes. *J. Med. Chem.* **1994**, *37*, 2206−2215.

(22) Oprea, T. I.; Gottfries, J.; Sherbuhin, V.; Svensson, P.; Kühler, T. Chemical Information Management in Drug Discovery: Optimizing the Computational and Combinatorial Chemistry Interfaces. *J. Mol. Graphics Model.* **2000**, *18*, 512−524.

(23) More information on the Navstar Global Positioning System information can be retreived from http://gps.laafb.af.mil.

(24) 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.

(25) Sadowski, J.; Kubinyi, H. A Scoring Scheme for Discriminating between Drugs and Nondrugs. *J. Med. Chem.* **1998**, *41*, 3325−3329.

(26) Oprea, T. I.; Gottfries, J. ChemGPS: A Chemical Space Navigation Tool. In *Rational Approaches to Drug Design;* Höltje, H.-D., Sippl W., Eds.; Prous Science Press: Barcelona, 2001, in press.

(27) Oprea, T. I.; Gottfries, J. Toward Minimalistic Modeling of Oral Drug Absorption. *J. Mol. Graphics Modell.* **1999**, *17*, 261−274.

(28) Zamora, I.; Oprea, T. I.; Ungell, A.-L. Prediction of Oral Drug Permeability. In *Rational Approaches to Drug Design*; Höltje, H.-D., Sippl W., Eds.; Prous Science Press: Barcelona, 2001, in press.

(29) Leo, A.; Weininger, A. CMR3 Reference Manual, 1995. Available from Daylight Chemical Information Systems, Santa Fe, New Mexico, http://www.daylight.com/.

(30) Glen, R. C. A Fast Empirical Method for the Calculation of Molecular Polarizability. *J. Comput.-Aided Mol. Des.* **1994**, *8*, 457−466.

(31) Olsson, T.; Shcherbukhin, V. *Synthesis and Structure Administration (SaSA)* 1997−2000; AstraZeneca, http://www.astrazeneca.com.

(32) Oprea, T. I. Property Distribution of Drug-related Chemical Databases. *J. Comput.-Aided Mol. Des.* **2000**, *14*, 251−264.

(33) Basak, S. C.; Balaban, A. T.; Grunwald, G. D.; Gute, B. D. Topological indices: Their nature and mutual relatedness. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 891−898.

(34) Balaban, A. T. Topological and Stereochemical Molecular Descriptors for Databases Useful in QSAR Similarity/Dissimilarity and Drug Design. *SAR QSAR Environ. Res.* **1998**, *8*, 1−21.

(35) Raevsky, O. A.; Grigor'ev, V. Yu.; Kireev, D.; Zefirov, N. S. Complete Thermodynamic Description of H−Bonding in the Framework of Multiplicative Approach. *Quant. Struct.-Act. Relat.* **1992**, *11*, 49−64. HYBOT is available from pION Inc., Cambridge, MA, http://www.pion-inc.com.

(36) Van de Waterbeemd, H.; Camenisch, G.; Folkers, G.; Raevsky, O. A. Estimation of Caco-2 Cell Permeability Using Calculated Molecular Descriptors. *Quant. Struct.-Act. Relat.* **1996**, *15*, 480−490.

(37) Gasteiger, J.; Marsili, M. Iterative Partial Equalization of Orbital Electronegativity: A Rapid Access to Atomic Charges. *Tetrahedron* **1980**, *36*, 3219−3222.

(38) Leo A. Estimating LogP_{oct} from Structures. *Chem. Rev.* **1993**, *5*, 1281−1306.

(39) CLOGP is available from Biobyte Inc., Claremont, CA, http://clogp.pomona.edu/.

(40) ACDLogP is available from ACD Labs, Toronto, Canada, http://www.acdlabs.com/.

(41) SIMCA-P 8.0 is available from Umetrics, Umeå, Sweden, http://www.umetrics.com/.

(42) Wold, S. Cross-validatory Estimation of the Number of Components in Factor and Principal Components Models. *Technometrics* **1978**, *20*, 397−405.

(43) Wold, S.; Ruhe, A.; Wold, H.; Dunn, W. J., III. The collinearity problem in linear regression. The partial least squares approach to generalised inverses. *J. Sci. Stat. Comput.* **1984**, *5*, 735−743.

(44) Höskuldsson, A. PLS regression methods. *J. Chemom.* **1988**, *2*, 211−228.

(45) Oprea, T. I.; Waller, C. L. Theoretical and practical aspects of three-dimensional quantitative structure−activity relationships. In *Reviews in Computational Chemistry*; Lipkowitz, K. B., Boyd, D. B., Eds.; New York, Wiley: 1997, Vol. 11, pp 127−182.

(46) Available from MDL Information Systems, http://www.mdli.com/dats/pharmdb.html.

(47) Linusson, A.; Gottfries, J.; Lindgren, F.; Wold, S. Statistical Molecular Design of Building Blocks for Combinatorial Chemistry. *J. Med. Chem.* **2000**, *43*, 1320−1328.

(48) Gottfries, J.; Oprea, T. I. N-Dimensional Modeling of Objects Within A Hypervolume. Patent Application SE 9804127-0, 1998.

(49) Goodford, P. J. Computational Procedure for Determining Energetically Favourable Binding Sites on Biologically Important Macromolecules. *J. Med. Chem.* **1985**, *28*, 849−857. GRID is available from Molecular Discovery Ltd, http://www99.pair.com/grid/.

(50) Cruciani, G. Personal communication, 2000.

(51) Livingstone, D. J. The Characterization of Chemical Structures Using Molecular Properties. A Survey. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 195−209.

(52) Mandelbrot, B. How Long is the Coast of Britain? Statistical Self-similarity and Fractional Dimensions. *Science* **1967**, *156*, 636−638.

(53) Cruciani G.; Crivori P.; Carrupt P. A.; Testa B. Molecular fields in quantitative structure-permeation relationships: The VolSurf approach. *J. Mol. Struct. (THEOCHEM)* **2000**, *503*, 17−30.

(54) Zamora, I.; Oprea, T. I.; Ungell, A. L. ChemGPS/VolSurf: A pharmacokinetically based mapping device for chemical space. *Eur. J. Pharm. Sci.*, submitted.

CC0000388