Expanding the ChemGPS Chemical Space with Natural Products

Josefin Larsson,[†] Johan Gottfries,[‡] Lars Bohlin,[†] and Anders Backlund*,[†]

Division of Pharmacognosy, Department of Medicinal Chemistry, Biomedical Centre, Uppsala University, PO Box 574, S-751 23 Uppsala, Sweden, and Department of Medicinal Chemistry, AstraZeneca R&D Mölndal, S-431 83 Mölndal, Sweden

Received November 1, 2004

Recently various attempts have been made to increase the efficacy and precision of chemical libraries used in high-throughput screening (HTS) drug discovery approaches. One such approach is ChemGPS, which provides a defined chemical space for prescreening evaluation of chemical compound properties or virtual dereplication. In the present study, ChemGPS has been applied to a set of natural products shown to exhibit cyclooxygenase-1 and/or -2 (COX-1/2) inhibition. With the purpose of defining chemical properties and linking these to the observed mode of enzyme inhibition, this resulted in two lines of reasoning. On one hand several specific features of these compounds have been identified and discussed. Overall COX inhibition was frequently correlated with the presence of at least one ring in the structure, fragments exhibiting structural rigidity, and a relatively large molecular size. The concept "size" includes several parameters, e.g., molecular volume, weight, and number of bonds. On the other hand, and possibly even more important, was the unexpected finding that the natural products studied to a large extent fell outside the defined ChemGPS chemical space. Therefore, we also propose an expanded space for natural products: ChemGPS-NP.

Molecules of natural origin are today used as drugs, precursors for semisynthesis, templates for synthesis, and pharmacological tools. Furthermore, many drug classes on the market have natural products as their origin, as leads or models. So far nature has proven to be by far the richest source of novel compound classes. It has been estimated that there are more than 300 000 different plant species and, on top of these, additional sources such as fungi, bacteria, marine invertebrates, and insects, with another million species or more. The vast majority of this biological diversity is still unexplored, and yet the chemical diversity within nature brings an estimated number of possible drug-like molecules exceeding 10^{60}.[1] Most certainly nature will also in the future be of great importance as a source for discovering both new leads and targets for the development of new drugs.

The concept "chemical space" is often used instead of "multidimensional descriptor space", which is a region defined by the descriptors chosen to describe a set of chemicals.[2] At present, numerous efforts are made at pharmaceutical companies worldwide attempting to identify from their growing reference collections of chemical compounds the nuggets showing effects on a given target under study. It can be easily demonstrated that, even with unlimited resources, with regard to energy and time (not to mention money), only a minute part of chemical space can be explored. Thus, a crucial aspect of all modern drug-screening efforts is to limit chemical space to minimize the number of stops demanded along the route. This can be done in different ways, of which we will demonstrate one, with applications on a set of experimental data based on natural products affecting the cyclooxygenase (COX) enzyme system.

COX is an endogenous key enzyme that mediates the two first steps in the synthesis of prostaglandins (PGs).[3] So far, there are three known isoforms of the enzyme: the well-known COX-1 and COX-2 and the recently reported but poorly known third isoform, COX-3, which has been proposed to be inhibited by paracetamol.[4,5] COX-1 is constitutively expressed and present at a constant level in almost all tissues, where it performs many physiological functions.[6] COX-2, on the other hand, is normally unexpressed but can be induced by inflammatory stimuli.[7] It has been suggested that COX-2 also plays a central role in carcinogenesis.[8,9] Numerous naturally occurring compounds have been investigated for their ability to affect COX-2, either by a direct effect on the enzyme activity or by influencing expression of the genes coding for COXs or translation of their mRNAs.[10]

Oprea and Gottfries (2001)[11] suggested the term "chemography", in resemblance to geography, as "the art of navigating in chemical space". In geography, with conventional mapping systems such as Mercator, it is possible to project, on the same plane, objects located on a geosphere.[12] By applying analogous conventions it is possible to create a drug space map over the chemical space occupied by drug-like molecules. The rules, in geography represented by meridians, would in chemography be dimensions including general properties such as size, lipophilicity, and hydrogen bond capacity. Objects mapped would, when compared to cities, be instead molecular structures, including so-called satellite and core structures. The main objective of chemography is to provide a consistent chemical mapping device, namely, the chemical global positioning system, ChemGPS. This approach makes it possible to avoid extrapolations and instead use interpolations when positioning the properties of new drug-like molecules, much in the same way as geographical positions can be more accurately defined in charted areas. The ChemGPS data set consists of 531 structures selected to provide a balanced chemical space largely defined by Lipinski's rule of five[13] and general drug-like properties, defined here by 60 descriptors.[11] Like the NavstarGPS satellite system, the satellites are intentionally placed outside the space of interest, i.e., in this case the drug-like space.[14] These satellite structures have extreme values in at least one of their properties but still contain drug-like fragments. The core structures are required to retain a focus on drug-like space and to balance

* To whom correspondence should be addressed. Tel: +46-18-4714498. Fax: +46-18-509101. E-mail: Anders.Backlund@fkog.uu.se.
[†] Uppsala University.
[‡] AstraZeneca R&D Mölndal.

10.1021/np049655u CCC: $30.25
Published on Web 06/24/2005

the model. Core structures of the present model have been selected primarily from a list of orally available drugs.[11] Chemographic map coordinates are extracted, in the ChemGPS model, by principal component analysis (PCA) from a fixed list of molecular descriptors that evaluate the above-mentioned rules on a selected set of molecules. The results can be presented as observation- or variable-related projections visualized in two-dimensional plots called score and loading plots, respectively.[15,16]

After establishing the chemographical map, PCA-score prediction is used to project new molecules on this map, providing a consistent and systematic method to explore and map chemical property space. This underlying map does not change with time or the chemistry under evaluation. The principal properties of the compounds are predicted rather than recomputed, and the method consequently has the advantage not only of increasing speed and the possible scope of the study but also of significantly reducing the number of outliers due to the preselected satellite compounds. In the present study, ChemGPS was used to identify and characterize trends among substances of natural origin with an experimentally demonstrated effect on COX-1 and/or COX-2.

Results and Discussion

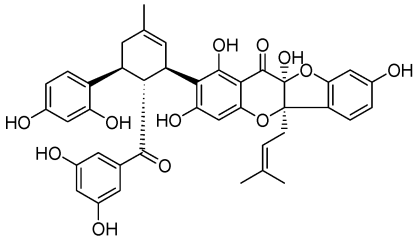
In the literature there are numerous natural products documented as affecting COX. Of the natural products in the data set used in this study around 30% were flavonoids, 20% terpenoids, 10% phenylpropanoids, and 5% alkaloids.

The majority of the COX-2 enzyme inhibitors are located in the lower part of the score plot when plotting the principal components (PCs) PC1 and PC3 (Figure 1), whereas descriptors for rigidity, size, and rings can be found in the corresponding loading plot (Figure 2). This indicates that relatively large, rigid structures with at least one ring are of importance for achieving this activity. The prenylated flavonoid sanggenon D (**1**) from *Morus mongolica* and the polyketide derivative radicicol (**2**) from *Penicillium luteoaurantium* are examples of substances from this group.

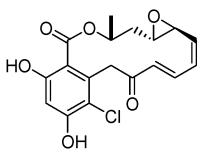
A large proportion of the substances inhibiting COX-2 mRNA activity are sited in the fourth quadrant when looking at PC1/PC3 (Figure 1). The corresponding quadrant in the loading plot contains descriptors for rigidity, ring, and size (Figure 2). A majority of the compounds are rigid and contain at least one ring, a concept that seems to be favorable for inhibition at transcriptional level. The diterpenoid triptolide (**3**), from *Tripterigum wilfordii*, is one of the representatives of this group.

When observing PC1/PC3 of multidimensional space occupied by the substances inhibiting COX-2 protein expression, several occur drawn toward the lower right-hand side of the plot (Figure 1). On this side of the corresponding loading plot descriptors for size, ring, and rigidity are situated (Figure 2). None of the molecules appear to be extremely small, with all consisting of 11 or more carbons and having a molecular weight of 270 or more. The biflavonoid ginkgetin (**4**), from *Ginkgo biloba*, is an example of a compound from this group.

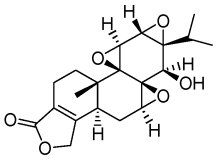
Almost all COX-1 enzyme inhibitors are situated in the upper half of the score plot when observing PC1 and PC5. The descriptors indicate that the distance between a donor and an acceptor or between two acceptors is frequently large. They also signify that there are few negatively charged substances and that nitrogen is not preferred in the structures, and strong H-bond donors are uncommon for inhibition of the COX-1 enzyme. The majority of the COX-1 inhibitors are also positioned in parts of the plot corresponding to low flexibility and a small number of rings. A representative from this group is the monoterpenoid carvacrol (**5**) from *Alpina officinarum*.



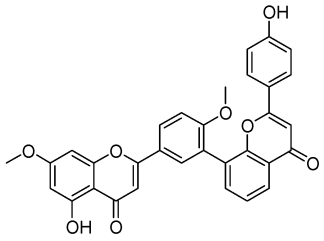
1



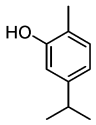
2



3



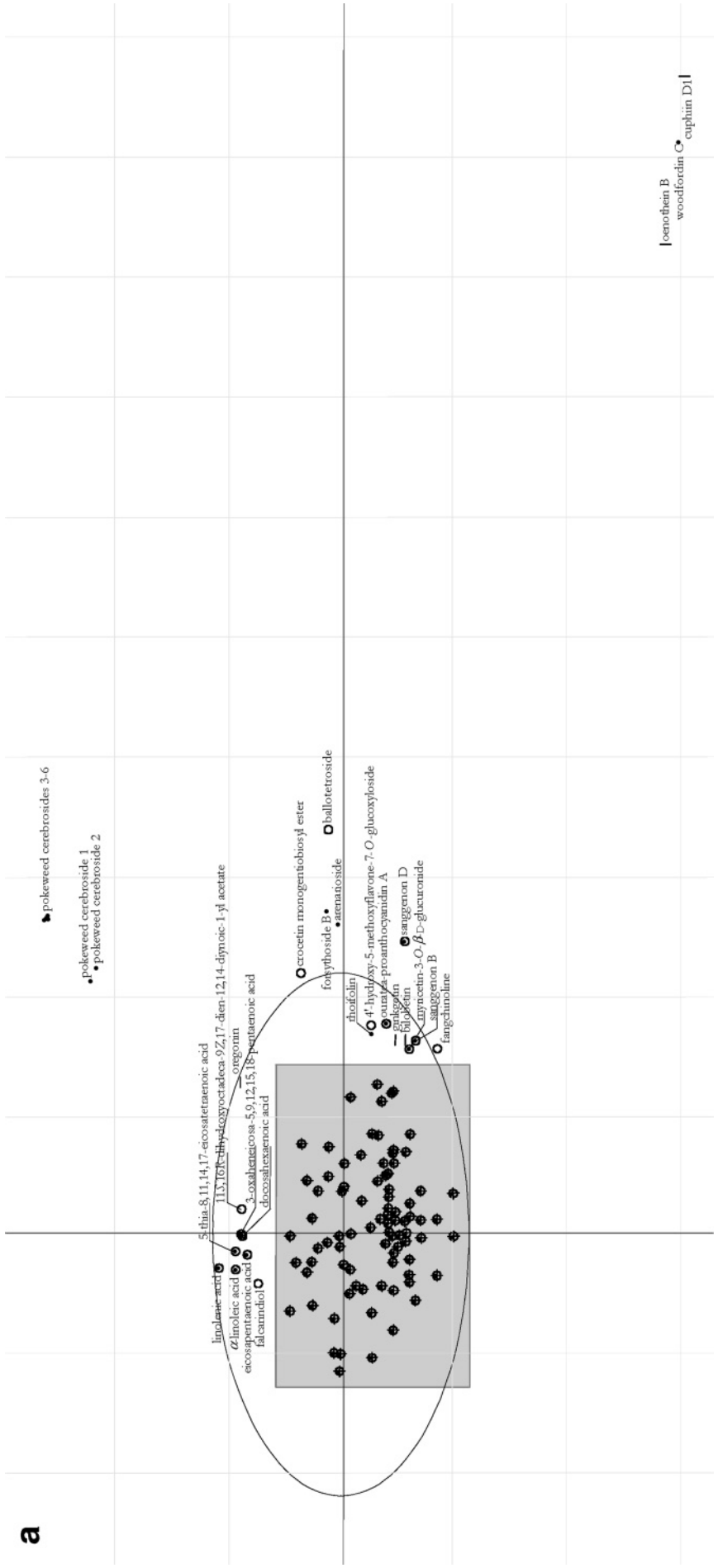
4



5

In many of the plots a cluster consisting of fatty acids can be discerned. These differ from the majority of the compounds included by having no ring systems and by being more flexible. It is well known from studies that some fatty acids show a strong effect on COX activity.[17] The notable difference compared to the chemical properties representative for most other compounds included in the study support the hypothesis[10] that there are different mechanisms of selective COX-2 enzyme inhibition, with flexibility favorable in at least one of these. In other mechanisms, however, it appears to be more functional with flat, rigid structures. This difference appears relevant in light of the structural studies on differences between COX-1 and COX-2.[3,18] The COX active site consists of a long, narrow channel, reaching from the outside of the membrane-binding domain deep into the interior of the enzyme, and has three different regions: a hydrophobic pocket under the heme group, the orifice of the active site, and a side pocket that is larger in COX-2 than in COX-1.[19] Access to the extra volume of the side pocket in COX-2 makes COX-2 drug selectivity possible. This extra volume is formed by an amino acid exchange of valine at position 523 in COX-2 for isoleucine and at the same position in COX-1. Isoleucine is, because of a longer side chain, more bulky and restricts the access to the side pocket in COX-1.[20] The binding site in COX-1 is 316 Å^3, and the volume of the primary inhibitor binding site and the secondary pocket in COX-2 is 25% larger, 394 Å^3. The secondary pocket contributes also to the larger volume, but the central channel is wider in COX-2.[3]

It may be concluded that, overall, COX inhibition was frequently correlated with the presence of at least one ring


a
pokeweed cerebrosides 3-6
pokeweed cerebroside 1
pokeweed cerebroside 2
oenothein B
woodfordin C
cuphiin D1
crocetin monogentiobiosyl ester
ballotetroside
forsythoside B
arenariioside
5-thia-8,11,14,17-eicosatetraenoic acid
11S,16R-dihydroxyoctadeca-9Z,17-dien-12,14-diynoic-1-yl acetate
oregonin
3-oxaheneicosa-5,9,12,15,18-pentaenoic acid
docosahexaenoic acid
linolenic acid
α-linoleic acid
eicosapentaenoic acid
falcarindiol
dhonfolin
4'-hydroxy-5-methoxyflavone-7-O-glucoxyloside
ounalia proanthocyanidin A
ginkgetin
bilobetin
myricetin-3-O-β-D-glucuronide
sanggenon D
sanggenon B
fangchinoline

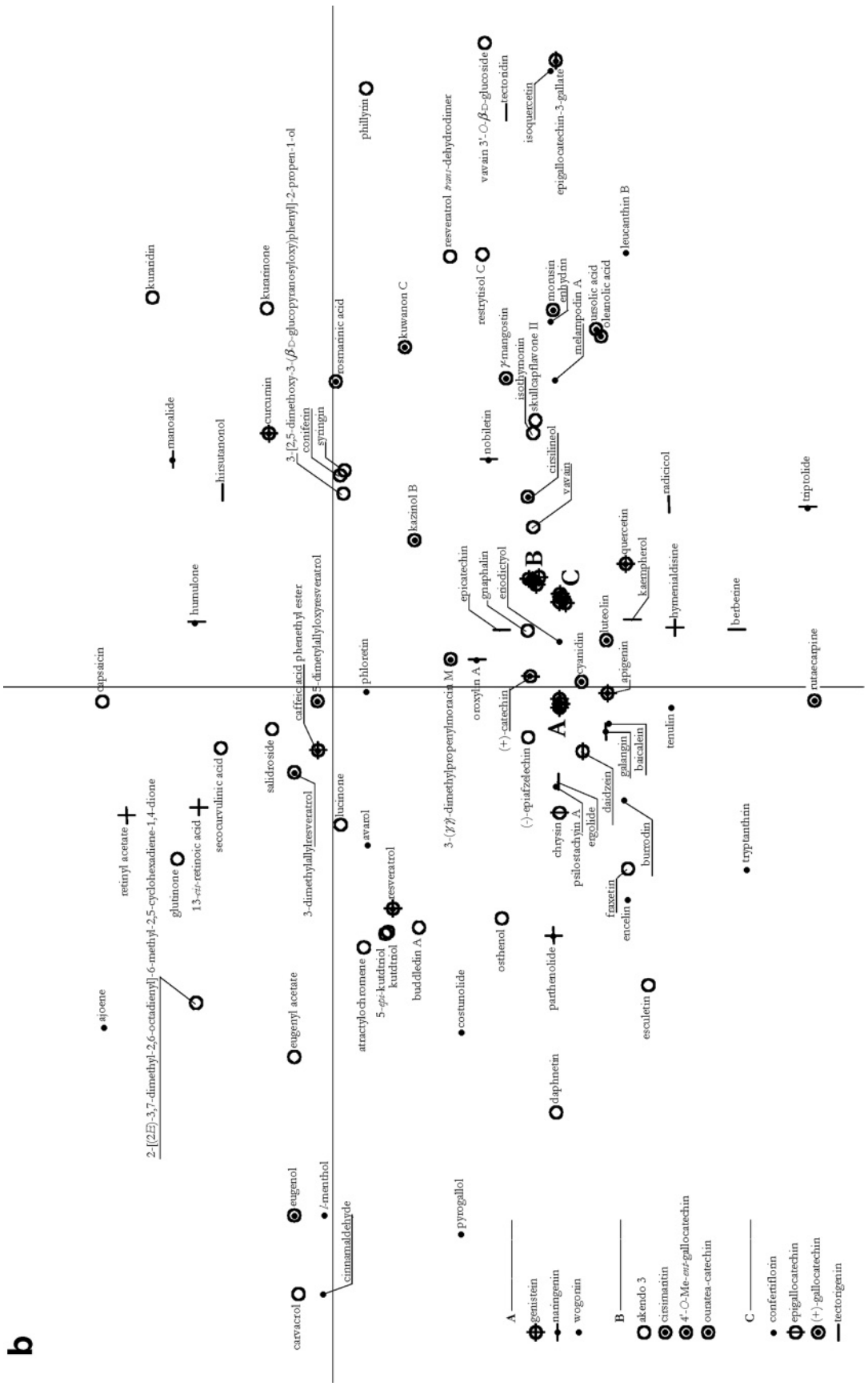


Figure 1. Predicted score plots of the first and third PCs of the COX inhibitor data set. The first component comprises 32% of the variation in the molecular properties, and the second component accounts for 9%. Each combined symbol represents one compound. ○ represents inhibition on COX-1 enzyme level, ● on COX-2 mRNA level, and — on COX-2 protein level. The circle in the plots indicates the critical distance to the model. Observations outside the critical distance are considered outliers. (a) Entire dataset. (b) Enlargement of the shaded area in part a. A, B, and C in part b represent three groups of tightly clustered substances. Substance names and biological activities for these clusters are indicated in the lower left corner.

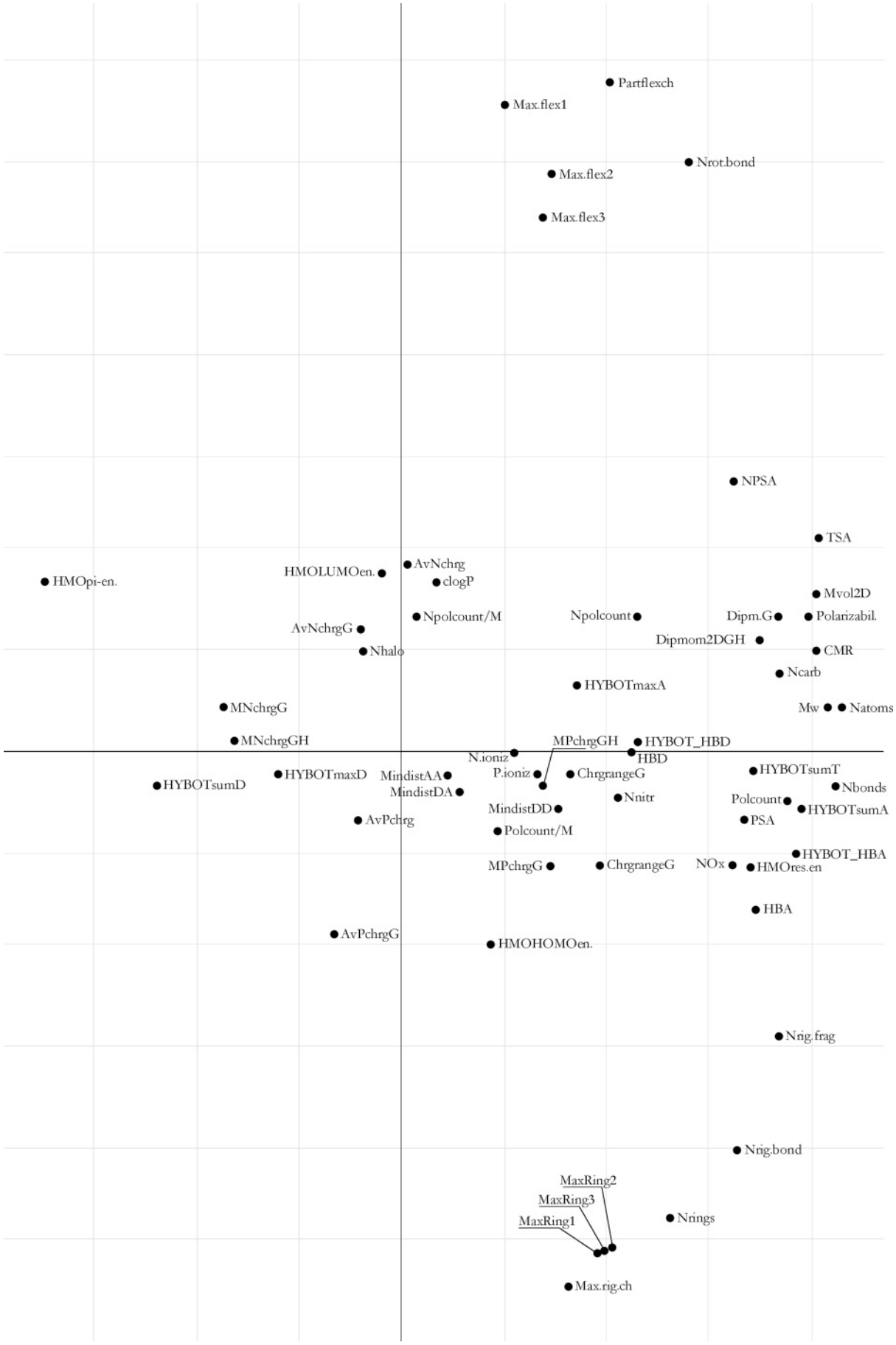


Figure 2. Loading plot for the first and third PCs of the ChemGPS training set.

in the structure, fragments exhibiting structural rigidity, and a relatively large molecular size. The concept "size" includes several parameters, e.g., molecular volume, weight, and number of bonds. This correlates with the fact that flavonoids are frequently identified as active compounds in different COX assays. There are over 6000 naturally occurring flavonoids found in a wide variety of different plants, of which a large number are unexplored in terms of their COX-inhibitory activities.[21]

The results of this ChemGPS prediction should be valuable in the future search for natural and novel COX inhibitors. The prediction of all of the compounds included in this study also reveals approximately 10 moderate outliers that may be difficult to accurately characterize since they represent apparent extrapolation outside the ChemGPS chemical space. Since these are predictions from ChemGPS, the outliers do not cause distortion of the model, and they will not change the characterization of the structures covered by the model. The characteristics of the outliers in the model are large, complex structures with many rings and high polarizability (Figure S1).

Due to the rapid technical development and instrumentalization of chemical sciences, many variables can today be measured simultaneously and the masses of data obtained are often organized in multivariate tables. It can be very difficult to visualize relationships between the tabulated variables by just observing a table. There is an urgent need for computer-based methods to extract meaningful information from raw data within a short time. For this purpose multivariate data analyses of different kinds have proven to be valuable tools.

Various methods have been developed to foster the integration of experimental and *in silico* screening and maximize their output in drug discovery. HTS has become a key technology of pharmaceutical drug discovery research. Improving the quality of screening libraries and HTS assays, rather than their quantity, is a factor of increasing importance for the identification of active compounds. The odds that the "hits" selected will progress through the development stages toward a new drug need to be improved. By increasing the quality of the screening libraries fewer and smarter experiments need to be performed, and thereby time, resources, and substantial amounts of money can be saved. Diverse methods have been introduced over the past few years to analyze screening data, extract knowledge from experiments, and derive predictive models of activity. Integrating these methods with for instance HTS can help transform random screenings into more focused efforts.[22]

The PCA-based model, ChemGPS, is a tool for such focused experiments. It is evident already from this initial study that ChemGPS can be used to cull complex characteristics of compounds affecting the biological activity investigated. It is also possible to rapidly predict and position large numbers of new molecular structures in chemical space and, from this, to draw conclusions concerning their odds of influencing the system studied, a feature that can be used as a means of virtual dereplication. Completely different or additional descriptors and other programs for calculating descriptor sets can be chosen according to research focus, consequently providing virtually unlimited possibilities to investigate new correlations. Due to its scalability, ChemGPS is also well suited to become an easily expandable reference system for comparing multiple combinatorial libraries and for keeping track of the explored regions of chemical space.

The reason we encounter outliers when natural products are being analyzed, despite the design of the ChemGPS, is that some of these natural products are very different in terms of structure and chemical properties from the drug-like molecules for which the system was initially designed. Predicted outliers result in extrapolation and thus uncertainty in accuracy and precision. On the other hand, the outliers are an important observation. Although being a useful model for characterization of molecules, the present ChemGPS is not an ideal framework to use in the search for novel molecules of natural origin and does not suit some of the complex and atypical chemical structures encountered. We therefore propose the expansion of ChemGPS into ChemGPS-NP, where NP stands for natural products. This will become a future tool, tuned for natural products research and better at avoiding outliers when predicting the biological activity of natural products. Work with a complementary ChemGPS-NP with a modified inner structure more focused on the diverse chemical structures of nature is at present in progress.

Experimental Section

Literature Survey. Data were compiled from PubMed and Biological Abstracts and from in-house studies. Information, including name and source of substance, system or assay in which the substances have been tested for activity, type of effect on COX, i.e., inhibition or stimulation on gene, translational, or enzymatic level, definition of observed effect (most frequently IC_{50} value), molecular structure, and assay reference substance, was organized in a database.

Computations. The molecular structures of the active substances were transformed to structure description files (SDF files), from which an AstraZeneca in-house program calculated the ChemGPS set of 60 descriptors (Table S1). The principal components of the training set, i.e., ChemGPS, were calculated using the software SIMCA-P+, version 10.5.[23] All data were mean centered and scaled to unit variance. The COX inhibitors (listed in Table S2) were then positioned in chemical space on the basis of predicted PCA scores using ChemGPS as the model.

Acknowledgment. The authors are grateful to E. Johansson at Umetrics AB, Umeå, Sweden, for help with literature and software and to Professor T. Oprea, University of New Mexico School of Medicine, Albuquerque, NM (previously at Astra Zeneca R&D, Mölndal, Sweden), for valuable discussions. This work was supported, in part, by the Swedish Research for Environment, Agricultural Science and Spatial Planning.

Supporting Information Available: Tables listing the molecular descriptors with explanations (Table S1), information on the substances plotted in figures (Table S2), and the chemical structure of a ChemGPS outlier (Figure S1) are available free of charge via the Internet at http://pubs.acs.org.

References and Notes

(1) Bohacek, R. S.; McMartin, C.; Guida, W. C. *Med. Res. Rev.* **1996**, *16*, 3−50.
(2) Dobson, C. M. *Nature* **2004**, *432*, 824−828.
(3) Luong, C.; Miller, A.; Barnett, J.; Chow, J.; Ramesha, C.; Browner, M. F. *Nature Struct. Biol.* **1996**, *3*, 927−933.
(4) Willoughby, D. A.; Moore, A. R.; Colville-Nash, P. R. *Lancet* **2000**, *355*, 646−648.
(5) Botting, R.; Ayoub, S. S. *Prostaglandins Leukot. Essent. Fatty Acids* **2005**, *72*, 85−87.
(6) Otto, J. C.; Smith, W. L. *J. Lipid Mediat. Cell Signal.* **1995**, *12*, 139−156.
(7) Ristimäki, A.; Garfinkel, S.; Wessendorf, J.; Maciag, T.; Hla, T. *J. Biol. Chem.* **1994**, *269*, 11769−11775.
(8) Hinz, B.; Brune, K. *J. Pharmacol. Exp. Ther.* **2002**, *300*, 367−375.
(9) Huss, U. Studies on the Effects of Plant and Food Constituents on Cyclooxygenase-2: Aspects in Inflammation and Cancer. Ph.D. Thesis, Uppsala University, Uppsala, Sweden, 2003, p 12.

(10) Ringbom, T. Bioassay Development for Identification of Cyclooxygenase-2 Inhibitors of Natural Origin. Ph.D. Thesis, Uppsala University, Uppsala, Sweden, 2002, p 11.
(11) Oprea, T. I.; Gottfries, J. *J. Comb. Chem.* **2001**, *3*, 157−166.
(12) Snyder, J. P. *Map Projections a Working Manual*, USGS Professional Paper; U.S. Government Printing Office: Washington, DC, 1935.
(13) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeny, P. J. *Adv. Drug Delivery Rev.* **1997**, *23*, 3−25.
(14) Pike, J. Navstar GPS available at: http://www.fas.org/spp/military/program/nav/gps.htm.
(15) Pearson, K. *Philos. Mag.* **1901**, *2*, 559−572.
(16) Eriksson, L.; Johansson, E.; Kettaneh-Wold, N.; Wold, S. *Multi- and Megavariate Data Analysis Principles and Applications*; Umetrics AB: Umeå, Sweden, 2001; pp 43−70.
(17) Ringbom, T.; Huss, U.; Stenholm, Å.; Flock, S.; Skattebøl, L.; Perera, P.; Bohlin, L. *J. Nat. Prod.* **2001**, *64*, 745−749.
(18) Picot, D.; Loll, P. J.; Garavito, R. M. *Nature* **1994**, *367*, 243−249.
(19) Llorens, O.; Perez, J. J.; Palomer, A.; Mauleón, D. *J. Mol. Graph. Mod.* **2002**, *20*, 359−371.
(20) Dannhardt, G.; Kiefer, W. *Eur. J. Med. Chem.* **2001**, *36*, 109−126.
(21) Harborne, J. B.; Williams, C. A. *Phytochemistry* **2000**, *55*, 481−504.
(22) Bajorath, J. *Nat. Rev. Drug Discovery* **2002**, *1*, 882−894.
(23) SIMCA-P+ 10.5; Umetrics: Box 7960, S-90719 Umeå, Sweden. Available from www.umetrics.com.

NP049655U