ChemGPS-NP and the Exploration of Biologically Relevant Chemical Space

JOSEFIN ROSÉN
Dissertation presented at Uppsala University to be publicly examined in B42, BMC, Husargatan 3, Uppsala, Friday, March 27, 2009 at 09:15 for the degree of Doctor of Philosophy (Faculty of Pharmacy). The examination will be conducted in English.

Abstract

Chemical space is basically infinite, and comprises all molecules that could possibly exist. Intelligent ways to efficiently navigate through chemical space and to select promising compounds in drug discovery are important tasks, and the focus of this thesis.

In this work a new model for chemical space navigation, ChemGPS-NP, was developed. This model is based on a methodology where a global chemical space map is defined through principal component analysis of physico-chemical properties of a reference set of compounds. Through interpolation from the reference set, positions of novel compounds can be defined on this map and interpreted as chemical properties.

ChemGPS-NP was demonstrated to be able to chart the entire biologically relevant chemical space, including both drug-like and natural compounds. This is an important improvement considering the present interest in natural products (NPs) in the pharmaceutical industry, as well as the track record of NPs to serve as basis for more than 50% of all marketed drugs. ChemGPS-NP proved able to handle and process large data sets, to aid in efficient selection of test objects, and to extract useful information from the results of high-throughput screening campaigns. Using ChemGPS-NP, it was shown that NPs occupy unique regions of chemical property space in comparison to drug-like compounds, and a number of features distinguishing NPs from medicinal chemistry compounds were identified.

ChemGPS-NP was also shown to be able to reliably predict mode of action of anticancer agents based on chemical structure, a finding that has potential to improve cancer research efficiency. Applying a property based similarity search based on calculated eight dimensional Euclidean distances from ChemGPS-NP rendered a tool to identify NP inspired potential leads for drug discovery.

Furthermore, ChemGPS-NPweb, an online version of ChemGPS-NP, was developed, which provides scientists with open access to the tool via http://chemgps.bmc.uu.se/.

Keywords: ChemGPS-NP, chemical space, chemography, selection

Josefin Rosén, Department of Medicinal Chemistry, Division of Pharmacognosy, Box 574, Uppsala University, SE-75123 Uppsala, Sweden

© Josefin Rosén 2009

ISSN 1651-6192
ISBN 978-91-554-7431-7
urn:nbn:se:uu:diva-89364 (http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-89364)
To my family
This doctoral thesis is based on the following papers, which will be referred to in the text by their Roman numerals (I-V):


* née Larsson

Reprints were made with permission from the publishers.
## Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Introduction</td>
<td>11</td>
</tr>
<tr>
<td>Pharmacognosy</td>
<td>11</td>
</tr>
<tr>
<td>Natural products in drug discovery</td>
<td>12</td>
</tr>
<tr>
<td>The unique properties of natural products</td>
<td>13</td>
</tr>
<tr>
<td>Chemometrics</td>
<td>14</td>
</tr>
<tr>
<td>Chemical structure representation</td>
<td>14</td>
</tr>
<tr>
<td>Calculation of molecular descriptors</td>
<td>15</td>
</tr>
<tr>
<td>Pfizer’s Rule of Five</td>
<td>16</td>
</tr>
<tr>
<td>Similarity</td>
<td>16</td>
</tr>
<tr>
<td>Structure-activity relationships</td>
<td>17</td>
</tr>
<tr>
<td>The concept of chemical space</td>
<td>17</td>
</tr>
<tr>
<td>Chemography and the ChemGPS methodology</td>
<td>18</td>
</tr>
<tr>
<td>Aims of the thesis</td>
<td>20</td>
</tr>
<tr>
<td>Materials and Methods</td>
<td>21</td>
</tr>
<tr>
<td>Chemical structure representation</td>
<td>21</td>
</tr>
<tr>
<td>Structure data files</td>
<td>21</td>
</tr>
<tr>
<td>SMILES</td>
<td>21</td>
</tr>
<tr>
<td>Daylight fingerprints</td>
<td>22</td>
</tr>
<tr>
<td>Similarity metrics</td>
<td>23</td>
</tr>
<tr>
<td>Tanimoto index</td>
<td>23</td>
</tr>
<tr>
<td>Euclidean distance</td>
<td>23</td>
</tr>
<tr>
<td>Molecular descriptors</td>
<td>24</td>
</tr>
<tr>
<td>Multivariate data analysis</td>
<td>26</td>
</tr>
<tr>
<td>Pre-processing multivariate data</td>
<td>26</td>
</tr>
<tr>
<td>PCA</td>
<td>27</td>
</tr>
<tr>
<td>Cross-validation</td>
<td>28</td>
</tr>
<tr>
<td>OPLS-DA</td>
<td>29</td>
</tr>
<tr>
<td>SIMCA</td>
<td>30</td>
</tr>
<tr>
<td>Statistical molecular design</td>
<td>30</td>
</tr>
<tr>
<td>D-optimal design</td>
<td>31</td>
</tr>
<tr>
<td>D-optimal onion design</td>
<td>31</td>
</tr>
<tr>
<td>Data sets</td>
<td>32</td>
</tr>
<tr>
<td>Cytotoxicity estimation</td>
<td>33</td>
</tr>
<tr>
<td>Activity profiles</td>
<td>33</td>
</tr>
</tbody>
</table>
DModX ....................................................................................................34
Determination of lead-likeness.................................................................34

Results and discussion ..................................................................................35
The need for a new ChemGPS model (Paper I) ........................................35
The development of ChemGPS-NP (Paper II) .........................................36
  Selection of descriptor array ................................................................37
  Comparison between ChemGPS and ChemGPS-NP ............................38
  Validation with other descriptor set .....................................................39
  Potential applications of ChemGPS-NP .............................................39
ChemGPS-NP online (Paper III)..............................................................40
Predicting anticancer mode of action (Paper IV) .................................45
Exploration of biologically relevant chemical space (Paper V) ..........49

Conclusions and future perspectives .............................................................53

Populärvetenskaplig sammanfattning ...........................................................56

Acknowledgements .......................................................................................60
References .....................................................................................................62
Abbreviations

0D, 1D, 2D, 3D Zero, one, two and three Dimensional
ADME Absorption, Distribution, Metabolism, and Excretion
AID Assay ID (in PubChem)
ChemGPS Chemical Global Positioning System, Chemical Global Property Space, or Chemical Global Property Scores
COX Cyclo-OXygenase
DA Discriminant Analysis
DModX Distance to Model in X space
DNP Dictionary of Natural Products
DoE Design of Experiments
DOOD D (determinant)-Optimal Onion Design
ED Euclidean Distance
HTS High-Throughput Screening
IC50 Inhibitory Concentration 50%
MIF Molecular Interaction Field
MDL Molecular Design Limited
MOA Mode Of Action
NCI National Cancer Institute
NCI60 Cancer cell line panel at the NCI
NP Natural Product
OPLS Orthogonal Partial Least Squares
PC Principal Component
PCA Principal Component Analysis
PLS Partial Least Squares
PRESS Prediction Error Sum of Squares
Q SAR Quantitative Structure-Activity Relationships
QSPR Quantitative Structure-Property Relationships
Ro5 Rule of Five
RSS Residual Sum of Squares
RT Reverse Transcriptase
SAR Structure-Activity Relationships
SMD Statistical Molecular Design
SPR Structure-Property Relationships
SDF Structure Data File
SLN SYBYL Line Notation
SMILES Simplified Molecular Input Line Entry Specification
SIMCA Soft Independent Modelling of Class Analogy
UAS10 Cancer cell line panel at Uppsala University hospital
WOMBAT World of Molecular BioAcTivity
Introduction

In modern drug discovery vast amounts of data are produced, driven by e.g. the omics technologies, such as genomics and proteomics, and high-throughput screenings (HTS). Numerous variables can today be measured and collected simultaneously, and the pharmaceutical industry tends to face a data-overload [1]. This could potentially slow down the productivity of drug discovery companies if its management and interpretation is not efficiently dealt with. Improving the quality of e.g. screening libraries, rather than their quantity, is a factor of increasing importance for the identification of active compounds. Meaningful information needs to be efficiently extracted from the raw data to enable rational prioritization and selection of which compound to synthesize or test next out of thousands or millions of possibilities. Initial high-throughput evaluation of compounds in silico has considerable impact on the success rate of screening programs and has the ability of transforming random screenings into more focused efforts. There is a need for rational methods and tools to handle this information and to subsequently move from data management to knowledge management. The work presented in this thesis is intended to aid in that process.

Pharmacognosy

The present work has been conducted at the Department of Medicinal Chemistry, Division of Pharmacognosy, Uppsala University. The word pharmacognosy originates from two Greek words, pharmakon (drug) and gnosis (knowledge), and can thereby be translated as ‘the knowledge of drugs’. A more modern definition of pharmacognosy is ‘a molecular science that explores naturally occurring structure-activity relationships with a drug potential’ [2]. Studies in medicinal chemistry, and more specifically regarding the chemical properties and functions of natural products (NPs), have a long tradition at Uppsala university since at least the Linnaean disciple Laurentius Hiortzberg defended his thesis in 1754 [3].

Pharmacognosy has evolved as an interdisciplinary field of science that spans a wide range of fields, and a model, see Figure 1, was recently proposed to illustrate this [4].
Natural products in drug discovery

NPs, in particular secondary metabolites, are essential sources for the discovery and development of new promising compounds in drug discovery. Secondary metabolites are, in contrast to primary metabolites, compounds that are not directly involved in the normal growth, development or reproduction of organisms. They are instead e.g. used as protection against potential threats like predators, parasites, and diseases, for competition with other species, and as attractants in reproductive processes (e.g. colouring agents and attractive scents).

NPs can be regarded as pre-validated by Nature. They have been optimized by evolutionary forces in a natural selection process for optimal interactions with biological macromolecules. NPs also surprisingly often have advantageous pharmacokinetic properties as a result following their mission of identifying targets in the organism [5].

NPs also have a long history in medicine and were for long the only medication available. Approximately 50% of the drugs in clinical use today are of natural origin [6-8]. Indeed, many of the currently marketed drugs are NPs or of NP origin, particularly in the therapeutic fields of cancer and infectious disease [7-11].

Despite the potentials and beneficial properties of NPs, there was a decline in the development of natural compounds as potential drug candidates
in the pharmaceutical industry in the 1990s in favour of high-throughput synthesis of large compound libraries [12]. New technologies, including e.g. combinatorial chemistry, HTS, and the omics techniques, improved the efficiency of the drug discovery process and gave NP research competitive disadvantages [13]. This since investigation of NPs involved time-consuming extract-library screening, bioassay-guided isolation, followed by laborious structure elucidation. Even though these new technologies have accelerated the synthesis and delivered hits, they unfortunately have not fulfilled the expectations of an increasing number of new drug candidates. In the drug discovery we now instead find a renewed interest in NPs as a promising and dependable source of novel bioactive compounds [12,14-16]. Several start-ups focusing entirely on NP-based drug discovery have emerged [17], and traditional pharmaceutical companies are increasing their investments in the NP based drug discovery.

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. Natural compounds also display an exceptional chemical diversity that can inspire synthetic chemists when developing combinatorial libraries [18,19]. One such example is diversity oriented synthesis, which intends to increase the chemical diversity within a combinatorial library so that the compound collection differs as much as possible in molecular properties and chemical structures [13,20].

The unique properties of natural products

NPs occupy a different and larger space than that normally dealt with in medicinal chemistry, and several statistical analyses have identified a number of distinguishing properties of NPs.

The distribution of molecular weight seems to be comparable for synthetic drugs and NPs [21].

Natural compounds are in general more unsaturated than synthetic drugs, but even though they, in average, contain about two more rings, NPs contain considerably fewer aromatic rings [22,23]. Approximately every second atom in drugs is aromatic, but on average only every sixth atom in NPs [24].

NPs typically have circa five times higher average number of chiral centers than synthetic drugs [22,24]. Furthermore they often contain more O atoms, alcohols are e.g. three times more frequent in NPs compared to synthetic drugs, but less N atoms, and much less S and halogen atoms [21-24].

NPs are more rigid than synthetic drugs. Their number of rotatable bonds is in average two less than in drugs [22,23].
NPs also differ by having a higher number of hydrogen bond donors and acceptors per molecule.

Additionally the topological polar surface area, regarded as critical for drug bioavailability [25-27], is greater for NPs than for synthetic drugs, even though they are both within the suggested range for oral bioavailability [27].

Chemometrics

Chemometrics is the application of mathematical and statistical methods to chemical data. Chemometric methods include multivariate projection methods, such as principal component analysis (PCA) [28,29], partial least squares (PLS) [30], and orthogonal partial least squares (OPLS) [31,32], as well as statistical experimental design, also known as design of experiments (DoE) [33]. Projection methods are important tools for identifying trends, clusters and outliers in data sets, while DoE is the methodology of how to conduct and plan experiments in order to extract the maximum amount of information in the fewest number of experiments. DoE can also be used to select representative, non-redundant sample sets for further characterization [34,35]. Even though chemometric tools, as the name suggests, initially were used mainly for chemical problems, they are today widely used in many fields of science, e.g. in various biological [36,37] and pharmaceutical [38] fields.

Chemical structure representation

To be able to use chemical structures in computer programs and databases, they need to be represented in a computational form. Simply using the compound name, its two dimensional (2D) drawing or the three dimensional (3D) molecular model as input data is impractical. There are several ways of representing chemical structures in suitable form, e.g. linear notations, connection tables or 3D structure representations.

Linear notations represent the structure of a chemical compound as a linear sequence of characters and numbers, where common examples are SMILES (Simplified Molecular Input Line Entry Specification) [39,40] and SLN (SYBYL Line Notation) [41].

A connection table contains information regarding the structural relationships and properties of a collection of atoms. Examples of exchange format based on a connection table is the Molfile format and the Structure Data File (SDF), developed by Molecular Design Limited (MDL) Information Systems [42].
A 3D structure representation can be obtained from the connection table of a compound. The most well-known 3D structure generators are CORINA [43,44] and CONCORD [45].

Fingerprints, e.g. the Daylight fingerprints [46], can also be seen as a way of represent the chemical structure. Fingerprints are binary bit string representations that indicate specific substructure fragments in a compound with a binary bit (1 as present, 0 as absent) at specific positions. Their design concepts and lengths can vary largely.

**Calculation of molecular descriptors**

A molecular descriptor can be defined as ‘the final result of a logic and mathematical procedure transforming the chemical information encoded within a chemical structure representation into numerical values that characterize properties of molecules’ [47].

Several thousands of molecular descriptors that are computable by using different software have been defined so far. They can be used in virtual screening e.g. to define correlations between the chemical structure and properties or biological activities, for clustering of compound databases [48], and for database comparisons [49,50]. Generating descriptors is often also the first step in e.g. similarity searches.

Molecular descriptors are calculated from a computational structure representation of the compound, and can generally be classified into e.g. zero, one, two and three dimensional (0D, 1D, 2D, and 3D) descriptors depending on the level of the molecular representation that the descriptors require. 0D and 1D descriptors only depend on the empirical formula of the molecule to calculate e.g. number of different atoms and molecular weight. 2D descriptors rely on the connectivity of bonds between atoms and give information about e.g. presence or absence of a certain functional group or information about branching and cycles. 3D descriptors also depend on the stereochemistry and geometry of the molecule. Typical descriptors calculated from the 3D structure are properties related to the molecular surface area, and the volume of the molecule, properties that also could be estimated from the 2D structure [47]. Examples of descriptors that starts from 3D structures are the alignment independent VolSurf descriptors [51,52] that transforms 3D molecular interaction fields (MIFs) from a GRID [53] calculation into quantitative numerical volume and surface descriptors.

One would think that 3D descriptors would perform best when it comes to selecting active compounds, since it is the 3D structure that constitutes the binding interface between a ligand and a target. However, several studies have shown that 2D descriptors perform better than or equally as well as 3D descriptors [54-57]. When studying NPs it is many times necessary to work with 2D representations since the absolute and relative configuration not
always have been determined, and attempting to use 3D information in such cases may introduce noise in the data set.

**Pfizer’s Rule of Five**

Some descriptors have become more important than others in drug discovery. Based on an analysis of the World Drug Index, Lipinski and co-authors at the drug company Pfizer published what has been known as the Rule of five (Ro5). Pfizer’s Ro5 provided guidelines to evaluate if a chemical compound has properties that would make it likely to be orally available in humans [50]. All numbers are multiples of five, which makes up the origin of the name. Ro5 has raised the awareness about properties and structural features that make compounds drug-like, and it has been adopted by the pharmaceutical industry as an initial guide as to whether compounds fit the known profile of drugs that succeed in development. Ro5 states that poor absorption after oral administration is more likely if two or more of the following is true:

- molecular weight is greater than 500 Da
- the lipophilicity is high (expressed as an octanol-water partition coefficient, logP, greater than 5)
- there are more than 5 hydrogen bond donors (N or O atoms with one or more H atoms)
- there are more than 10 hydrogen bond acceptors (N or O atoms)

The name *Rule* of five is a bit misleading, since it is not a rule, but more of a general guideline. NPs are often cited as an exception to Pfizer’s Ro5, but even Lipinski himself noted [58] that many NPs are bioavailable despite violating the Ro5. In a recent paper [59] NPs, that each led to an approved drug between 1970 and 2006, were analyzed and found to be divided into two equal subsets. One is Ro5 compliant, while the other one violates Ro5 criteria. The two subsets had an identical success rate in delivering an oral drug. The Ro5 should therefore be applied with care, since there is a tendency otherwise that interesting lead compounds may be filtered out.

**Similarity**

In drug discovery it is often of interest to find out how similar one molecule is to another. Similarity searches are very important for e.g. virtual screening of large compound collections [60] and have found particular favour in the pharmaceutical industry [61]. Similarity searches have their philosophical basis in the so called *similarity principle* [62] stating that similar molecules are likely to have similar physico-chemical properties and therefore may
have similar biological activity [63-65]. If the goal is to identify novel lead compounds it is therefore a waste of resources to test molecules that are too similar, but when looking for compounds with a certain activity it is efficient to test compounds that are similar to a known active compound. The expectation is then to find alternative molecular structures that preserve the properties required while enhancing for instance the patentability or optimized pharmacokinetic profiles.

Chemical structure representations, structural fingerprints and molecular descriptors are common inputs in similarity calculations and similarity searches can thereby be both structure based and property based. The input data is compared using a similarity index, where the most widely used metrics are simple distance measures such as the Euclidian distance, and association coefficients such as the Hamming and Tanimoto coefficients. There are also many other similarity measures, see for example, the work of Willett and co-workers [66,67].

Structure-activity relationships

Huge amounts of data are produced in drug discovery today, and the essential information in the data is often hard to extract without e.g. chemometric methods. An important mission for medicinal chemists is to identify interesting chemical structures, which have the potential to become approved drugs. Such research aims to understand the relationship between a compound’s structure and its properties or biological activity. The similarity principle mentioned in previous section is also the basis for structure-property or structure-activity relationships (SPR or SAR) or even finding such relationships in a quantitative manner (QSPR or QSAR). These can then be utilized to help guide e.g. chemical synthesis. Many such relationships are multivariate, where several variables together influence the activity. Multivariate methods are very useful for understanding these relationships and to guide research towards compounds with enhanced biological activity.

The concept of chemical space

Douglas Adams, famous author of the book ‘The Hitchhiker’s Guide to the Galaxy’ [68], quoted that:

Space is big. You just won’t believe how vastly, hugely, mind-bogglingly big it is.

Even though he referred to cosmic space, these words might very well be used to describe chemical space. Chemical space can also be described as a multi-dimensional region, like a coordinate system with multiple axes, repre-
senting a number of physico-chemical properties [69]. The chemical structures are positioned in this coordinate system based on their respective values on the different axes. Chemical space is definitely enormous and basically infinite. By definition it contains all possible molecules, which has been estimated to exceed $10^{60}$, already when only small (with a molecular weight less than 500 Da) carbon-based compounds are considered [70]. The size of chemical space becomes even more impressive if one considers that the average size of a natural protein is 300 residues. Using the 20 amino acids this results in $10^{390} (20^{300})$ tangible proteins. To put this in perspective a typical compound collection at a top pharmaceutical company today contains around a few million compounds [71]. Such compound collections only offer a very modest sampling of all the potential compounds that comprise chemical space and, furthermore, it most presumably comprises historic bias regarding diversity. Following the discussion above, it is not possible, during a lifetime, to thoroughly explore the entire chemical space. Guidance towards more relevant regions is needed.

To make analysis of, or more popularly expressed navigation in, chemical space easier, it can with advantage be divided into smaller sections. One possibility is to reduce the vast theoretical chemical space by looking at the region encompassing only small molecules. Another even more challenging possibility is to identify the limited part of chemical space referred to as biologically relevant chemical space, which is interpreted as the fraction of space where biologically active compounds reside, and where we with higher probability also can find future leads for drug discovery.

Intelligent ways to efficiently navigate through chemical space, extract meaningful information from the large amounts of data obtained in drug discovery, and to select and prioritize which compounds to test for a certain activity out of the endless possibilities of compounds are important tasks [72] and the focus of this thesis.

**Chemography and the ChemGPS methodology**

In 2001 Oprea and Gottfries introduced the concept chemography and the ChemGPS methodology. ChemGPS is a computer based technique for the global investigation of property spaces, and an acronym, which can be interpreted as Chemical Global Positioning System, Chemical Global Property Space, or Chemical Global Property Scores [73-75].

Chemography, defined by the authors as ‘the art of navigating in chemical space’, resembles geography and basically means mapping of objects using chemical descriptors. Their idea was to construct a map, ChemGPS, over the chemical space by using the same principles as Mercator did for his maps [76]. The multi-dimensional space, just like the spherical Earth, can with ChemGPS be illustrated and studied in 2D projections. Instead of longitudes
and latitudes, rules, equivalent to dimensions, are used as different directions, i.e. different axes, on the map. Some of these rules include flexibility, size, hydrogen bond capacity and lipophilicity. Objects on the map are chemical compounds instead of e.g. houses, cities and villages.

ChemGPS is based on a reference set of compounds, for which a fixed list of molecular descriptors is computed. Chemographic map coordinates are extracted by PCA on these molecular descriptors. The results are presented as observation or variable related projections called score and loading plots, respectively. The reference set includes one set of so-called satellite structures and one set of core structures. The satellites are, as the name suggests, compounds that intentionally are different from the core structures and that have extreme values in one or more of the properties. Subsequently they are positioned outside, marking the limits of, the chemical space of interest. The core structures provide a balanced inner region by having average values with respect to the defined rules. The loadings and directions in space are consistent and model rotations are avoided. ChemGPS represents drug-like space and the core structures are therefore mostly orally available drugs which have properties in agreement with Pfizer’s Ro5 [50]. Together, the core and satellite structures are selected to cover the chemical space of interest. This is important since the compound sets analyzed with ChemGPS receive their positions on the map by interpolation from the reference set using PCA score prediction. The whole concept very much resembles Navstar GPS, where satellites are used as references for triangulation of a position on Earth.

One important feature of ChemGPS is that it is a global model, i.e. the same molecule would be positioned on the same spot of the map every time analyzed. The principal properties are consistent and do not change over time. With local models like those normally obtained with PCA, it is difficult to compare different compound sets, especially if the descriptors and data sets used to derive the principal components are not the same. Local models tend to be outdated as additional compounds are added and need to be recalculated (resulting in potentially new loadings) as soon as a new molecule is added or removed to avoid severe projection errors. This is avoided using ChemGPS.
Aims of the thesis

The work presented in this thesis is part of the research at the Division of Pharmacognosy, Department of Medicinal Chemistry, Uppsala University, aimed at exploring means of rational selection, and eventually prediction, when studying biologically active compounds.

The specific objectives of the present thesis are:

- to explore and develop an in silico model for chemographic analyses of biologically relevant chemical space (Paper I & II)

- to implement ways to make the developed chemographic model (ChemGPS-NP) widely available for the scientific community, which in a uniform way then can analyze and chart their chemical compound libraries (Paper III)

- to apply ChemGPS-NP to central scientific issues at different levels in the drug discovery process, such as finding correlations between chemical structures/properties and biological activities (Paper III, IV & V)
Chemical structure representation

As mentioned earlier the chemical structures need to be represented in a suitable form to be useful in computational endeavours. Chemical structures were, when needed, drawn using ChemDraw [77], or extracted from the online databases PubChem [78], ChemSpider [79], or DrugBank [80]. For conversion between various file types e.g. from SDF to SMILES the software ISIS/Base [81] (Paper I) and the command line program MolConverter implemented in MarvinBeans v. 4.0.4 [82] (Paper II-V) were used.

Structure data files

In Paper I all chemical structures were transformed to SDF, which is a common chemical table file format developed by MDL [42]. The connection table is fundamental to all of the chemical table file formats. A connection table contains information about the structural relationships and properties of a collection of atoms. SDF can contain the structural information and associated data items in the file for one or more compounds.

SMILES

In Paper II-V all chemical structures were transformed to linear notations as SMILES [39,40,83]. This format was chosen because it is compact and does not require excessive space to store, it can be imported by most molecule editors for conversion back into 2D drawings, and it is used as input in many software and databases.

SMILES notations consist of a string of characters with no spaces. Chiral or isotopic indications are optional. A notation without this information is called generic SMILES. If chiral and isotopic specifications are included, it is referred to as isomeric SMILES. Algorithms have been developed to generate so called canonical SMILES. A single structure can be represented with several different SMILES depending on in which order the structure elements are described. Canonical SMILES are formed in a specified way, such that for a specific compound only one single SMILES is canonical. These are e.g. used for indexing and to ensure uniqueness of molecules in a
database. In this work generic canonical SMILES have been used throughout.

In the SMILES format non H atoms are represented by their atomic symbols enclosed in square brackets, [ ], with the exceptions of B, C, N, O, P, S, F, Cl, Br, and I. H atoms only have to be specified if they are charged (protons), connected to another H, connected to more than one atom (a bridging H), or if an isotope is used, e.g. heavy water. Atoms in aromatic rings are specified by lower case letters (see benzene, Table 1). Charge can be indicated with one of the symbols + or -. Double and triple bonds are represented by the symbols = (see e.g. carbon dioxide, Table 1) and # (see hydrogen cyanide, Table 1) respectively. Bonds between atoms are, if not specified, assumed to be single. Branches are described with parentheses, where the connection is to the left of the parenthesis (see acetic acid, Table 1). Cyclic structures are represented by breaking one bond in each ring. The bonds are numbered in any order, designating ring opening (or ring closure) bonds by a digit immediately following the atomic symbol at each ring closure (see e.g. cyclo-hexane in Table 1). Disconnected compounds are written as individual structures separated by a ‘.’ (period).

Table 1. Examples of names, empirical formulas, 2D chemical structures and corresponding SMILES.

<table>
<thead>
<tr>
<th></th>
<th>ethane</th>
<th>carbon dioxide</th>
<th>hydrogen cyanide</th>
<th>acetic acid</th>
<th>cyclo-hexane</th>
<th>benzene</th>
</tr>
</thead>
<tbody>
<tr>
<td>empirical formula:</td>
<td>C₂H₆</td>
<td>CO₂</td>
<td>CHN</td>
<td>C₂H₂O₂</td>
<td>C₆H₁₂</td>
<td>C₆H₆</td>
</tr>
<tr>
<td>2D chemical structure:</td>
<td>H₃C−CH₃</td>
<td>O=C=O</td>
<td>HCN</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SMILES:</td>
<td>CC</td>
<td>O=C=O</td>
<td>C#N</td>
<td>CC(=O)O</td>
<td>C1CCCCCCCl</td>
<td>c1cccccl</td>
</tr>
</tbody>
</table>

Daylight fingerprints

In Paper II Daylight fingerprints were used for the initial cluster analysis. Generally, fingerprints are vectors or bit-wise representations of fragments that indicate specific substructure fragments in a compound with a binary bit (1 as present, 0 as absent) at specific positions. The Daylight fingerprinting algorithm generates a pattern for each atom; a pattern representing each atom and its nearest neighbours (plus the bonds that join them); a pattern representing each group of atoms and bonds connected by paths up to two bonds long, continuing with paths up to three, four, five, six and seven bonds long, depending on the size of the molecule (i.e. the path length limit). For example, the molecule below would generate the following patterns:
23-bond paths (i.e. atoms): C, N, O
1-bond paths: C-C, C-N, C=O, C-O
2-bond paths: C-C-C, C-C-N, C-C=O, C-C-O, O=C=O
3-bond paths: C-C-C-O, C-C-C=O, N-C-C-O, N-C-C=O

etc…

In this way the number of possible patterns becomes vast, and it is therefore not possible to assign one particular bit to each pattern. Instead, each substructure pattern is represented (hashed) so that a given number of bits in the bit string are set for each pattern. Each pattern hence generates its particular set of bits. Collisions in the bit string are possible, since the number of bits is limited, but as long as at least one of those bits is unique (not shared with any other substructure pattern present in the molecule), it is possible to tell if that specific substructure pattern is present or not.

Similarity metrics

Tanimoto index

A Tanimoto coefficient, with 0.7 as similarity cut-off to define the cluster sizes, was in Paper II used to assess molecular similarity in an initial cluster analysis of Dictionary of Natural Products (DNP) [84] based on Daylight fingerprints [46].

The Tanimoto index between two molecules A and B described by a fingerprint is calculated using the following expression:

\[ T_{AB} = \frac{c}{a+b-c} \]

In this expression \( c \) is the number of bits common to the two molecules, while \( a \) denotes the number of bits present in A, but not in B, and \( b \) denotes the number of bits present in B, but not in A. Tanimoto may be regarded as the proportion of present bits that is shared. For binary fingerprints the Tanimoto index range is 0 to 1, where 1 means that the two molecules are identical.

Euclidean distance

In Paper V Euclidean distances (EDs) were calculated, using an in-house script written in awk, to determine nearest NP neighbours for a set of regis-
tered drugs. The EDs were calculated between points \( P = (p_1, p_2, \ldots, p_n) \) and \( Q = (q_1, q_2, \ldots, q_n) \) in the Euclidean \( n \)-dimensional space, as defined by the following expression:

\[
\sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \ldots + (p_n - q_n)^2} = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2}
\]

**Molecular descriptors**

The descriptors used in Paper I were calculated based on 2D representations of molecules as SDF using an AstraZeneca in-house program. Size-related descriptors included molecular weight, volume, total surface area, number of atoms, number of C atoms, and the calculated molecular refractivity [85]. Polarizability was estimated by the calculated molecular refractivity and by an atom-based polarizability scheme implemented in the in-house program [86]. Flexibility and rigidity were estimated by: counting the total number of bonds, rings, rotatable bonds, rigid bonds, rigid fragments, C in the longest partially flexible chain, C in the three longest chains only containing rotatable bonds, C in the longest rigid chain, and C in the three largest rings. Hydrogen bond capacity was estimated using: seven HYBOT [87] descriptors including maximum free energy hydrogen bond donor factor (\( C_d \)), sum of \( C_d \) values, the maximum free energy hydrogen bond acceptor factor (\( C_a \)), sum of \( C_a \) values, total strength of hydrogen bond donors and acceptors in the molecule, and the count of hydrogen bond donors and acceptors respectively based on HYBOT definition. In addition, simple counts of O and N atoms, hydrogen bond donors, and hydrogen bond acceptors, together with calculation of the smallest distance between two hydrogen bond donors, between a donor and an acceptor and between two acceptors respectively were included. Charge was estimated using the Gasteiger-Marsili method (G) [88,89], by counting the positive and negative ionization groups, as well as the maximum positive and negative charge, the average positive and negative atomic charge, the differences between the highest and the lowest atomic charges, and the topological dipole moment. Lipophilicity was estimated by calculating the logarithm of the octanol/water partition coefficient, \( \text{clogP} \). Polarity was estimated using the calculated polar surface area, non polar surface area, number of polar and non polar atoms, and number of polar and non polar atoms respectively divided by the molecular weight. Additionally the energy of the \( \pi \)-electron system in \( \beta \)-units based on Hückel molecular orbital method (H) [90-93], the resonance energy of \( \pi \)-electron system in \( \beta \)-units based on H, the energy of the highest occupied molecular orbital in \( \beta \)-units based on H, the energy of the lowest unoccupied molecular
orbital in β-units based on H, the highest positive atomic charge based on combined π-charges from H and σ-charges from G, the lowest negative atomic charge based on G and H as above, the difference between the highest and the lowest atomic charges based on G and H, the average positive atomic charge based on G and H, the average negative atomic charge based on G and H, the topological dipole moment based on G and H, and number of halogens were calculated.

In Paper II and IV all molecular descriptors were calculated using the software Dragon Professional [94], and in Paper III they were calculated using DragonX [95]. Lipophilicity was in these papers estimated by the Ghose-Crippen logarithm of the octanol/water partition coefficient, AlogP [96-98]. Polarity was estimated using descriptors calculating topological polar surface area using N and O contribution or N, O, P, and S contribution [99], the hydrophilic factor [100], as well as the counts for O and N atoms, and aliphatic/aromatic hydroxyl groups. Size and shape related descriptors included molecular weight, counts of atoms, C atoms, non H atoms respectively, and the Ghose-Crippen molar refractivity [96]. Hydrogen bond capacity was measured by counting the number of N and O as donor atoms for hydrogen bonds, and the number of N, O, and F as acceptor atoms for hydrogen bonds excluding N atoms with positive formal charge, higher oxidation states and the pyrrolyl form of N. In addition the counts for O and N atoms, and aliphatic/aromatic hydroxyl groups were used to estimate also this capacity. Polarizability was taken into account through summing atomic polarizabilities and calculating the Ghose-Crippen molar refractivity. Flexibility and rigidity were estimated by counting the total number of bonds, rings, and rotatable bonds [25], as well as calculating the rotatable bond fraction. Upon these the constitutional descriptors sum of atomic van der Waals volumes, sum of atomic Sanderson electronegativity, mean atomic van der Waals volume, mean atomic Sanderson electronegativity, number of non hydrogen bonds, number of multiple bonds (double, triple and aromatic bonds) in a molecule, aromatic ratio, number of double bonds, number of aromatic bonds, number of halogens, and number of benzene like rings, were added together with the functional group counts, number of aromatic carbons (sp²), counts for aliphatic and aromatic primary, secondary and tertiary amides respectively, and the molecular property descriptor Lipinski alert index [50]. The later is a dummy variable taking the value 1 if two or more of the Ro5 properties are out of range. The separate counts for aliphatic and aromatic primary, secondary, and tertiary amides were summed and defined as a new descriptor called ‘number of amides’.

Additionally, VolSurf descriptors, as described by Cruciani and co-workers [51,52], were used in Paper II. Chemical structures were first converted to 3D representations using CONCORD [45], without further processing. VolSurf descriptors were obtained in two steps where the MIFs for the H₂O, DRY and O probes were first computed with GRID [53] and the MIFs
were then automatically converted into simpler molecular descriptors using VolSurf. GRID is a computational procedure for determining energetically favourable binding sites on molecules of known structure. In 3D QSAR, the initial alignment of the compounds in the series is one of the most time-consuming steps. VolSurf produce alignment independent descriptors related to pharmacokinetic properties. The 72 molecular descriptors obtained refer to molecular size and shape, to hydrophilic and hydrophobic regions, and to the balance between them.

In Paper V the same molecular descriptors as in Paper II-IV, excluding the VolSurf descriptors, were used. An additional descriptor, the logarithm of the intrinsic aqueous solubility, was calculated, to distinguish lead-like [101,102] compounds, using the online software ALOGPS 2.1 [103,104].

Before all calculations duplicates, salts, hydration information, and counter-ions were removed and the remaining charges were neutralised using the Daylight toolkit [46] (Paper I and II) and an in-house Perl script (Paper III-V). Another in-house Perl script was used to remove information about stereochemistry and isotopes.

Multivariate data analysis

Pre-processing multivariate data

Before performing chemometric analyses the data was scaled to unit variance and mean-centered [105,106] as implemented in the SIMCA-P+ package [107]. Pre-treatment of data is often needed in chemometrics to make the data suitable for analysis.

Variables commonly have different numerical ranges. A variable with a large range also has a large variance, while it is the opposite for variables with small ranges. Since PCA is a maximum variance projection method, it follows that a variable with a large variance is more likely to be expressed in the modelling than a low variance variable. The standard deviation is calculated for each variable (column), and each variable is then divided by its standard deviation so that all variables get equal variance and equal opportunity of influencing the data analysis. Unit variance scaling is the most common way to scale data and is useful when variables are of different kinds and not directly comparable numerically. It should be noted that sometimes no scaling or other scaling than unit variance scaling work better for some data types. One example of this is when all variables have the same unit, e.g. with spectroscopic data.

In mean-centering the mean of each variable is calculated and subtracted from each value of the respective variables in the data set. After mean-centering, the average value of each variable taken over all observations will be zero, which improves the interpretability of the model. The objective is to
model the variation in the data rather than to determine the offset of the variables. It is used to focus on the fluctuating part of the data and leaves only the relevant variation between the samples for analysis.

PCA

PCA [28,29] was calculated employing the software SIMCA-P+ [107] (Paper I, II, IV and V) and SIMCA-QP [108] (Paper III). PCA is a mathematical method, which has been widely used in drug discovery to transform a multidimensional descriptor space into a more manageable low dimensional space, by filtering out noise while keeping the characteristics that contribute most to the variance of the data. PCA is a good starting point for analyzing multivariate data that rapidly provides an overview of the information hidden in the data. The dimensionality of a data set is the number of variables that are used to describe each object. Correlated variables are compressed into a smaller number of new and per definition uncorrelated variables called principal components (PCs). The PCs can be seen as new variables that summarize the original ones. Hereby PCA provides an overview of all observations or samples in the data table where patterns like clustering, trends, and deviating observations, outliers, can be found. PCA is visualized using two kinds of plots. The score plot shows the relations between the objects (e.g. chemical structures). The loading plot shows the relations between the variables (e.g. molecular descriptors like molecular weight, number of hydrogen bond donors, or number of C atoms). The relative distance between the objects in chemical space is an estimation of how similar they are with regard to the selected variables.

The starting point for PCA is a data table $X$ with $N$ rows and $K$ columns. Each row in the data table corresponds to an observation, while each column in the data table corresponds to one dimension or coordinate axis in a $K$-dimensional space. When discussing chemical space, the rows are most frequently chemical compounds, while the columns often are molecular descriptors. Each row will be represented as a point in the $K$-dimensional space and together all rows (observations) will form a swarm of points. The first PC is the line representing the maximum variation in the data in the $K$-dimensional space (i.e. the data have their greatest “spread” along this first PC). Thereby it follows the length of the swarm and passes through origin. Each observation will be projected down to this line to receive a coordinate, a score, along the first PC.

One PC is often not enough to demonstrate the systemic variation in the multivariate data table. The second PC is perpendicular to the first and accounts for the maximum variation in the data that is not already explained by the first PC, and so for the third, fourth etc. The objective function in PCA is to explain as much as possible of the (remaining) variation in $X$ by each latent variable. A latent variable is defined as a variable that is not directly
observed but inferred from other variables, referred to as manifest variables. To study more PCs give more details, but since the components describe decreasing variation, more and more minute details are revealed. All PCs passes through origin and each observation will be projected on to this line as well, and receive a score in the same way as for the first PC. With two PCs it is possible to create a plane. This plane can be compared to a window, through which it is possible to look into chemical space. All observations can be projected on to this plane as they have coordinates, scores, along both PCs.

Similar objects are situated close to each other or will be grouped, clustered, in the score plots. Variables containing similar information are situated close to each other if positively correlated or diagonally opposite to each other if they are negatively correlated, in the loading plot. The meaning of the scores is explained by the loadings. The loadings for each PC express how the variables are linearly combined to form the scores. The loading values range from \(-1\) to \(+1\). The absolute value and sign of loadings represent the extent and type of correlation (i.e. positive or negative) in which the original variables contribute to the scores. Variables not contributing to a specific PC have loading values of zero. The relation between loading and score plot can be analyzed by comparing objects with high scores in one quadrant with variables in the corresponding or diagonally opposite quadrant in the loading plot.

Together the scores, loadings, and residuals (the variation in \(X\) that cannot be modelled) explain all of the variation in \(X\) and the model is expressed as below, where \(T [N \times A]\), represents the score matrix, where \(A\) denotes the number of latent variables, \(P [K \times A]\) the loading matrix, and \(E\) the residual matrix containing the unexplained \(X\) variation.

\[
X = TP' + E = t_1p_1' + t_2p_2' + t_3p_3' + \ldots + t_Ap_A' + E
\]

The variation of \(X\) that is explained by \(TP'\) is expressed as \(R^2X\). A value close to 1 indicates that most of the variation has been explained.

Cross-validation

Cross-validation as implemented in the SIMCA-P+ package [107] was used to determine the number of significant PCs in the chemometric analyses. In PCA it is very important to know how many PCs that should be included in the model in order to account for most of the data variability but to avoid over-fitting of the model.

In cross-validation a portion of the data matrix is divided into a number of groups. These groups are then successively kept out of the model development in a row-wise/column-wise way. New models are developed from the reduced data and the left-out data is then predicted by the different models,
to be able to compare the predicted values with the actual ones. The prediction error sum of squares (PRESS) is then calculated. PRESS is the squared differences between observed and predicted values for the data kept out of the model fitting. This procedure is repeated for each left-out group, followed by the summation of all partial PRESS-values in terms of an overall PRESS-value. The final PRESS then has contributions from all data and is a measure of the predictive power of the tested model [28, 109].

In SIMCA-P+, cross-validation is performed for each successive model dimension starting with dimension 0. For each additional dimension, cross-validation gives a PRESS, which is compared with the error one would obtain by just guessing the values of the data elements, i.e. the residual sum of squares (RSS) of the previous dimension. When PRESS is not significantly smaller than RSS, the tested dimension is considered insignificant and the model building is arrested. If a new PC enhances the predictive power compared with the preceding PC, the new PC is kept in the model.

Normally, the performance of a PCA model in SIMCA-P+ is evaluated by looking at both the explained variation $R^2_X$ (goodness of fit) and the predicted variation $Q^2_X$ (goodness of prediction). Without a high $R^2_X$ it is impossible to get a high $Q^2_X$. As a rule of thumb, $Q^2_X > 0.5$ is regarded as good, and a $Q^2_X > 0.9$ as excellent, and the difference between $R^2_X$ and $Q^2_X$ should preferably not exceed 0.2-0.3, this, however, largely depends on the source of data.

OPLS-DA

Orthogonal partial least squares discriminant analysis (OPLS-DA) [31, 32] was used in Paper IV to distinguish between the different modes of action of anticancer agents. OPLS-DA models were obtained employing the software SIMCA-P+ [107].

Regression methods can be used to identify relationships between one matrix $X$, usually comprising either spectral or chromatographic data or computed molecular descriptors, and another matrix $Y$, containing quantitative values such as biological responses. OPLS is a latent variable based regression method for relating two separate data tables, $X$ and $Y$. The major improvement in OPLS compared to other multivariate regression methods relates to the separation of systematic variation in $X$ in two parts, one linearly related to $Y$, $T_p [N \times A_p]$, and one unrelated (orthogonal) to $Y$, $T_o [N \times A_o]$. OPLS removes any orthogonal variation from the descriptor matrix $X$ as guided by cross-validation. Partitioning of the $X$-data makes the interpretation of the model easier with the additional benefit that structure interpretation of the non-correlated variation is possible as well. The OPLS model of $X$ is defined as follows, where $T_p$ denotes the predictive scores matrix for $X$, $P_p$ the predictive loading matrix for $X$, $T_o$ is the corresponding $Y$-orthogonal
score matrix, $P_o$ the loading matrix of $Y$-orthogonal components and $E$ denotes the residual matrix of $X$:

$$X = T_pP_p + T_oP_o + E$$

OPLS can be used for discrimination (OPLS-DA) [110-112], to e.g. generate models able to distinguish between different classes. In OPLS-DA all observations are assigned a class-specific numerical value, the $Y_{hat}$, based on the final model. The $Y_{hat}$ value should be as close to 1 as possible to belong to the class specified or as close to 0 as possible not to. In Paper IV a cut-off at $Y_{hat} > 0.5$ was used as the definition of membership to a specified mode of action. It should be noted that if the classes show inhomogeneity or if they differ considerably in size, this is not optimal, since it could lead to overfitting of the model. One way to overcome that problem is to utilize a resampling strategy insensitive to skewed class sizes, as described in the work by Bylesjö and co-workers [110]. Classification by the estimated OPLS-DA model is accomplished in two steps. First, the $Y$-orthogonal variation is removed from the data matrix $X$ as shown below, where $T_o$ is the $Y$-orthogonal score matrix for $X$ and $P_o$ is the $Y$-orthogonal loading matrix.

$$X_p = X - T_oP_o$$

Second, the $Y_{hat}$ is estimated using the updated $X_p$ and the predictive components for the training set.

**SIMCA**

Soft Independent Modelling of Class Analogy (SIMCA) [113] is a method (not to be confused with the software SIMCA-P+) that, beside OPLS-DA, was used in Paper IV for classification of anticancer agents based on mode of action. SIMCA was calculated using the software SIMCA-P+. SIMCA classification uses PCA to generate significance limits for the classes in the scores and the residual direction. One separate PCA model is calculated for each known class of observations. Unknown observations are then predicted as belonging to one or more of these classes. The limits for each class are defined by a 95% confidence interval. The SIMCA output can be, and was in Paper IV, visualized with Cooman’s plot [114], which compares class distances to each other for two classes at a time.

**Statistical molecular design**

Statistical molecular design (SMD) [115,116], also referred to as multivariate design, is frequently used in drug discovery for selecting representative
subsets or training sets of compounds, out of the often enormous chemical libraries available, for e.g. various screening endeavours and for QSAR modelling [72,117,118]. SMD can be seen as a combination of DoE and multivariate analysis. In DoE a statistical design is made in orthogonal and measurable factors, but in SMD the principles of DoE are applied to the PCs. In SMD the compounds are first described using e.g. molecular descriptors. The variables form the matrix, which is subsequently subjected to PCA. The resulting scores are then used as design variables. Subsets selected with SMD have a good chemical diversity and spread in the latent variables. Two kinds of selection methods were used in this thesis, D-optimal design (Paper II and IV) and D-optimal onion design (Paper II). All D-optimal (onion) designs were generated with the software MODDE [119]. In MODDE, the design factors are scaled and centered by default prior to the design.

D-optimal design

D-optimal designs select the most extreme points of the candidate set and give a minimal set of selected compounds with maximum diversity. The D-optimal criterion is based on the theory that a subset of compounds from the candidate set spans a maximum space when the determinant (D) of the matrix is maximized [34]. Selections are made in the periphery of the experimental space to investigate linear terms, and with a center point to detect curvature.

D-optimal onion design

One limitation of D-optimal designs is that the inner regions of the investigated chemical space are not well sampled. Olsson and co-workers have developed D-optimal onion design (DOOD) [35,120] to overcome this limitation of D-optimal design. DOOD divides the candidate set into a number of layers (see Figure 2) where one separate D-optimal selection is made in each layer. Thereby it samples more even throughout the spanned region.
Figure 2. A DOOD with three layers coloured red, green and blue. Selected points are shown as black triangles.

Data sets

In the present work we employed several data sets of different origin.

In Paper I the ChemGPS data set of 531 compounds, as defined by Oprea and Gottfries [74], was used in conjunction with an in-house compiled set of 248 NPs affecting the cyclo-oxygenase (COX) enzymes.

In Paper II 46 different data sets comprising more than one million unique compounds of different origin were evaluated to successively expand the emergent model ChemGPS-NP. Among these 46 data sets were the ChemGPS data set and DNP, released October 2004 [84].

In Paper III three different data sets were used to illustrate how interpretation of large data sets can benefit from ChemGPS-NP. Two of the sets were publicly available via PubChem [78]. The first set of close to 50,000 compounds (assay ID in PubChem (AID):361) had been tested for activity in a pyruvate kinase assay, and the second, a combined set of more than 190,000 compounds (AIDs: 950-952 and 1007-1009), had been tested in a HTS assay to identify regulators of protein interactions between Bim and six Bcl-2 family members, which all have a role in apoptosis regulation. The third set was an in-house set of 40 betalains and two muscaflavines.
In Paper IV three sets of known anticancer agents with previously defined mode of action were used. These were an in-house set of 57 compounds tested on the cell line panel at Uppsala University Hospital, UAS10 [121], the NCI set of 122 compounds tested on the cell line panel of the National Cancer Institute (NCI), NCI60 [122-124], and the kinase inhibitor set, a selection of 48 kinase inhibitors from Biomol [125].

In Paper V three different data sets were used to compare the chemical space coverage by different classes of compounds and to perform a property based similarity search. These included the WOMBAT (World of Molecular BioAcTivity) database [126], version 2007.01, the DNP released October 2004 [84], and the GVKBIO Drug Database version June 2008 [127]. For sorting and organizing of data Filemaker Pro 8.5 [128] and ISIS/Base [81] were used.

WOMBAT is a medicinal chemistry database containing chemical structures and associated experimental biological activity data on 1,820 targets (receptors, enzymes, ion channels, transporters and proteins) for 203,924 records, or 178,210 unique structures [129,130]. The version of DNP used includes 167,169 compounds (126,140 unique compounds) of natural origin, covering large parts of what has been isolated and published in terms of NPs up until the release date. GVKBIO Drug Database contains data on 3,211 drugs approved by the FDA and other authorities extracted from pharmacological journals and other sources.

Cytotoxicity estimation

The cytotoxicity of compounds tested on UAS10 in Paper IV, was measured with the fluorometric microculture cytotoxicity assay, previously described by Larsson and co-workers [131,132]. For experiments, cells were seeded in drug-prepared 96-well or 384-well microplates. Cell survival was presented as survival index defined as fluorescence in test wells in percent of control wells with blank values subtracted. The IC\textsubscript{50} value was defined as the concentration giving a survival index of 50%.

Activity profiles

To obtain an activity profile for the compounds tested on UAS10 cell lines in Paper IV, a custom-made database with functions similar to those of COMPARE [133], used by the Developmental Therapeutics Program at the NCI, was used. The IC\textsubscript{50} value for each drug on every cell line was converted to its logIC\textsubscript{50} value. Average logIC\textsubscript{50} of all cell lines in the panel was subsequently calculated and a delta graph was constructed indicating for each cell line if they proved to be more or less resistant than the average cell line.
DModX

Distance to model in $X$-space, DModX [113], expressed as normalized distances in units of standard deviations, as implemented in SIMCA-P+ [107] was in Paper II calculated for each compound and used to detect moderate outliers for incorporation in the emergent ChemGPS-NP model. With PCA it is possible to discover both strong and moderate outliers. Strong outliers are found in plots of PCA scores and moderate outliers are found by inspecting the model residuals ($E$).

In SIMCA-P+ an ellipse can be seen in the 2D score plots. This denotes the normal area corresponding to 95% confidence by default. Strong outliers are positioned outside of this area and are detected in score plots. They conform to the overall correlation structure of the data, but they have an extreme character and may account for one PC on their own.

Moderate outliers, on the other hand, are found in residual plots and break the correlation structure. A moderate outlier does not have the same effect on the PCA model as a strong outlier does, but they are important to identify because they indicate lack of homogeneity in $X$. Moderate outliers are detected with DModX [113]. DModX gives the distance of a given observation to the model plane.

Determination of lead-likeness

A lead is a new chemical entity that has properties, which makes it valuable as a starting point in drug development. To distinguish lead-like compounds in Paper V the following set of computational cut-off criteria were used, based on previous studies [101,102]: molecular weight less than or equal to 460, the logarithm of the octanol/water partition coefficient between -4 and 4.2, the logarithm of the intrinsic aqueous solubility larger than -5, number of rotatable bonds less than or equal to 10, number of rings less than or equal to 4, number of hydrogen bond donors fewer than or equal to 5, number of hydrogen bond acceptors fewer than or equal to 9.
Results and discussion

The need for a new ChemGPS model (Paper I)

The ChemGPS model, as defined by Oprea and Gottfries [74], was in Paper I applied to NPs using an in-house compiled set of natural cyclo-oxygenase (COX) inhibitors as an example. COX is an endogenous enzyme involved in e.g. inflammation. There is an ongoing research project at the Division of Pharmacognosy, focused on identifying COX-2 inhibitors of natural origin. The structural variability among the identified inhibitors is staggering and it is difficult to see any correlations among the structures by visual inspection. The starting point of this study was therefore to aid in understanding the chemical properties and structural features linked to the enzyme inhibition.

The most interesting, but maybe not surprising, result of this study turned out to be the observation that several natural compounds were projected outside the ChemGPS chemical space, where their properties were rather extrapolated than interpolated from the reference set. This suggests that NPs might occupy unique regions of property 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 (i.e. projected within the ChemGPS chemical space). ChemGPS was, as previously mentioned, mainly based on Pfizer’s Ro5 [50] referring to a number of guidelines describing orally bioavailable drug space, derived from an analysis of the World Drug Index. Additionally, the core structures of ChemGPS have been selected primarily from a list of orally available drugs. In other words, ChemGPS was designed for the drug-like space. Natural compounds are, as supported by several studies, e.g. [21,22,24,134], in many aspects different from many orally available drugs.

Following this, the need for a new model was proposed, to be able to chart the entire biologically relevant chemical space, i.e. both the drug-like and the NP chemical space.

Despite the shortcomings of ChemGPS when used in conjunction with NPs, some general features associated with COX inhibition could be identified. Overall, COX-inhibition was found to be, with a few exceptions, frequently correlated with presence of at least one ring in the structure (in average 3 rings), fragments exhibiting structural rigidity, and relatively large molecular size (average molecular weight was around 300 Da, and average
Those are properties matching most drug-like compounds indicating that it is difficult to distinguish COX inhibitors as a homogenous group. The general properties identified correlate well with the fact that e.g. flavonoids (an example is shown in Figure 3) are frequently identified as active compounds in different COX-assays. There are over 6,000 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 [135].

Additionally, a number of fatty acids inhibiting COX [136] were included in the study. They differ from the majority of the compounds included by their lack of ring systems and by being more flexible indicating that there are different mechanisms of COX inhibition where also flexibility is favourable in at least one of them.

Figure 3. Wogonin, a flavonoid inhibiting the COX-2 enzyme, isolated from *Scutellaria baicalensis*.

The development of ChemGPS-NP (Paper II)

In Paper II the new model proposed in paper I, ChemGPS-NP, was introduced and its development described. With ChemGPS-NP the attempt was to better represent the entire biologically relevant chemical space, including both drug-like compounds and bioactive NPs. ChemGPS-NP defines a map in the same way as ChemGPS, but with a new reference set of satellite and core structures, and a new descriptor array chosen and evaluated from a number of criteria.

A large amount of the core structures were retrieved from a representative portion of DNP [84], and used as starting set. The DNP compounds, represented as SMILES, were pruned of duplicate and erroneous data, and compounds with elements other than H, C, N, O, F, P, S, Cl, Br, or I. Subsequently, cluster analysis of the remaining compounds was performed based on Daylight fingerprints and a Tanimoto index of 0.7 as definition of cluster membership. This resulted in 10,859 clusters when 2,307 singletons, i.e. compounds not similar enough to cluster with any other compound, were included. 1,376 clusters contained more than 50 substances, and the cluster
seeds of these were selected as a preliminary reference set for the emergent ChemGPS-NP.

Selection of descriptor array

The temporary reference set established by these initial steps were used to evaluate a suitable descriptor array. ChemGPS used an in-house program for descriptor calculation. With a long term goal of making the model available for the scientific community, we chose to use a descriptor calculating software that was easy to access. Dragon Professional [94] was used for all descriptor calculations. The 926 descriptors that could be computed from SMILES were calculated for the temporary reference set. This array of descriptors was then trimmed based on a number of criteria. For inclusion a descriptor needed to:

✓ have a comprehensible physical meaning.
✓ reveal loading in at least one component of the PCA model when terminated with cross-validation criterion.
✓ be able to distinguish between the compounds.
✓ encode relevant aspects of molecular complexity.
✓ describe at least one of the following intuitively important molecular properties: lipophilicity, polarity, size/shape, hydrogen bond capacity, polarizability, flexibility, and rigidity.

The resulting descriptor array had 35 descriptors. For this study 46 data sets, comprising in total around one million compounds of different origin, were compiled. For all of these compounds the 35 selected descriptors were calculated. As a first expansion of ChemGPS-NP, a portion of ChemGPS [74] was included to the reference set to cover this overlapping space. ChemGPS was mapped with ChemGPS-NP and 283 compounds which had a DModX larger than the critical value at a probability level of 5%, here 1.17 standard deviations were included in the reference set. Satellites were then chosen from the remaining data sets. This was accomplished by projecting the data sets, one at a time, using the evolving ChemGPS-NP to successively include subsets of the deviating compounds (outliers) to the model, and in this way consecutively enhance the coverage of the biologically relevant chemical space. Compounds in a prediction set with a predicted DModX larger than four standard deviations were considered outliers. If there were less than 20 outliers in a data set they were all added to the reference set. If the data set contained more than 20 outliers a subset of outliers was selected via the D-
optimal design or, if they were more than 100, via DOOD. This procedure was iterated until all data sets had been projected and no additional outliers were retrieved.

The final ChemGPS-NP model has eight PCs, as determined by cross-validation, which explains 92% of the variation ($R^2_X$ was 0.92 and $Q^2_X$ was 0.73).

In practice the first four PCs have been most frequently used. These explain 77% of the variation and can be interpreted as follows; PC1 represents size, shape and polarizability, PC2 corresponds to aromatic and conjugation related properties, PC3 describes lipophilicity, polarity and H-bond capacity, while PC4 expresses flexibility and rigidity.

Comparison between ChemGPS and ChemGPS-NP

Interestingly, in ChemGPS, size and shape were explained in PC1, lipophilicity related parameters described in PC2, while flexibility versus rigidity and polar variables were explained in PC3 [74]. In the present model, ChemGPS-NP, a different order of explained properties was revealed. Lipophilicity was not explained until in PC3, replaced as PC2 by aromaticity and conjugation related properties of the compounds, while flexibility and rigidity were explained in PC4. There can be several possible explanations to this, including the fact that medicinal chemists more often tend to explore hydrophobic interactions between ligands and biological targets [50]. Large distributions in related properties would subsequently lead to an increase in variation in lipophilicity, resulting in a comparably larger variation in this respect dealing with man-designed molecules. Natural compounds in general and secondary metabolites in particular function in a generally hydrophilic environment. In order to retain e.g. supposed defence substances in solution, highly lipophilic substances must be avoided and hence, the variation in lipophilicity is reduced by a functional constraint. With a lower degree of variation follows a lower order of the corresponding PC. A schematic summary of the property interpretation for the three first dimensions of ChemGPS and ChemGPS-NP respectively is given in Figure 4.

The development from ChemGPS to ChemGPS-NP and the following ability to analyse the whole biologically relevant chemical space is an important improvement considering the renewed interest in NPs in the pharmaceutical industry as well as the track record of NPs to serve as inspiration and/or basis for more than 50% of all marketed drugs.
Validation with other descriptor set

As a step in the model validation ChemGPS-NP descriptors were successfully replaced with VolSurf descriptors [52]. VolSurf descriptors are based on 3D representations of the included molecules and their surface properties, which constitute an alternative and different starting point for the molecular description. In this validation process VolSurf descriptors were calculated for ChemGPS-NP reference set. Latent variables were extracted by applying PCA to the calculated VolSurf descriptors and the main properties that could be calculated with VolSurf were compared with the DRAGON related scores. This comparison, using 2D and 3D based description respectively, indicated that similar molecular property dimensions were found by both approaches, which in turn validated that a robust molecular principal property space had been established [74].

Potential applications of ChemGPS-NP

ChemGPS-NP can be applied in several kinds of drug discovery related endeavours. A number of specific examples are provided in Paper III-V. In general terms, ChemGPS-NP can assist in prioritization and selection of suitable lead compounds in drug discovery. With reference to the similarity principle [62] known inhibitors of a certain target could be mapped together with a number of available compounds from which those situated close to the known inhibitors (neighbourhood mapping) can be selected for further
testing, thereby increasing the possibilities of hit generation. An alternative selection procedure could be performed if, for instance, only a small number of compounds from an initial large set can be selected for testing, e.g. because of high screening costs, against a target without prior knowledge of active compounds. The initial set can then be mapped using ChemGPS-NP and a diverse set of compounds can be selected using sampling techniques based on cluster analysis and neighbourhood mapping [54] to explore as large parts of chemical space as possible and at the same time avoiding testing too similar compounds. ChemGPS-NP can also be used for cluster analyses, for evaluation of molecular similarity, and for characterization of large data sets. The coordinates could also be used as unique identifiers, e.g. similar to the CAS number, with the additional advantage that it contains information relating to chemical properties.

ChemGPS-NP can handle the processing of very large data sets, which makes it useful for the analysis of results of HTS campaigns or for instance the chemical diversity in Nature. This could serve as a source of inspiration for the design of new compound libraries.

ChemGPS-NP also gives possibilities to interpret evolutionary driven changes in series of chemical compounds, effectively tracking the evolution of physical chemical properties.

ChemGPS-NP online (Paper III)

Internet has become an important source for information, tools, and services to support medicinal chemists and drug discoverers. Over the last years, a growing number of web-based tools for data analysis in chemistry have been made available [104,137]. In Paper III the ChemGPS-NPWeb, an open access online version of the ChemGPS-NP, was introduced, and the design and features described. ChemGPS-NPWeb enables the scientific community to, via http://chemgps.bmc.uu.se/, analyze and compare chemical libraries in a consistent manner. A screen shot of the web-site is given in Figure 5.

Technically ChemGPS-NPWeb includes a number of different programs and libraries that interact with each other according to the traditional UNIX-model, where each part performs a well defined task and together they produce the desired output. Three main elements are included in the system: DragonX [95] for calculation of molecular descriptors; SIMCA-QP [108] for multivariate predictions; and Batchelor [138], the queue manager web interface that makes it possible to upload SMILES and to download results from the ChemGPS-NPWeb runs, and that allows jobs with long run times to be submitted to the web server and scheduled for later execution by its batch queue. The programs exchange information with the web interface by storing information on the file system, which is a database.
The work flow is as follows: the user uploads the compounds as SMILES. A Perl script makes sure that erroneous SMILES, as well as information about stereochemistry and isotopes are removed, and that the maximum number of compounds (at present 8000) is not exceeded. The SMILES are then submitted to DragonX, where the molecular descriptors are calculated. ChemGPS-NPWeb retrieves 40 molecular descriptors from DragonX. Six of these are subsequently summarized into a new descriptor, ‘number of amides’, by a second Perl script. In total the initial model as well as the final score predictions are based on 35 molecular descriptors, the same ones that were described in Paper II. A third Perl script then organizes and prepares the calculated descriptor matrix. This is subsequently used as input data to a client (cgpsclt) that connects to a server (cgpsd) to run SIMCA-QP, which performs the PCA score prediction, i.e. the actual mapping, via the library (libchemgps). The server subsequently returns the eight coordinates for each compound back to the client, which stores them in the database. The extra step with client/server was incorporated to avoid having to load the ChemGPS-NP reference set for each job. As an additional benefit it also enables predictions to be performed by one or more computers on the network. If the server is not available the prediction will instead be performed locally by a standalone program. Figure 6 describes how the different computing elements interact with each other.
Figure 6. Schematic description of the interaction between the different elements of ChemGPS-NPWeb. Figure reprinted (from Paper III) with permission from Springer.

Users can check the status of their submitted jobs (pending, running or finished) and later download the result from the queue. At the moment, a visualizer for ChemGPS-NPWeb that will e.g. be able to plot the coordinates in three dimensions at a time, colour them according to user defined classifications, and calculate distances between the compounds, is under construction. Until then the coordinates can be plotted using any preferred software. There is also a possibility to access statistics based on results from each of the successive computational steps.

Three different data sets were used to illustrate how interpretation of large data sets can benefit from ChemGPS-NP. The first set of close to 50,000 compounds had been tested for activity in a pyruvate kinase assay [139]. The same set had also been evaluated by Schuffenhauer and co-workers from a scaffold-based viewpoint [140]. The scaffold-based method identified eleven active substances with 2-phenyl-benzooxazole scaffolds that they refer to as a privileged group, i.e. a group with high proportion of active compounds. Mapping with ChemGPS-NP of the same data indicated that the physico-chemical properties of active, non-active and inconclusive compounds tested in this bioassay were largely overlapping. Compounds within the privileged group discussed above all fall well within these parameters, with one exception. In addition to the aberrant physico-chemical properties demonstrated
for this compound, it was also concluded from the bioassay data that it was one of the least active compounds in the privileged group. Colour coding the eleven privileged structures of this group according to their assay responses, emphasize and further narrow the volume of the potentially most interesting active compounds (Figure 7).

![ChemGPS-NP mapping of the privileged group identified by Schuffenhauer and co-workers, colour coded according to legend (green lowest and red highest activity). Figure reprinted (from Paper III) with permission from Springer.](image)

The second set of more than 190,000 compounds, had been tested in a HTS assay to identify regulators of protein interactions between Bim and six Bcl-2 family members, which have a role in apoptosis regulation [141]. The assay results were overviewed with ChemGPS-NP, and the volumes of chemical space in which the confirmed active compounds could be found were identified. From the predicted score plots it was clear that the inactive compounds spanned a much larger volume, where the outer volumes of this space, with no exceptions, were solely populated by inactive compounds, as illustrated in Figure 8.

The earlier discussed similarity principle stipulates that compounds with similar structures and properties in many cases have similar biological activities. Following this it was implied that this information could assist in future selection of test compounds. Compound libraries could be positioned with ChemGPS-NP, and those compounds residing outside the active volume could be excluded for the benefit of compounds located closer to the active compounds. To demonstrate this we selected a subset of 42 of the 140 confirmed active compounds using DOOD. The compounds were positioned in ChemGPS-NP and the ‘active volume’ was defined. Subsequently the
remainder of this compound set was positioned in ChemGPS-NP. After eliminating those compounds positioned outside of the active volume, only 52% of the compound set remained. This set included 91% of the active compounds. Accordingly, if this were a set aimed for future screening in this assay, 48% of the compounds could have been left out after initial ChemGPS-NP positioning, while still finding 91% of the actives, saving both time and resources.

Figure 8. ChemGPS-NP mapping (three first dimensions, PC1-PC3) of all compounds tested in the Bcl-2 assay (A) and confirmed actives (B). All active compounds are localized to one sector of chemical space, forming a cluster with distinct borderlines. Figure reprinted (from Paper III) with permission from Springer.

The third set was an in-house compiled set of betalains, for which an assumed chemical similarity issue was discussed. The betalains are found in nine of the eleven families of the plant order Caryophyllales, where they constitute red, violet, and yellow pigments [142]. Despite their small number and apparent structural homogeneity, the betalains span a comparably large sector of the chemical property space. The two main biosynthetic groups, betacyanins (red and violet pigments) and betaxanthins (yellow pigments), are clearly separated by their physico-chemical properties as demonstrated in Figure 9. The muscaflavins, are identified as yellow pigments in e.g. the common toadstool Amanita muscaria. On biosynthetic grounds, they have been suggested as chemical relatives of the betalains, as they both are biosynthesized from compounds produced by the activity of DOPA-dioxygenase on DOPA (3-hydroxytyrosine) [143,144]. In this study two muscaflavins were included, and they clustered together with the other yellow pigments, the betaxanthins. This supports a long-standing suggestion that the muscaflavins show a ‘chemical similarity’ to the betalains, although originating from very distantly related organisms and partly different biosynthesis routes.
Predicting anticancer mode of action (Paper IV)

In Paper IV it was demonstrated that the mode of action (MOA) of an anticancer agent not only is correlated to its activity profile, as earlier demonstrated and frequently used in predictions [122,133,145-147], but also to the physico-chemical properties derived from chemical structure, as illustrated in Figure 10. The activity profile is a combined graph describing the growth inhibition values from human cancer cell lines. The findings that the models for prediction of MOA based on chemical structure data in ChemGPS-NP chemical space perform well as an initial indication of MOA is an important result since it opens a way to save a considerable amount of work, time, and resources. This implies that it is not always necessary in early evaluation to test the cytotoxic compounds in several concentrations on 60 or 10 cell lines. Neither would it be necessary to initially calculate their activity profiles. In some cases it could instead be adequate to initially test new compounds on only one, or a few, sensitive cell line to confirm cytotoxicity and then use the chemical structure in silico to calculate descriptors and subsequently to predict MOA.
Data from two different cell line panels were used: the cell line panel of 60 human tumour cell lines at the NCI, the NCI60 [124], and the labour efficient alternative to NCI60 with a limited number of ten cell lines, the UAS10 [121], developed at Uppsala University hospital. The data from NCI60 were obtained from the anticancer agent mechanism database [122,123] and the UAS10 data were obtained in-house. The anticancer agents were classified as belonging to one of the following MOA classes: antimetabolites [148-150], alkylating agents [151,152], topoisomerase I- and, topoisomerase II inhibitors [153], and tubulin active agents [154]. PCA was applied to investigate the correlations between MOA and activity profile, where the activity profiles of the in-house and the NCI set were treated as ten and 60 continuous variables respectively. Possible correlations between MOA and chemical structure were investigated with global chemographic [74] mapping of the in-house and the NCI set using ChemGPS-NP. OPLS-DA was used to generate models able to distinguish between the different MOA classes. Additional details on the classification of anticancer agents, the usage of PCA and OPLS-DA, and the sources of the data are provided in the text.
tionally another classification method, SIMCA, was compared with OPLS-DA in terms of classification performance.

A procedure, outlined in Figure 11, for prediction of MOA of anticancer agents was suggested, where ChemGPS-NP is used to predict MOA based on physico-chemical properties derived from chemical structure, followed by subsequent verification by cell line based prediction.

The first step implies global chemographic mapping of novel compounds together with a training set consisting of available compounds with experimentally determined MOA, using ChemGPS-NP. Already from this step, a primary indication is obtained for novel compounds positioned near or in any of the MOA clusters. To be able to distinguish between possibly over-

![Figure 11](image-url)
lapping MOAs the compounds are subsequently predicted in the corresponding OPLS-DA model, providing a secondary and more strongly corroborated indication. These first two steps could in many cases be sufficient, both of which can be performed completely in silico in e.g. a virtual screening endeavour. In a complementing step, compounds are positioned based on their activity profile using the PCA model of the activity profile data from the training set. A statement of MOA is thus obtained for novel compounds positioned in accordance with any of the known clusters. The compounds could then subsequently be predicted in the corresponding activity profile-based OPLS-DA model to distinguish between the proposed MOAs in cases of unclear cluster affinities. Compounds that were not correlated with any of the MOAs using either training set should be further investigated. One possibility is that these compounds comprise activities via multiple MOAs. Another possibility is that these compounds might exhibit novel MOAs [121,122,133,145,146], and could therefore be considered as interesting for further investigation and possibly unique mechanism determination.

A validation was done to confirm accurate prediction results by the proposed procedure. Training sets were selected using D-optimal designs. The training sets were used to build the models and the remainder of the compounds served as external test set. Since all compounds had previously known and documented MOAs, the classification performance could be assessed. OPLS-DA models were calculated for all pair wise combinations of MOA with the activity profile data or the ChemGPS-NP descriptors of compounds from the training set as $X$ matrix (e.g. data for alkylating agents and topoisomerase-I inhibitors in the model aimed to differ between these two MOA classes) and a binary coded $Y$ matrix in which 0 (zero) denotes compounds that do not belong to the specified MOA, and 1 signifies compounds that do belong to the specified MOA. The compounds in the test set, for which MOA could not be determined from the visual inspection, due to positioning between two or more clusters, were predicted in appropriate OPLS-DA model/models to get a $Y_{\text{hat}}$ value. A cut-off at $Y_{\text{hat}} > 0.5$ was used as the definition of membership to a MOA. By using D-optimal designs for selection of diverse and homogenous training set, and since the class sizes of the training set are the same, we found the use of this cut-off to be applicable. In total 94% of the compounds were correctly predicted in the prediction based on activity profile, and in total 87% of the compounds were correctly predicted in the prediction based on chemical structure.

OPLS-DA in this case appeared to have better classification performance than SIMCA analysis. One possible reason for this is that SIMCA analysis, which is based on PCA, model the within cluster variation while OPLS, applied as DA, focus on the between classes variation.

To test the procedure’s ability to discern compounds with novel mechanisms we allowed a set of kinase inhibitors [155] to pass through the different steps. The same procedure with training set selected with D-optimal de-
signs and test set was applied also in this step. The results indicated that the kinase inhibitors may form their own cluster and that they possibly could belong to a different MOA class than those specified in the training set.

In a time with increasing resistance problems in cancer therapy, when the need for new anticancer agents with novel MOAs is urgent, rapid virtual screening based on chemical structure should be of general significance in cancer research.

Exploration of biologically relevant chemical space (Paper V)

In Paper V ChemGPS-NP was used to explore the biologically relevant chemical space in different ways.

The chemical space occupancy by NPs from the database DNP and bioactive medicinal chemistry compounds from the database WOMBAT, here referred to as the medicinal chemistry compounds, was compared and the difference was obvious as illustrated for the first three dimensions in Figure 12.

*Figure 12. ChemGPS-NP projection illustrating the difference in coverage of biologically relevant chemical space by NPs (green), and bioactive medicinal chemistry compounds from the database WOMBAT (black), in the first three PCs. Figure reprinted (from Paper V) with permission from American Chemical Society.*
This indicates, what was also concluded in Paper I, that NPs populate unique regions of biologically relevant chemical space. Interpretation of the ChemGPS-NP scores revealed that NPs are generally more structurally rigid, and less aromatic than the medicinal chemistry compounds. Additionally the distribution of size, lipophilicity, and polarity addressed in PC3 appears to be very similar between the two sets.

Volumes sparsely populated by the medicinal chemistry compounds were many times evolutionarily explored by Nature. Interestingly some of these regions were found to contain possible (or tangible) [101] lead-like [101,102] NPs that could serve as inspiration for combinatorial libraries.

Finally, property based similarity searches were performed to identify NP neighbours of approved drugs by calculation of Euclidean distances (EDs) in eight dimensions based on the ChemGPS-NP coordinates. Selecting a molecule that is similar to those already on the market is a well-tried strategy that certainly increases success rates [27]. Interestingly 99.5% of all drugs had a NP neighbour closer than ED=10, and 85% of the drugs had a NP neighbour closer than the ED=1. This forms a strong argument that NPs has the potential to serve as an important source of inspiration for medicinal chemists. As a comparison, ‘within group’ EDs were calculated between known drug pairs with the same MOA. The average within group ED was 1.8, the median was 1.6, and the standard deviation was 0.9. Naturally, a portion of drug/NP pairs had ED equal to 0, as could be expected since many drugs are of natural origin.

Non-identical NPs with very short EDs to any of the approved drugs were proposed for further testing as potential lead compounds against the target in question. Several of the NPs revealed by this method, were confirmed to exhibit the same activity as their drug neighbours, which supports the use of near neighbours as a good starting point for drug discovery. A number of examples are given in Figure 13. The numbers in parentheses below relate to the numbers in this figure.

For instance the drug/NP pair formestane (1a)/ testolactone (1b) was one interesting pair captured by this method. Testolactone from the NP set, transformed from e.g. progesterone by the fungus *Aspergillus tamarii*, had the ED 0.15 to formestane from the drug set. Both testolactone and formestane are approved aromatase inhibitors used to treat e.g. breast cancer [156]. Also the two NPs 10-epi-8-deoxycumambrin B (1c) and 11βH,13-dihydro-10-epi-8-deoxycumambrin (1d) both isolated from *Stevia yaconesis* had short EDs, of 1.11 and 1.04 respectively, to the approved aromatase inhibitor formestane. The compound 1c is moderately active while 1d has been found to have a pronounced activity [157] as aromatase inhibitor.

Also the HIV-1 integrase inhibiting drug elvitegravir (in phase III clinical trials) (2a) had a close NP neighbour with similar MOA. Integrastatin A (2b), isolated from the fungus *Ascochyta* sp., inhibits HIV-1 integrase [158] and the ED between the two compounds is 2.7.
Another example of an interesting drug/NP pair revealed was 4',5,7-trimethoxyisoflavone (3a), isolated from Ouratea hexasperma, which had the ED 0.4 to the well known anticoagulant drug warfarin (3b). 3a has been shown to exhibit anticoagulant activities [159], just like its drug neighbour. Also, both 1,3-dimethoxy-2-(methoxymethyl)-anthraquinone (3c), isolated from Coussarea macrophylla and galangin from e.g. Helichrysum nitens (3d) were close neighbours to warfarin (ED=0.34 and 0.36 respectively). Any studies performed regarding anticoagulant properties of these two compounds could not be found in literature.

Numerous interesting drug/NP pairs with short EDs, where the activity of the NP remains to be investigated, were highlighted by this method. The neuraminidase inhibitor zanamivir (4a), used to treat e.g. avian flu, was derived from the NP 2-deoxy-2,3-didehydro-N-acetylneuraminic acid (4b) [160,161], a NP widely distributed in animal tissues as well as in bacteria. The ED between these two compounds was 1.9. Zanamivir has a close NP neighbour, N-[2-(Acetylamino)-2-deoxy-β-D-glucopyranosyl]-L-asparagine (4c), within ED 0.4. These two structures do have very similar fragments, but their relative arrangement is very different.

The HIV-1 reverse transcriptase (RT) inhibiting drug lamivudine (5a) had an active NP neighbour in littoraline A (5b), isolated from Hymenocallis littoralis. The ED between the compounds in this drug/NP pair was 3.4. Both of them inhibit HIV-1 RT [162]. Littoraline A was also a close neighbour (ED=3.3) of the HIV-1 RT inhibiting drug zalcitabine (5c). Zalcitabine in turn had three close NP neighbours that, to our knowledge, has not yet been tested for HIV-1 RT inhibiting activity; the structurally very similar NPs pentopyranine A (5d) isolated from Streptomyces griseochromogenes (ED=0.4); clavinimic acid (5e), isolated from Streptomyces clavuligerus (ED=0.4); and dioxolide A (5f) isolated from Streptomyces tendae (ED=0.3). The ED between zalcitabine and lamivudine is 0.2.

The method used here to identify the drug/NP pairs is derived from ChemGPS-NP scores and is thus property based, in contrast to the frequently used fingerprint based similarity search methods. While some of the drug/NP pairs revealed by this property based method are structurally very similar, others are not. Methods based on structural fingerprints would risk missing some of the compound pairs which are structurally dissimilar, but here show up as property neighbours with similar biological activities. One highly appealing feature of property based methods would be an ability to assist in finding new scaffolds for scaffold-hopping or solely as inspiration. Since the revealed neighbours not necessarily are structurally similar it could be possible to overcome toxicological problems, synthetic feasibility issues, and unfavourable ADME properties.
Figure 13. Chemical structures of a selection of the drug/NP pairs revealed in this study. The ED between the compounds is given in parentheses under the corresponding drug/NP pair. The proportion of NPs in DNP with similar or shorter EDs than the selected examples is given in parentheses, in percent, after each NP example, where 0 means that this was the single closest NP. (A) The aromatase inhibitor formestane (1a) and its NP neighbours 1b-d (0, 0.6, and 0.4% respectively). (B) The investigational new HIV-1 integrase inhibiting drug (in phase III clinical trials) (2a) with NP neighbour 2b (8.4%). (C) The anticoagulant drug warfarin (3b) and its NP neighbours 3a (0%), 3c (0%), and 3d (0%). (D) The antiviral drug zanamivir (4a) with close NP neighbours 4b (0.05%), from which zanamivir was derived, and 4c (0%). (E) The HIV-1 RT inhibiting drugs lamivudine (5a) and zalcitabine (5c) and their NP neighbours 5b (14%), and 5d-f (0%).
Conclusions and future perspectives

Referring to the aims, the conclusions of this thesis may be summarized as follows:

✓ Employing ChemGPS [74], an *in silico* model for chemographic analyses of drug-like space, revealed that NPs occupy, in comparison to drug-like compounds, unique regions of property space (Paper I).

✓ A new chemographic model, ChemGPS-NP, with a new reference set, and a modified descriptor array was developed, which proved to be better tuned to chart the whole biologically relevant chemical space, which includes both the drug-like and the NP chemical space (Paper II).

✓ ChemGPS-NP_web, a public online version of ChemGPS-NP, was developed and introduced. This resource now enables the scientific community to analyze and compare chemical libraries in a consistent manner via http://chemgps.bmc.uu.se/ (Paper III).

✓ ChemGPS-NP was applied to several drug discovery related issues (Paper III-V):
  
  o ChemGPS-NP proved to be able to handle and process large data set, and to extract useful information from the results of HTS campaigns (Paper III).

  o By using ChemGPS-NP to identify the active volume of chemical space, a considerably smaller number of compounds could be selected for further testing, while still finding most of the active compounds (Paper III).

  o ChemGPS-NP was successfully used for prediction of mode of action of anticancer agents (Paper IV).
A number of features distinguishing NPs from medicinal chemistry compounds could be revealed after mapping these two sets in ChemGPS-NP chemical space. Several regions, sparsely populated by medicinal chemistry compounds, contain lead-like NPs that could serve as inspiration for combinatorial libraries (Paper V).

Calculation of Euclidean distances based on ChemGPS-NP scores was found to be a useful tool to interpret results and for finding interesting NP inspired leads for drug discovery (Paper V).

The identification of promising compounds from a drug discovery point of view is of great importance in the pharmaceutical industry. ChemGPS-NP has in this work proven to be able to aid in this process, as well as to give useful insights valuable in several stages of drug discovery. With the potential data overload in drug discovery mentioned earlier, models like ChemGPS-NP will be increasingly valuable to keep track of the ever growing vast amounts of data collected in drug discovery.

One appealing thing that makes ChemGPS-NP different from many other models is that it is readily available to the research community through the online implementation.

For some of the studies included in this thesis there are remaining questions of potential importance. Regarding the property neighbours with unknown bioactivity identified in Paper V it would be very appealing to test for, and hopefully confirm, the corresponding activities. It was also illustrated how ChemGPS-NP could differentiate between different modes of action with regard to anticancer agents. It would be interesting to investigate whether this is applicable when it comes to differentiation between other subtypes of target inhibition. Preliminary studies (supporting information of Paper V) indicate that this is the case for e.g. antidepressants, antihypertensives and HIV-1 inhibitors.

As mentioned earlier, the work with a visualizer for ChemGPS-NPWeb is in progress. We believe that this will also greatly improve this resource, since the user instantly, apart from the list of coordinates, will obtain a graphical illustration. Furthermore, an implementation is in progress that will enable property based similarity searches. The user will be able to choose to visualize the five closest neighbours of a selected reference compound based on calculations of Euclidean distances over the eight dimensions.

So far ChemGPS-NP has mostly been used to explore issues related to drug discovery and medicinal chemistry. We believe that ChemGPS-NP have the potential to continue to aid in extracting new information in a diverse range of scientific fields. One possibility is to use ChemGPS-NP in the
field of botany as an aid in achieving a phylogenetic classification. Preliminary studies on the genus *Arnica* [163] showed that it is possible to cluster these species based on chemistry using PCA on gas chromatography-mass spectrometry data. The next step would be to identify the compounds encountered in the different species and use the global map ChemGPS-NP for similar classifications.

Another suggestion is to apply ChemGPS-NP in the field of chemical ecology. In an ongoing project we use ChemGPS-NP to study the advancement of evolution, a context not commonly associated with chemometrics applications. ChemGPS-NP is applied for interpretation of evolutionary driven changes in physico-chemical properties from primitive to more developed species in series of chemical compounds. By mapping the compounds in ChemGPS-NP and then colouring them according to different kinds of information we hope to gain insight regarding several issues, e.g. if there are any trends in chemical complexity from primitive to more developed species along different development lines, or between photosynthesizing organisms and others, and if there are any larger trend changes at the origin of insects, the colonization of land, at the origin of green plants or at the break up of Pangaea and Gondwanaland.

Other possible application areas include e.g. toxicology, where ChemGPS-NP could be used for characterization of toxic compounds, or in the food industry for fast identification of unwanted or wanted properties.

I believe that the work presented in this thesis will be of general scientific interest. The applicability of the model in numerous fields is especially appealing. So is also the finding that anticancer agents cluster in ChemGPS-NP based on their mode of action, a finding that has potential to improve cancer research efficiency.
Många av er har säkert läst boken 'Liftarens guide till galaxen'. Där skriver författaren Douglas Adams apropos rymdens storlek:

Rymden är stor. Ingen skulle kunna tro hur ohyggligt vansinnigt stor den är.


Inom läkemedelsindustrin ackumuleras idag enorma mängder data för vilka mängder av variabler och effekter kan beräknas och mätas samtidigt. Följaktligen kan detta resultera i stora komplexa datatabeller, som är mycket svårtolkade utan hjälp från statistiska metoder. Med hjälp av principalkomponentanalys (PCA) kan man skapa sig en översikt över alla observationer i en datatabell och till exempel hitta grupperingar och avvikande objekt. Tänk dig att varje rad i tabellen motsvarar ett kemiskt ämne och varje kolumn en beräknad eller uppmätt egenskap som utgör en axel i det mångdimensionella koordinatsystem som nämndes ovan. De kemiska ämnenas placeras i koordinatsystemet baserat på vilka värden de har på de olika axlarna. Som sagt så mäter och beräknar man ofta väldigt många variabler. Eftersom människan har svårt att se fler än tre dimensioner i taget använder man sig...
av PCA för att underlätta tolkningen av koordinatsystemet. Egenskaper som beskriver liknande information (som korrelerar) kommer då att slå samman och resultatet blir att man får ett mindre antal nya variabler som kallas för principalkomponenter och som kommer att utgöra de nya axlarna i koordinatsystemet. Dessa nya variabler numreras i storleksordning efter hur stor variation de beskriver. Om ämnen man har analyserat till exempel skiljer sig mest i storlek så kommer den principalkomponent som beskriver storlek att blir den första och så vidare.


I ett av delarbetena i avhandlingen beskrivs användandet av ChemGPS för att karakterisera ett set med antiinflammatoriska substanser med ursprung i naturen. Den viktigaste lärdomen från den här studien var att många av naturprodukterna hamnade utanför kartan. Detta stöder, vad flera andra för-
fattare också kommit fram till, att naturprodukter många gånger är annorlunda och mer komplexa än läkemedelsmolekyler, och det är därför naturligt att deras positioner hamnar i andra delar av den kemiska rymden. Vi föreslog därför ChemGPS-NP som en vidareutveckling av ChemGPS. Med denna ville vi kunna kartera hela den kemiska rymden som täcks av alla biologiskt relevanta strukturer, inklusive både naturprodukter och syntetiska läkemedel.


ifrån varandra som möjligt, för att inte riskera att man testar alltför lika substanser utan utforskar en större mångfald.


Med hjälp av ChemGPS-NP undersökte vi också skillnaden i utbredning i rymden mellan naturprodukter och vanliga läkemedelskemimolekyler. Skillnaden visade sig vara tydlig även om de naturligtvis delvis överlappade varandra. Områden där bara naturprodukter täcker rymden indikerar att man inte utforskat dessa inom läkemedelsindustrin.


Acknowledgements

During my time as a PhD student I have met many wonderful people that in one way or another contributed to this thesis. I would like to express my sincere gratitude to all of you. I would especially like to acknowledge the following persons:

My supervisor Anders Backlund – thank you for your inspiring enthusiasm, for your encouragement, for sharing your thoughts and ideas on scientific and non scientific issues, for believing in me both as a scientist and as a university teacher, and for your friendship.

My supervisor Johan Gottfries – thank you for introducing me to the field of chemometrics, for being a true scientific role model, for taking good care of me during my visits to AstraZeneca, for always giving valuable feed-back on my work, and for being a good friend.

Professor Lars Bohlin – thank you for giving me the opportunity to do my work in this field, for your support, critical reading of papers and of this thesis, and for valuable discussions concerning my work.

Past and present colleagues at the Division of Pharmacognosy for all the good times together and for contributing to the special pharmacognosy spirit during these years; Anders Herrmann – for our talks about all kinds of subjects and for being a good friend; Catarina Ekenäs – for nice collaboration on the Cladistics paper; Erik Hedner – for good times in the gym and for the delicious ballerina crackers on your duty weeks; Jenny Felth – for being such a good friend, for critical reading of this thesis, for co-authorship and for sharing a lot of nice memories from our trips to Holland and Japan; Mariamawit Yeshak Yonathan – my room-mate, for being such a sweet friend and for introducing me to Ethiopian food and tradition; Petra Lindholm – my former room-mate for introducing me to what it is like to be a PhD student and what to watch out for..., it helped me a lot!; Robert Burman – you are really a good guy, and so are the things you bake :-) ; Sonny Larsson – for being a walking dictionary, your knowledge about all those little details really impresses me, thank you also for critical reading of this thesis; Teshome Leta Aboye – for bringing some Ethiopian spirit to the division, and for being a survivor; Erika Svedlund; Martin Sjögren; Sofia Ortlepp; Ulrika...
Huss; Ulf Göransson – for always being helpful and for being a very competent scientist; Hesham El-Seedi – for our nice conversations and your kind support; Kerstin Ståhlberg – for being such a caring and supportive person; Maj Bladh and Siv Bergström – for administrative help during hectic teaching periods; and our new colleagues – Cecilia Alsmark, Christina Wedén, and Sunithi Gunasekera.

Past and present PhD colleagues at our neighbouring department – for nice company and discussions in the fika room; Anita Persson, Annica Tevell Åberg, Henrik Lodén, Ingela Sundström, Jakob Haglôf, Matilda Lampinen Salomonsson, Marika Reis, Viktoria Barclay, and Ylva Hedeland.

My co-authors and collaborators – for scientific contributions and valuable discussions concerning my work; Sorel Muresan and Tudor I. Oprea – you are true mentors and great scientists, it has been a pleasure to work with you; Thierry Kogej and Anders Lövgren – for a great job with ChemGPS-NPWeb; Joachim Gullbo, Rolf Larsson, and Linda Rickardson.

All my students over the years.

Umetrics and Talete – for software support to ChemGPS-NPWeb.

Småland’s nation and Swedish Academy of Pharmaceutical Sciences – for travel grants.

All of my friends outside academia – for being just that.

My wonderful sisters Lisa and Julia, my mother Gerd and my father Per - for always being there for me, for your endless support and for always believing in me. I love you!

Last but not least, Daniel, the love of my life, my best friend, ‘my gooding’ and my greatest support in all situations. You mean the world to me!

Thank you all!

Uppsala, February 2009


51 VolSurf is available from Molecular Discovery Ltd., http://www.moldiscovery.com/


ChemSpider: http://www.chemspider.com/

DrugBank: http://www.drugbank.ca/

ISIS/Base 2.5 SP2. MDL Informations Systems, Inc. http://www.mdl.com/


Hückel, E. (1931) Quantentheoretische Beiträge zum Benzolproblem. II. Quantentheorie der induzierten Polaritäten. Z *Phys* 72, 310-337.


Dragon Professional 5.3 software, Talete srl, Milano, Italy. http://www.talete.mi.it

DragonX software, linux version, Talete srl, Milano, Italy, http://www.talete.mi.it/


VCCLAB, Virtual Computational Chemistry Laboratory, http://www.vcclab.org/


SIMCA-P+ 11.5 software, Umetrics AB, Umeå, Sweden, http://www.umetrics.com/


119 MODDE 7 software, Umetrics AB, Umeå, Sweden, http://www.umetrics.com/


125 Biomial: http://www.biomial.com/


Filemaker Pro 8.5. Filemaker Inc., http://filemaker.com/


Batchelor (2008), http://it.bmc.uu.se/andlov/proj/batchelor/


duced by the atherogenic oxidized phospholipid oxPAPC. *Thromb Res* **120**, 71-79.


Acta Universitatis Upsaliensis

Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Pharmacy 89

Editor: The Dean of the Faculty of Pharmacy

A doctoral dissertation from the Faculty of Pharmacy, Uppsala University, is usually a summary of a number of papers. A few copies of the complete dissertation are kept at major Swedish research libraries, while the summary alone is distributed internationally through the series Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Pharmacy. (Prior to January, 2005, the series was published under the title “Comprehensive Summaries of Uppsala Dissertations from the Faculty of Pharmacy”.)