CHAPTER 2

Substructure-Based Approaches to GPCR Drug Discovery and Design

This chapter is based upon:

van der Horst, E.; IJzerman, A. P. Computational Approaches to Fragment and Substructure Discovery and Evaluation. In *Fragment-Based Drug Discovery: A Practical Approach*; Zartler, E. R.; Shapiro M. J., Eds.; John Wiley & Sons, Ltd: Chichester, West Sussex, U.K., 2008.

2.1 Introduction

Nowadays, large molecular databases are easily accessible to the research community. This is illustrated by the advent of free online resources such as PubChem[1] and eMolecules[2]. These publicly available databases consist of structure and property data for millions of small molecules. Both databases are accessible through web-based search tools, and are thus an unprecedented source of small molecule data. Outside the public domain, a similar progress takes place. Large molecular databases are becoming available that include bioactivity data, for example WOMBAT[3] (WOrld of Molecular BioAcTivity). Molecular data from these sources may be used to construct predictive models, such as Structure Activity/Property Relationships (SARs/SPRs), or classification models. These models can be based on molecular properties, such as lipophilicity, solubility, and molar weight, but also on molecular structures *per se*. *In silico* fragmentation of molecular structures is often used to provide a dataset of structural elements of the intact molecule. Analysis of the resulting fragments is useful to derive novel classifiers *e.g.* for predicting activity of new molecules. What is meant by the term *fragment* depends on the context. In the chemical sense, a fragment is a small, low-molecular weight substance with weak affinity often used to 'build' a higher affinity lead compound. This is different from the computational sense. In the computational context, the term fragment, or substructure, denotes some structural part of the 2D structure of a molecule. It is the result of fragmentation of the molecule according to some "breaking rules". This chapter focuses on the *computational fragment*. We review fragment discovery and evaluation in the context of large molecular databases as described in current literature. Definitions, use and applications of fragments are addressed as well as fragmentation methods. Fragmentation of 3D molecular structures will not be discussed.[4] In the first part (section 2.2), we will discuss the ways in which fragments can be derived. In the second part (section 2.3), a few examples of what can be learned from such fragmentation methods are presented together with their applications.

2.2 Fragmentation Methods

What is considered a fragment depends on the definition. A 'ring' could be a fragment, or a particular chain of carbon atoms could be a fragment. The definition follows from the breaking rules that are used. To find structural patterns in a database, molecules should be broken into manageable parts that are readily analyzed. Graph theory is extensively used to this end (see section 2.2.1). There are two approaches to molecule fragmentation. The first approach is to find all possible fragments that form some part of the molecular structure; the second is to dissect the molecule into fragments according to predefined (breaking) rules. The first approach allows a complete analysis of the fragments that exist in the set. However, the number of substructures for a single structure may then become very large, even for a moderately sized molecule. Several methods allow considering all (potential) fragments for analysis without generation of the full substructure set. The substructure approach will be the subject of sections 2.2.2.1 and 2.2.2.2. The second fragmentation approach generally has a lower yield of fragments per molecule. Fragments result from 'breaking' the molecular structure into non-overlapping, predefined parts. Thus, 'ring structures' may be defined as well as functional groups. Fragmentation into molecular building blocks according to predefined rules follows in sections 2.2.3.1 and 2.2.3.2.

2.2.1 Graph Representation

Graph theory plays an important role in fragmentation. The 2D structure of a molecule and its fragments are often represented as graphs.[5] A graph is a mathematical object that consists of a set of vertices, or nodes, and a set of edges that connect these nodes. The molecular structure conveniently translates into a graph, where vertices represent the atoms and edges represent the bonds.[5] This abstraction enables the use of generic methods that are under study in graph theory, such as the discovery of rings (cycles).

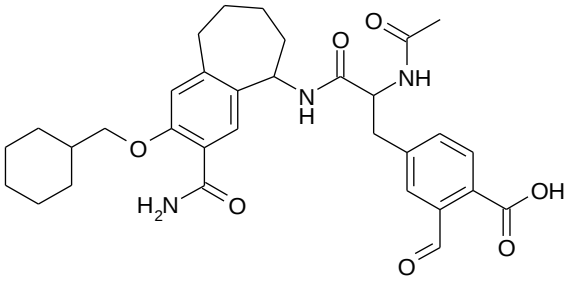


Figure 1. Example structure taken from the PubChem compound database.1 IUPAC name: 4-[(2S)-2-acetamido-2-[[(2S)-10-carbamoyl-9-(cyclohexylmethoxy)-2-bicyclo[5.4.0]undeca-7,9,11-trienyl]carbamoyl]ethyl]-2-formyl-benzoic acid, PubChem CID: 9959891.

To illustrate the representation of molecules as graph, let us consider the sample structure in Figure 1 (taken from the PubChem compound database[1], accession number CID9959891). Figure 2 shows the graph representation of the molecule in Figure 1. Hydrogen atoms even when connected to heteroatoms are omitted. Note that with standard graphs, representation of the molecule is limited to reproducing the connection pattern (connectivity) between the atoms. Any other information such as atom type or bond order is disregarded.

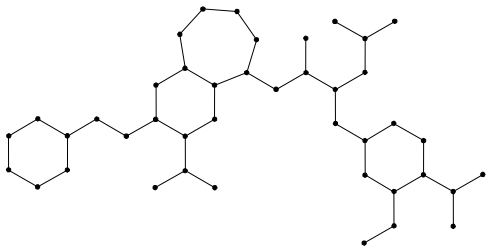


Figure 2. Graph representation of the example structure in Figure 1. Nodes (black dots) represent the atoms and edges (solid lines) represent the bonds. Note that standard graph representation disregards any extra information such atom type or bond order.

2.2.2 Substructure Methods

2.2.2.1 Frequent Subgraph Mining

Graph-based data mining aims to find interesting patterns in graph data. It has a variety of applications, such as analysis of literature citation networks, weblogs, and web searches. Frequent subgraph mining is the process of finding all frequently recurring topological patterns in a database. In drug discovery, frequent subgraph mining, or fragment mining in the context of molecular databases, can be used to find structural patterns that are frequent in one class of compounds and infrequent in the other. First, the general procedure of subgraph mining will be described. After that, a number of algorithms and tools for molecular fragment mining will be presented.

To find the frequently occurring fragments in a set of graphs, a typical algorithm would enumerate all possible fragments that exist in the set, and find for each fragment the graphs in which it occurs. The frequency of a fragment is the number of graphs in which it occurs. The process of testing whether a fragment is part of a graph is called subgraph isomorphism testing. It searches the graph for a subgraph that is isomorphic to the fragment. A typical example is the ethyl fragment (C-C) in n-propane (C-C-C); it occurs twice, and the one is 'isomorphic' to the other. In terms of computing steps, graph/subgraph isomorphism tests are relatively costly. This translates to prolonged computing time or memory requirements. It is one of the key issues in graph mining since there currently exist no efficient algorithms for isomorphism testing on general graphs. In the worst case, the number of computing steps is exponentially proportional to graph size, which contributes to the inefficiency of an algorithm. Therefore, most algorithms seek ways to avoid graph/subgraph isomorphism tests as much as possible.

Starting from an empty fragment, all possible fragment extensions (refinements) are generated, a process that will be explained below for the simple amino acid alanine. This is done by recursively adding edges and nodes to already generated fragments. In case of a ring closure, only an edge is added. Generated fragments are compared against the graphs in the database to check whether they occur. New refinements can

only appear in those graphs that already hold the original fragment. Accordingly, the algorithm keeps appearance lists to restrict isomorphism testing to the graphs in the lists only. The support for a fragment is the proportion, or percentage, of graphs in the database it occurs in. Obviously, found fragments are more relevant if they occur in at least a given minimum number/fraction of molecules. This minimum is called the minimum support value. Fragments are discarded if they occur in fewer molecules than allowed for the minimum support value, which is related to the significance of the found fragments. In general, lower minimum support values will yield higher numbers of fragments. Choosing a sufficiently high minimum support value will result in a comprehensible number of fragments while mining is completed within a reasonable timescale. By definition, the support value of a fragment never exceeds the support values of the fragments it contains. This restricts refinement generation further, starting only from fragments with sufficient support (*cf*. Apriori-rule[6]). To focus isomorphism testing, fragment-mining algorithms may keep a mapping of the nodes and edges of a fragment to the corresponding nodes and edges of the graph in which it occurs. This is known as an embedding.

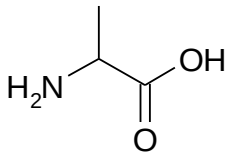


Figure 3. Chemical structure of alanine. Implicit hydrogens are omitted.

To illustrate the process, let us consider a graph mining experiment on a molecule database with alanine (Figure 3). For a single molecule in the database, such as alanine, a search tree can be constructed of all possible fragments. Figure 4 shows all these fragments for alanine with hydrogen atoms omitted as discussed before. On top is an empty fragment and each following fragment is a substructure of its descendants below. Fragments on the same level (six in total) have the same number of bonds (edges). For instance, the first level contains the elements N, O, and C, since these are the constituents of the molecule. The C-C fragment on the second level forms the common core for the C-C-N, C-C-C, C-C=O, and C-C-O fragment

on the third level. The arrows indicate the paths leading from an empty fragment to the complete structure, yielding one extension at a time.

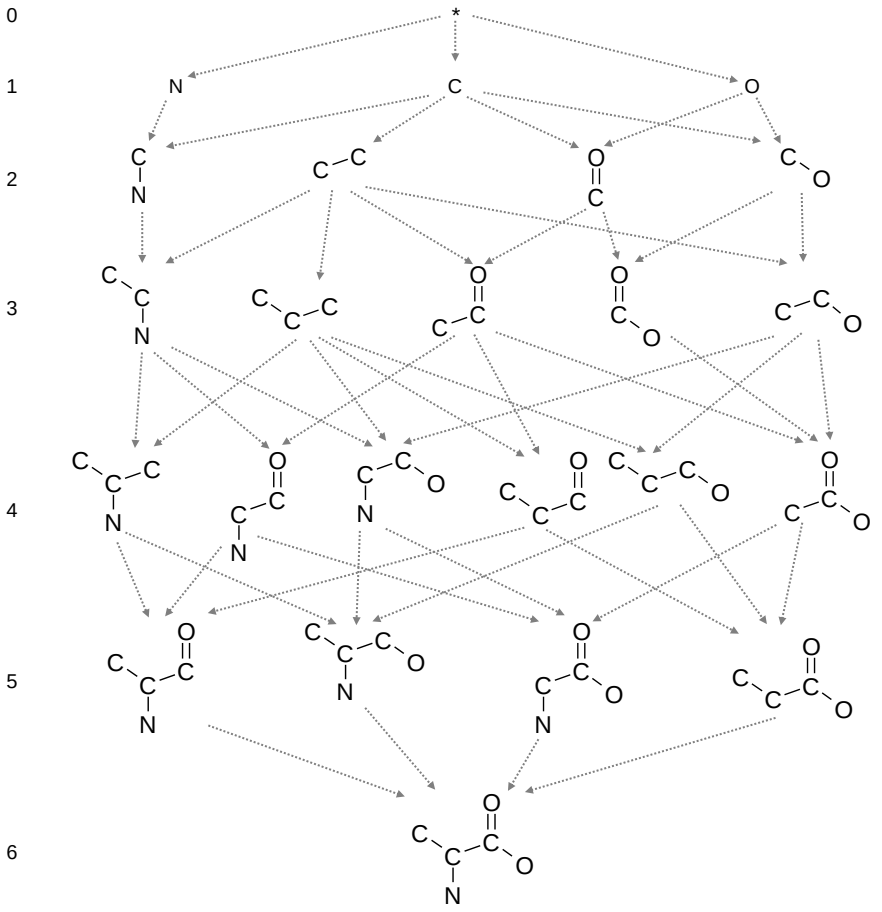


Figure 4. The complete lattice of substructures for alanine (bottom). Level numbers in the lattice increase with fragment size until the final structure of alanine is reached.

There are two ways to travel the subgraph lattice as the entire scheme in Figure 4 is called: breadth-first and depth-first. A breadth-first search considers all refinements at the same level before advancing to the next. For Figure 4 this means stepping through the lattice one row of fragments at a time. Storage requirements are proportional to the maximal number of subgraphs at one level. Depth-first searching requires less storage, since a graph is completely searched before advancing to the next. Therefore, it is proportional to the size of the biggest graph. Modern graph mining algorithms such as the ones described below, work in a depth-first manner.

There are three problems central to frequent subgraph mining; the difference between algorithms lies in how they address these problems. First, as was mentioned, subgraph isomorphism tests are expensive in terms of computation steps needed to perform the search. Second, the generation of refinements should be restricted. Third, since generated duplicates require isomorphism tests, their number should be kept to a minimum, *e.g.* by using a unique graph representation for testing.

The fragment miner MoFa (Molecule Fragment Miner)[7] was made especially for the purpose of molecule mining. All embeddings are stored and used for isomorphism testing and for restriction of fragment extensions to refinements that actually exist in the database. To reduce the number of generated refinements, MoFa sorts all nodes and edges of a fragment in the order in which they were added. Refinements may only occur at the same or newer nodes. Nonetheless, many duplicates are generated, with time-consuming isomorphism tests as a consequence. Two extensions exist for MoFa; the first treats rings as single units and the other treats chains of arbitrary length as a single unit. One of the advantages of treating rings as single units becomes clear when fragmenting steroid structures. Normally, MoFa considers more than 300,000 fragments per steroid, whereas the ring extension generates only 93 fragments. Another advantage is that the ambiguity of aromatic bond representations in rings, either single or double, is circumvented.

In gSpan (graph-based Substructure pattern)[8], a canonical graph representation is used, that is constructed from the concatenation of edge representations in the order in which they are visited. To generate unique representations, the algorithm dictates a strict, depth-first traversal of the subgraph lattice, hence the name 'depth-first search' code (dfs-code). Since the string of concatenated edge/vector representations resembles the sequence of letters in a word, graph representations are compared in the same way, that is, lexicographically. The elements of the string are sequentially compared until a mismatch is found or if one string ends. Lower edge/vector labels precede higher ones; if all labels match, the shorter string precedes the longer. The dfs-code of a fragment determines which nodes can be extended, thereby restricting the number of refinements for that fragment. Only those refinements are generated that have the smallest dfs-code. Appearance lists are used instead of embeddings; hence, subgraph isomorphism tests are still necessary for the graphs in these appearance lists.

FFSM (Fast Frequent Subgraph Mining)[9] uses a canonical code, the Canonical Adjacency Matrix (CAM) code, to identify isomorphic graphs and to restrict refinement generation. It is based on a matrix representation of the graph. By concatenating all entries of the matrix, a string is formed that is used for lexicographic ordering of the graphs. FFSM stores embeddings for the nodes only. In this way, embeddings are rapidly created for new fragments made from joining or extension.

Gaston (GrAph/Sequence/Tree extractiON)[10] exploits the fact that various types of substructures are contained in each other, and that for the simple types more efficient algorithms exist. First, only paths are considered in a substructure search. After that, paths are transformed to trees and trees are searched. Finally, trees are transformed to general graphs with cycles. This type of graph requires the most advanced and time consuming algorithms. As stated before, finding subgraph isomorphisms is a laborious task compared to other search problems, and therefore time consuming. Therefore, they are only used, when they are really needed. Gaston stores all embeddings, in

order to restrict generation of fragment refinements to those that actually appear in the database, and for isomorphism testing.

Table 1. Comparison of four frequent substructure-mining algorithms in terms of performance.[a]

	Total runtime (min)		Time / fragment (sec)		Memory (GB)	
Support	**5%**	**20%**	**5%**	**20%**	**5%**	**20 %**
Gaston[10]	**7.4**	**2.5**	**0.1**	**0.4**	**1.3**	**0.9**
gSpan[8]	**19**	**4.5**	**0.3**	**0.8**	**0.3**	**0.2**
FFSM[9]	**19**	**8.3**	**0.3**	**1.5**	**1.2**	**0.9**
MoFa[7]	**80**	**11.8**	**1.1**	**2.2**	**0.6**	**0.6**

[a] Performance was measured by applying each algorithm to the NCI HIV database (42689 compounds). Runtime and memory usage are provided for two support thresholds: 5% and 20%. The runtime per fragment found is also provided to correct for the runtime overhead due to the higher number of fragments at lower support values. Data taken from performance charts from Wörlein *et al*.[11]

The tools have recently been compared and evaluated in the context of molecule mining.[11] Wörlein *et al.* reimplemented all four methods (same code base, programming expertise and optimization effort). Benchmarks were carried out on a comprehensive set of graph databases, including molecular databases. The molecular databases used were the IC93 (1,283 compounds),[12] the HIV assays 1999 (42,689 compounds),[13] and the NCI (237,771 compounds)[14].

The IC93 database served to investigate how the algorithms behaved when the number of found fragments and the fragments themselves get large. For example, a support value of 4% resulted in 37,727 fragments of which the largest had 22 bonds. The HIV database served to measure performance, whereas the NCI was used to test how the algorithms scale with increasing database size. For this, molecules were

randomly divided into sets of various sizes. Sample measurements are provided for illustrating the quantitative comparison of the algorithms. Table 1 lists the performance measurements for the algorithms applied to the HIV data. The runtime of the algorithms increases with lower support values. Gaston was the fastest and MoFa the slowest algorithm. However, Gaston used the highest amount of memory, whereas MoFa needed less. gSpan had the lowest memory requirements. Note that these figures may differ for other data sets. Size and contents of the database, the minimum support value, as well as implementation details and even the underlying hardware architecture may influence performance of the algorithm. The data in Table 1 are indicative for the overall outcome of the quantitative comparison. For all algorithms, lower support values resulted in an exponential rise in runtime. This is probably due to the runtime overhead caused by the exponential rise in found fragments at lower support values. The benchmark results permitted a ranking of methods. gSpan needed the least memory, since it does not use embedding lists. MoFa, which stores only one subgraph embedding per node in the search tree, was also memory efficient. FFSM required more memory than gSpan and MoFa, probably because it stores the main subgraphs together in a node in the search tree. Gaston needed most memory, since with this method embedding lists for new fragments are based on those of 'parent' fragments. Extensions to the parent's list are stored with the 'children'. The size of embedding lists also depends on the number of children per fragment.

In terms of runtime, Gaston was always the faster algorithm, except at lower support values on the complete NCI. The gSpan algorithm was faster than FFSM for the large datasets, although FFSM was faster than gSpan for the IC93 dataset. Embedding lists are not used in gSpan, which, in fact, speed up testing, especially for larger fragments. MoFa was always the slowest algorithm. The authors suggested that the slowdown of Gaston at lower support values on the complete NCI was due to the large amount of bookkeeping related to the vast number of embeddings. This results in a slowdown due to memory operations. However, the authors found that this effect varies for different systems. Some memory architectures penalize the memory-intensive

operations of Gaston. Although MoFa was the slowest in all tests, it offers more functionality for molecular databases, *e.g.* there is an extension for treating rings as single entities as mentioned above.[15] Another extension offers finding fragments with carbon chains of varying length. This can be useful for the exploration of biochemical reactions where this length is less important.[16]

Interestingly, the four fragment miners mentioned above have been made available as a single package named ParMol (Parallel Molecular Mining)[17]. In addition to uniform access to MoFa, gSpan, FFSM, and Gaston, the authors included a 2D viewer for molecular structures, parallel (multiprocessor) search, and support for several file formats such as SMILES and SDF, and a number of options to customize mining.

Other algorithms for frequent fragment mining that are more database-centric include Molfea[18] and Warmr[19]. Molfea (Molecular Feature Miner)[18] is in essence an inductive database framework. It finds patterns based on first-order logic. Molecules are encoded as basic facts, and queries result in a combination of facts. The fragments that can be searched for or result from queries, are linear sequences of non-hydrogen atoms and bonds. The fact that Molfea only finds chains of atoms limits its usefulness since almost all molecules have rings or branching points. Warmr[19] is a general-purpose Inductive Logic Programming (ILP)* data-mining tool for finding frequently occurring patters in relational data.[20] It has been successfully applied to chemical data, for instance to find frequent substructures in carcinogenic compounds. First, molecules are described in a relational language. Atoms are related to molecules, and to other atoms through bonds. Algorithms such as Warmr perform multi-relational data mining,

* ILP (Inductive Logic Programming) is a machine learning technique used for knowledge discovery. The purpose of ILP is hypothesis generation, given some background knowledge, and a set of positive and negative examples. Examples and background knowledge are encoded as a facts and rules in a relational database. From this, possible hypotheses are generated through inductive learning. Logic programming is used to represent examples, background knowledge, and hypotheses, in a uniform way.

which means they are capable of finding patterns that span across multiple relations. Warmr searches the available patterns in a breadth-first manner, starting from the most general relations, and gradually increasing the level of complexity, to find patterns that are more specific. Candidates that are more specific are generated by pruning non-frequent patterns from the next level. Several meaningful relationships were reported for application of ILP on toxicity data. Although Warmr should be able to produce identical results compared to the fragment miners, it inherits some of the drawbacks related to ILP. First, a high level of expertise is required to encode the molecules, i.e. the graph and their properties, into relations that can be mined. Second, the complexity of relations queried, places high demands on computing resources [19]

2.2.2.2 Common Substructures

Fragments are also derived by comparing molecular structures. For a pair of molecules, a number of substructures/fragments may exist that occur in both structures. A "common substructure" is a set of atoms that two molecules have in common. Corresponding atoms should have the same atom type and the same topological distance to other common atoms, in both molecules. The topological distance is the number of bonds that form the shortest path between two atoms. The "maximum common substructure" (MCS) is a continuously bonded substructure that has the highest number of common atoms.[21] Note that there may be multiple MCS's for a pair of molecules. Figure 5 shows an example of the MCS of two molecules, of which the largest is the molecule from Figure 1. The "highest-scoring common substructure" (HSCS)[21] is similar to the MCS, but also allows discontinuous common substructures. Scores are based on the number of common atoms, and are corrected with a penalty for discontinuous pieces. In Figure 5, the HSCS and MCS are equal. Common substructure methods, such as the MCS and HSCS, are used to detect and visualize structural similarities between molecules.[21] In addition, the HSCS has been applied for discovery of common chemical replacements and to find fragments associated with multiple biological activities.[22,23] These applications will be described in section 2.3.3.

2.2.3 Building Blocks

The fragmentation methods described in the previous section all use the "full substructure set". Despite the high level of detail of these approaches, exhaustive study of all possible fragments can be costly, however. A more restrictive, still sensible, approach may be to focus on chemically meaningful fragments only, instead of including every single fragment in a study.

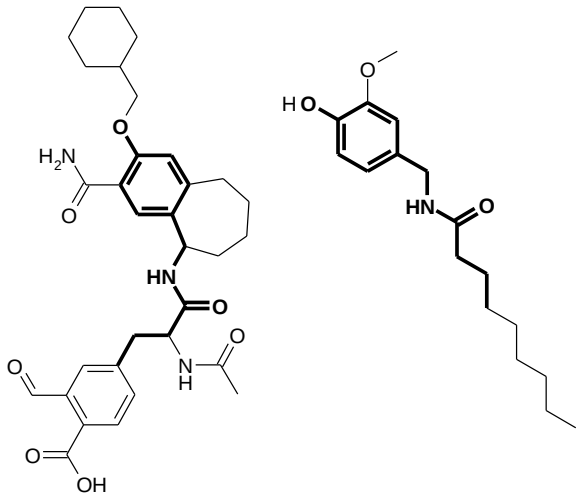


Figure 5. Maximum Common Substructure of two molecules (MCS, drawn in bold). The structure on the left is the example structure from Figure 1. The structure on the right is vanillyl-N-nonylamide, IUPAC name: N-[(4-hydroxy-3-methoxy-phenyl)methyl]nonanamide. PubChem CID: 2998.

2.2.3.1 Molecular Building Blocks

To accomplish this, compounds are dissected into molecular building blocks. This method splits molecules into non-overlapping structural parts according to a predefined set of breaking rules. These rules follow from the definition of individual building blocks. This approach yields (chemically) more intuitive fragments such as rings/ring systems, linkers, side chains, functional groups, *etc.* Figure 6 illustrates the

derivation of building blocks. A typical compound (Figure 6-a) is fragmented into molecular parts, according to the method described by Bemis *et al.*[28]. Three ring systems (Figure 6-d) are at the core of this compound, which are connected by two linkers (Figure 6-e). Together, ring systems and linkers form the molecular framework (Figure 6-c). Attached to this framework are the five side chains (Figure 6-b), yielding the complete molecule. There are many variations to this method; most methods differ in the precise definition of building blocks.

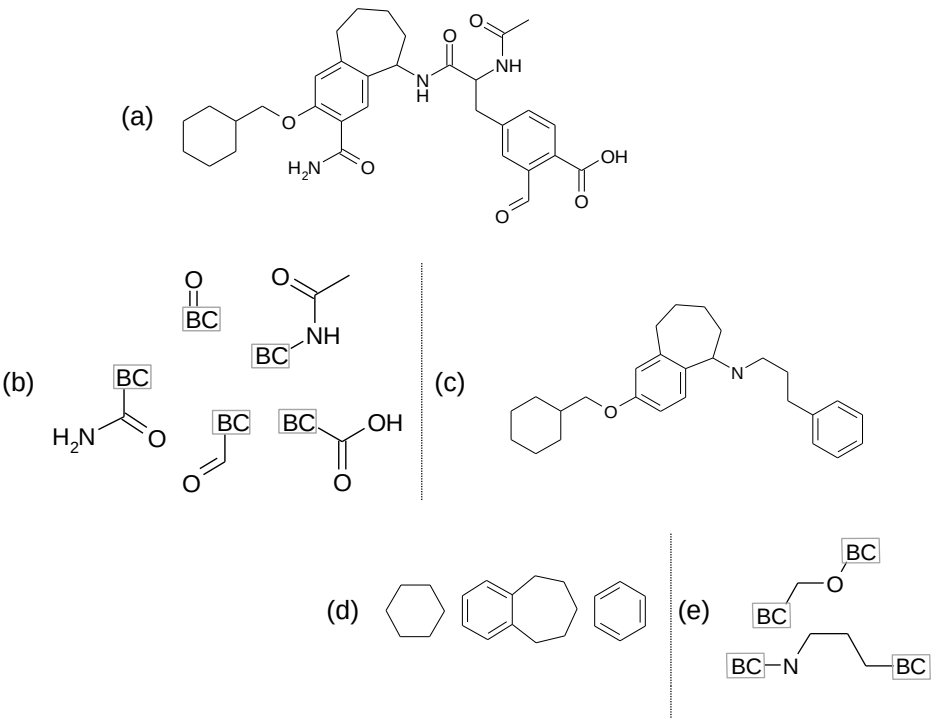


Figure 6. Molecular building blocks, according to Bemis *et al.*[28]. (a) The structure that will be fragmented (CID9959891, see Figure 1). By removing (b) the side chains from this structure, (c) the molecular framework is revealed. The framework consists of one or more (d) ring systems connected by (e) linkers. The connection point to the framework or rings is indicated by a rectangular label composed of the letter B and the atom type that it is connected to. For instance, the BC label means a carbon connection point in the framework.

2.2.3.2 Virtual Retro-Synthesis

Another way to split a molecule into smaller parts is by virtual retro-synthesis. This method applies a set of breaking rules based on chemical reactions. Bonds that are typically formed by one of these reactions, are cleaved, essentially reversing synthesis. The resulting fragments are precursors from which the molecule can be synthesized using the set of chemical reactions. Although this approach might seem useful from a chemical point of view, it is not so appropriate for precise analysis. A different choice of synthesis rules may result in a different set of fragments. Besides that, rules may conflict or the derived fragments may overlap. Moreover, there are indications that actual synthesis may not be reflected very well (*e.g.* Vinkers *et al.*).[24] For a general overview of retro-synthesis, the reader is referred to a recent review by Todd.[25] Furthermore, a recent application of this synthetic approach was described by Vieth and Siegel.[26] The authors investigated four sets of bioactive molecules, fragmented these, and analyzed fragment distribution within a single set, and between the four sets. An interesting example is the distribution of the β-lactam framework within antibiotics. This framework was prevalent in the older marketed drugs and absent in new ones. This may reflect the problem of the developing resistance observed against older antibiotics. Another example is the absence of amino acid scaffolds and side chains in marketed oral drugs. Likewise, the majority of amino acid scaffolds is exclusive to injectable drugs.

2.3 Learning from Existing Databases

There is a lot to be learned from existing (drug) compound databases in terms of fragments: which fragments exist, how frequent they are, and how the occurrence of one fragment is related to the occurrence of another, non-overlapping fragment.[27] For instance, one can find single fragments that occur extremely often (*e.g.*, a phenyl ring), or chemical templates some drug classes are based on (*e.g.*, benzodiazepines). Fragments which have low abundance might indicate barely explored parts of chemical

space,[27] potentially interesting for designing new compounds. Insight can be obtained in preferences regarding chemistry as well as in differences among databases. In the next paragraphs, we will further expand on this, discussing analysis and evaluation of such databases (sections 2.3.1 and 2.3.2) and applications of the findings thereof (sections 2.3.3, 2.3.4 and 2.3.5).

2.3.1 Analysis of a Single Database

In an effort to identify the common features present in drug molecules, Bemis *et al*.[28] analyzed the structures of 5,120 drugs extracted from the Comprehensive Medicinal Chemistry database (CMC)[29]. Two types of representation were used, in order to analyze structures at different levels of detail. At a more general level, properties of the molecular graphs were analyzed. Since the same graph may represent multiple molecules of similar shape, the common structure classes are revealed. For example, benzene, hexane, and pyridine are all represented by the same hexagonal graph. In a more detailed analysis, the authors also considered atomic properties such as atom type, hybridization, and bond order. The authors defined four non-overlapping structural units that form a hierarchical description of the molecule: ring systems, linkers, frameworks, and side chains as discussed in section 2.2.3.1. The authors justified their choice of this classification scheme by highlighting its useful features. For example, most frequent frameworks are easily identified, which may guide future drug design. Moreover, ring systems and linkers can serve as input for combinatorial library generation. In addition, the simple building blocks in existing drugs are already useful to check the overlap between compound libraries.

The graph theoretical approach as outlined in section 2.2 and in Figure 2, identified a set of 1,179 different frameworks, of which the six-membered ring was the most common one found. Of all these frameworks, 783 (66%) were unique, *i.e.* they occurred only once in the database. However, a small set of only 32 frameworks accounted for 50% of the drug molecules in the database. Analysis that also considered atomic properties logically resulted in a more diverse set of frameworks.

There were 2,506 different frameworks, of which 1,908 (76%) were unique. Not surprisingly, a small set of 41 frameworks accounted for 1,235 drug molecules (24%) in the database. Benzene was the most common framework found (8.5%). When we think of molecules as a common framework decorated with side chains, phenyl and other small rings may be considered side chains just as well, as in peptides. In this study, however, they were not; the few rings present in a small molecule are needed to derive a reasonable framework. In a continuation to this work, Bemis *et al*. focused on the various side chains found in drugs.[30] Additional information was included in the side chain description, *i.e.* the connection point and type of framework atom that the side chain was bonded to. Side chains consisting of a single (heavy) atom other than hydrogen, *e.g.* chlorine, were also considered. The set of molecules extracted from the CMC database was slightly smaller now, 5,090 molecules in size. From this set, 4,689 had side chains. The total number of side chains was 18,664, on average four side chains per scaffold. The average length of a side chain was two atoms. Side chains of one heavy atom in length were found most (66%). Since oxygen atoms double-bonded to a ring system have a profound effect on the ring's electronic properties, it may be reasonable to consider these as part of the ring. In this case, the number of side chains was reduced to 57%.

Lameijer *et al.* explored the possibility of gaining new insights solely from the structures that exist in the database.[27] For this, the NCI database[14] was mined. The authors reasoned that the substructures and the combinations they occur in, provide insight into synthetic feasibility and "chemical habits". These habits emerge from an analysis of compound types that are made frequently or substructures that are often found together. The most frequently occurring fragments and fragment combinations were denoted as "chemical clichés". Graph splitting was used to break the molecules into parts suitable for mining. For this, the method described by Bemis *et al*.[28] was adopted, with the extension that frameworks were further split into ring systems and linkers. Another difference was that only side chains connected to a ring counted as

side chain. Side chains attached to a linker were part of the linker. Figure 7 shows an example of a molecule split into molecular parts according to Lameijer *et al.*[27]

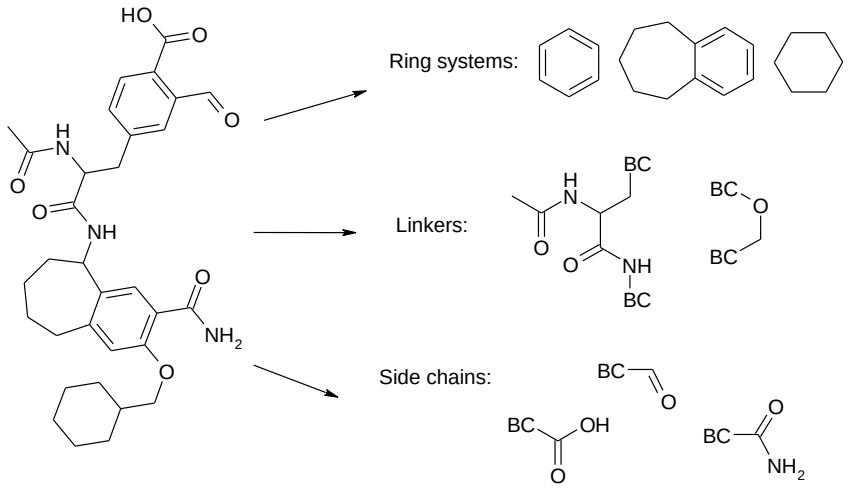


Figure 7. Example structure (see also Figure 1) split into ring systems, linkers, and side chains according to the algorithm of Lameijer *et al.*[27] In contrast to Figure 6, side chains in this figure stem from rings only. Side chains connected to a linker are considered part of the linker. Again, boxed 'B *& atom type*' labels are used to indicate a connection point to a ring.

By fragmenting 250,251 compounds from the NCI database, they found 65,612 fragments of the three different types of ring systems, side chains, and linkers. This already yielded useful information, for instance which ring systems occur, and which do not, *i.e.* finding an N_6-ring to be nonexistent may complement some chemical commonsense. In total, 13,509 ring systems were found, 18,015 side chains, 9,675 linkers with two ring systems, 2,531 linkers with three ring systems, and 2,280 linkers with four or more ring systems (up till 18 ring systems). In general, larger ring systems or branches occurred less frequently. Almost 70% of the three types of fragments occurred only once in the database. Branches with a higher number of attachment points seemed to have lower abundance. An exception to this rule was formed by linkers with six, or multiples of six, attachment points. These linkers occurred much

more frequent than their neighbors did. Inspection revealed these linkers were symmetrical.

The co-occurrence of fragments was also analyzed, to see whether the occurrence of one fragment in a molecule is related to the occurrence of another. This type of analysis can be compared to studying the contents of a shopping basket in a supermarket, a so-called Market Basket Analysis. Wine and olives may be frequently brought together as are beer and potato chips, where beer and olives might be rarely observed together. Market Basket Analysis is a data-mining tool for finding regularities in shopping behavior of customers of supermarkets, online shops, *etc.* A stochastic experiment was conducted first since for frequently occurring fragments the chance is higher that a relationship is found, even if there is none. A new "NCI" database was simulated using fragments that occurred in 20 or more molecules. Each fragment was used as many times as it occurred in molecules of the real NCI. Fragments were randomly divided over virtual molecules in the new database and each combination was counted. This process was repeated a thousand times, after which the expected occurrence of each fragment pair was calculated, together with the standard deviation of the occurrence. The expected occurrences were compared to actual co-occurrences in the NCI. A significant difference between the simulated/expected and the real co-occurrence implies that the fragments are correlated. Z-values were calculated and compared to detect that correlation.

Table 2. Some fragment pairs that occurred much more and much less often together than expected. The first row, consisting of the tetrahydrofuran and the $-CH_2OH$ group would be expected to occur 122 times together, but the pair appears in 2292 molecules leading to a multiple of 19 (see also text; data taken from Lameijer *et al.*[27]).

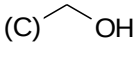
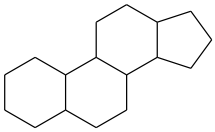
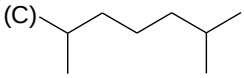
			Occurrence		
z-value	**Fragment 1**	**Fragment 2**	**Expected**	**Real**	**Multiple**
206			**122**	**2292**	**19**
117			**2.3**	**206**	**88**
-19			**544**	**139**	**0.26**
-67			**2653**	**270**	**0.10**

Table 2 presents some examples of fragment pairs that occur in the same molecule much more or much less frequently than expected. In the first row of Table 2 tetrahydrofuran and a CH_2OH group are together; they were expected to occur 122 times together, but do so much more frequently in 2292 molecules. This is 19 (2292/122) times more than expected, and very significantly different (z value of 206) from the simulated database. The explanation is that the combination is found in (substituted) nucleosides that have been tested for anti-tumor activity. The second row presents another example of frequently co-occurring fragments that present a single structure class, *viz*. dihydrocholesterol analogues.

Interestingly, the situation is opposite for the combination of a tetrahydrofuran and a phenyl group expected to occur in 2653 molecules. However, in the NCI there are only 270 of such instances, a factor of approx. 0.10 (270/2653). Apparently, this combination is underrepresented. A possible explanation for this effect might be that the 'avoiding' fragments belong to different compound classes with little overlap. Typical members from one class will be abundant in that class and scarce in others, adding to an overall reduction in co-occurrence frequency. Similarly, typical members from the same class are prone to be found together. Tetrahydrofuran-containing compounds generally differ in origin from phenyl-containing compounds. The tetrahydrofuran ring is often stemming from the ribose moiety of nucleosides, either natural or chemically modified, whereas the phenyl ring is often found in industrial chemicals.

The authors suggest that the derived fragment and co-occurrence lists are useful in creating new chemistry. For instance, these listings provide insight into the most popular and therefore most commonly used side chains and ring systems for synthesis. Rarer fragments also come forward through these lists, indicating less explored parts of chemical space. Finally, by looking at the fragments that do not occur together, new chemical space can be explored. The co-occurrences may be used to find a replacement for a structural feature. Examples of fragment pairs that are replacements of one another are chlorine and bromine, or naphthalene and benzene.[27] These fragment pairs rarely occur together,[27] possibly because of their comparable physicochemical properties.

2.3.2 Analysis of Multiple Databases

To facilitate the design of libraries for high throughput screening, Xue *et al*. extracted scaffolds and side chains and analyzed the distributions.[31] A "scaffold" was defined as a molecular fragment without side chains, essentially identical to the definition of frameworks (Figure 6). A "side chain" was defined as any acyclic chain or functional group with a single connection point to the rest of the molecule. As a source, the

authors used Optiverse (OV)[32], a combinatorial screening library designed for diversity, and the Maybridge collection (MB)[33], a library of compounds used in medicinal chemistry. Acyclic structures were removed prior to screening (1,214 from OV and 1,060 from MB). The remaining sets were 116,762 (OV) and 58,239 (MB) compounds in size. To isolate scaffolds and side chains, ring structures were detected first. Starting from these rings, all connected fragments were inspected. Acyclic fragments were removed from the structure and stored as side chains. The remaining structure was stored as a scaffold. Using this algorithm, the authors extracted 52,529 unique scaffolds and 4,486 side chains from OV, and 15,690 scaffolds and 2,851 side chains from MB. Only a minor overlap was observed: 2,945 scaffolds and 407 side chains occurred in both sets.

The ratios between the number of unique scaffolds and database size, suggest that on average one scaffold is found in 2.2 (OV) and 3.7 (MB) molecules, respectively. However, the authors observed an unequal distribution of scaffolds: 8% (OV) and 7% (MB) of scaffolds occurred in 50% of the molecules. Moreover, more than 90% of the scaffolds occurred only once or twice. Aromatic structures and heterocycles were found most. The distribution of side chains was similarly imbalanced. The ten most frequent side chains accounted for almost 75% occurrences, whereas the majority occurred only once. Among the top-ten were classic substitutions as halogens, the nitro group, the hydroxy group, and organic functional groups such as the methoxy group. The methyl group accounted for 25% (OV) and 20% (MB) of occurrences, respectively.

Xu[34] derived molecular scaffolds to evaluate chemical compound libraries in terms of diversity, distribution in chemical space, and differences/similarities with respect to existing drugs. The author used a Scaffold-based Classification Approach (SCA) that groups compounds into the same class if they share the same topological scaffold or so-called class center. The rationale behind this approach was that medicinal chemists intuitively group compounds based on scaffolds and functional groups, and not so much on structural descriptors that most classification algorithms use. Scaffolds were

derived similar to Xue *et al*.[31] and Bemis *et al.*[28] However, unsaturated bonds connected to a ring were considered part of the scaffold, since they change the chemical behavior of the ring system. Normally, scaffold analysis overlooks aliphatic compounds, since scaffolds are defined to consist of at least one ring. To overcome this, an extended definition of scaffold was adopted that also covered the aliphatic compounds. Double and triple bonds of acyclic compounds were treated as ring bonds, so part of the scaffold. For saturated acyclic compounds, the scaffold consisted of the heteroatoms and carbon atoms that connect them. In all other cases, the carbon backbone formed the scaffold. Although the purpose of this extended definition is to extract scaffolds from all possible compound classes, some compounds from the same class may appear unrelated. For instance, amino acids that possess a cyclic side chain are separated from those with an aliphatic chain. The structural scaffold derived will be the ring system in the first case and the characteristic amino/carboxyl group core in the second case.

First, a list of unique scaffolds was derived and sorted by complexity. The complexity was calculated from four structural descriptors, namely number of rings in the smallest set of smallest rings, number of heavy atoms, number of bonds, and the sum of heavy atomic numbers in the scaffold. Each scaffold, or class center, in the list was assigned an ID that corresponded to its position in the list. How much a molecule resembled its class center was determined by the amount of side chains attached to the scaffold. Fewer side chains will give a closer resemblance to the class center. The similarity of a drug with the class center was reflected in the *membership* value. The *membership* value was based on the sum of heavy atomic numbers, the number of rotating bonds, the number of one and two nodes, and the number of double and triple bonds in a molecule compared to its scaffold. Since the membership value indicated the contribution of rings in the class center for a certain molecule, this term was called *cyclicity*. The four databases ACD[35], NCI[14], CMC[29], and MDDR[36], were analyzed according to this scaffold-based classification approach. Only the orally available drugs of CMC and MDDR were used. A diversity map was constructed that mapped

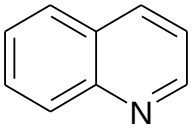
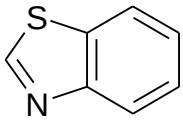
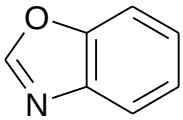
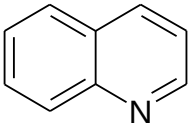
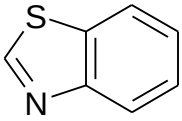
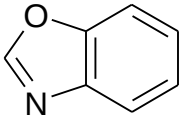
complexity values against cyclicity values for each compound. Libraries that are more diverse have a wider spread on this map. An interesting outcome was the ranking of the four libraries according to chemical diversity. The ACD was most diverse, followed by the NCI, then the CMC, and finally the MDDR. Two factors contribute to the low diversity of the MDDR: the majority of compounds are analogs, and all compounds comply with the 'drug-likeness' property values. Molecules contributing to the high diversity of the ACD included RNAs/DNAs and fullerene C60. Another interesting finding was that the orally active drugs from the CMC and MDDR were distributed in a narrower region than the other libraries.

2.3.3 Biological Activity

Sheridan[22] used common substructures to find fragment replacements in (drug-like) molecules. For this, 98,445 drug-like molecules from the MDL Drug Data Report (MDDR)[36] database were clustered according to similar biological activity, resulting in 556 clusters. Compounds from the same cluster were compared to find the "highest-scoring common substructure" (HSCS).[21] Only compounds with an HSCS significantly larger than two randomly selected molecules of the same size were used to extract the fragments pairs that differed. Two different methods were used to extract replacement fragment pairs. The first method used atom-wise comparison of fragments, *i.e.* based on element and hybridization of atoms. The second method also considered possible rings the atoms were in and adjacent functional groups, such as -NO_2, -CO, -SO_2, or -PO_3. Many of the classical replacements in medicinal chemistry were found.[22] With atom type, substitution of C with N in an aromatic ring (*e.g.* phenyl *vs*. pyridine) was the most common. The next most common was replacement of -O- with -S- in both rings and chains, followed by -N- with -O- in rings, chains, and esters *vs*. amides. Another interesting commonly found replacement was the change between a five- a six-membered ring. Also considering the context of atoms in the comparison, *e.g.* a ring or functional group yielded a qualitatively similar fragment list. For a more complete list of replacements, the reader is referred to Sheridan[22].

In a subsequent study, Sheridan *et al.* utilized the HSCS to identify fragments[23] that are associated with multiple biological activities. The authors considered activity in the widest sense, ranging from *in vivo* biological effects (*e.g.* anti-hypertensive) to *in vitro* measures (*e.g.* affinity for a receptor). Since high specificity is very much desired for new drugs, knowledge about multi-activity fragments may be useful to avoid chemical classes likely to have unwanted side effects. On the other hand, scaffolds that are active on a variety of receptors may form an attractive starting point in combinatorial library design. Pairs of molecules with similar structure and dissimilar activity were identified first. For each pair, the highest scoring common substructure (HSCS) was derived.[21] Again, only those HSCS's were kept that were significantly larger than would be expected for two randomly selected molecules. A "consensus substructure" was generated from each molecule and its HSCS. It consists of atoms that are considered to be "conserved", *i.e.* atoms that appeared relatively often in the set of HSCS's for that molecule. The most interesting consensus substructures are those that are found in many molecules and have many unique activities. Therefore, the generated consensus substructures were ranked according to both frequency of occurrence and number of unique activities. In case of structurally similar consensus substructures, only the highest in rank was kept. The steroid skeleton was found as a fine example of a multi-activity structure due to the many physiological processes steroid hormones are involved in. Similarly highly ranked were tricyclic structures as in imipramine and doxepine. They bind to many G protein-coupled receptors and transport proteins.

Table 3. Pairwise comparison of bicyclic rings (taken from Kho *et al.*[37]). The number is the logarithm of the odds ratio, and indicates the preference in terms of mutagenic potential of one ring system relative to the other. For instance, a value of -1.0182 (second row, second column from the right) means that the left ring system has higher odds of being found in Ames positive compounds, so the top ring system is preferred. The arrow points to the fragment that is more likely to be found in the Ames-negative class. Many more ring systems were considered, indicated by the (empty) third column.

		...		
	0.000	...	↑ -1.0182	↑ -3.0331
...			...	...
 Benzothiazole			0.000	↑ -2.0149
 Benzoxazole				0.000

2.3.4 Predictive Models

In an attempt to organize available data in mutagenicity databases, Kho *et al.* described an automated approach to extract and organize ring systems occurring in a mutagenicity dataset.[37] The authors suggested this method can be applied to any other set of molecules classified by some property, *e.g.* biological activity. A common assay for mutagenicity prediction is the Ames test, in which Ames-positive compounds are suspected to have mutagenic characteristics, whereas Ames-negatives are not. The database[38] was searched for the occurrence of ring types and their frequency in the Ames-positive and -negative categories. Emphasis was not so much on the development of predictive algorithms, but more on organizing the available data for use by chemists. Simple scaffolds were identified using a program that finds scaffolds by comparing all molecules in a set.[39] The results were presented as a hierarchy according to complexity. In this approach, simple rings are placed at the highest level and more complex ring systems that contain the parent rings, as descendants. An example hierarchy is presented in Figure 8. Note that the tetrahydronaphtalene branch (first child), having equal odds of being found in either set, leads to an Ames-positive and an Ames-negative scaffold. A selection of the bicyclic rings found is presented in Table 3. Such a two-way entry table may be useful for selection of (bio)isosteric replacements with higher odds in the Ames-negative set. Similar tables can be constructed for other properties. A general finding from these data was that an increase in aromaticity or extension of conjugation enhances the odds for mutagenic compounds. An increase in the aliphatic character of rings decreases the mutagenic potential. To evaluate the usefulness of the mutagenicity dataset (with a total of 6,039 compounds), the authors compiled a reference dataset consisting of 3,882 commercially available drugs. Analysis revealed that the chemical diversity within the mutagenicity dataset was significantly less than the diversity of the marketed drugs. For the smaller drug set, 750 ring systems were found in contrast to the 427 ring systems found in the Ames-test dataset. The two sets had 199 ring systems in common.

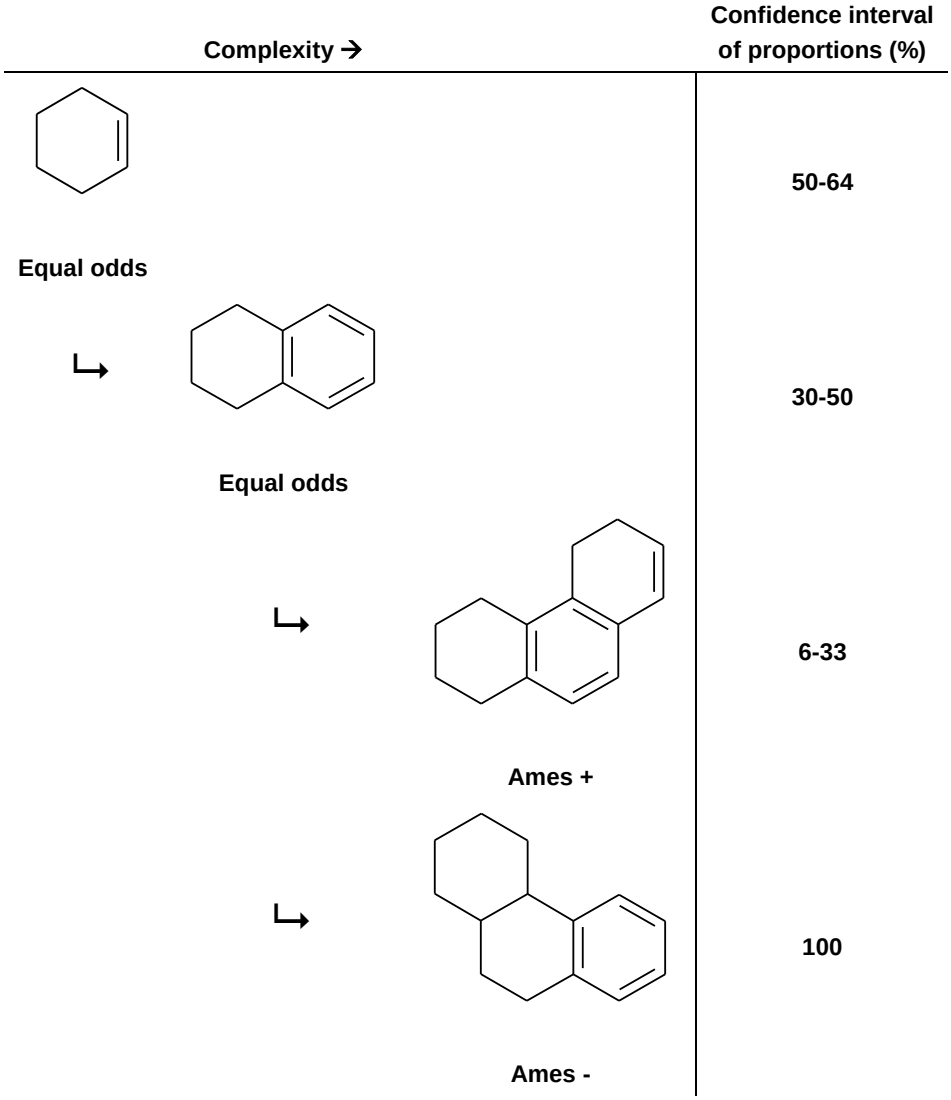


Figure 8. Scaffold (cyclohexene) hierarchy derived from mutagenicity data[38]. The proportion of Ames negative to Ames positive counts is qualitatively indicated below each scaffold. The confidence interval of the proportions is shown on the right of the scaffolds. Data taken from Kho *et al.*[37].

Instead of studying a limited set of structural features such as ring systems,[37] others have taken a more exhaustive approach. In such a scenario, all possible fragments are examined to find those discriminative for a certain property, *e.g.* toxicity. Kazius *et al.* used frequent fragment mining in order to derive toxicophores.[40] Similar to Kho *et al.*[37], structural elements were arranged according to mutagenic potential, thereby forming a decision list. Most substructure mining methods use only part of the chemical information in a molecule, *viz.* connectivity of the molecular graph (Figure 2), atom type labels, and bond order (sometimes including aromaticity). To increase the level of chemical detail that is considered, Kazius *et al*. used an extended chemical representation.[40] Figure 9 shows a typical compound in standard chemical notation and two types of elaborate chemical representation. Elaborate chemical representation uses atomic hierarchies in addition to atom type labels, thereby including both general and more specific information. Atomic hierarchies are tree-like structures that consist of a root of a general atom label representing an atomic property, and branches of more atom-specific labels (specifiers). Aliphatic nitrogen and oxygen atoms were labeled as “small hetero atom” with specifiers for the atom type and number of connected hydrogens, as shown in Figure 9. Aliphatic sulfur and phosphorus atoms were labeled “large hetero atom” with an additional specifier for the atom type. Chlorine, bromine, and iodine atoms were labeled “halogen” with atom type specifiers (Figure 9). For rings, two types of elaborate chemical representation were used. The “aromatic” setting used a special atom label and bond type to represent aromatic atoms and bonds, and attached a type specifier to aromatic heteroatom. Examples of aromatic atoms and bonds are shown in chemical representation **I** in Figure 9. The “planar” setting used a special atom label and bond type for atoms and bonds in aliphatic five- and six-membered rings or aromatic rings, including atom type specifiers. Planar atoms and bonds are shown in chemical representation **II** in Figure 9. All other atoms were labeled with the atom type. An additional atom specifier for the atom type was connected to heteroatoms and halogens, and a specifier for implicit hydrogens was connected to heteroatom. Standard and elaborate chemical representations were used to extract substructures

from mutagenicity data, both with and without considering non-linear fragments. The dataset consisted of 4,069 compounds from the Chemical Carcinogenesis Research Information System database[41]. Compounds were categorized as non-mutagens if all mutagenicity tests had a negative outcome. This resulted in 2,294 compounds classified as mutagens.

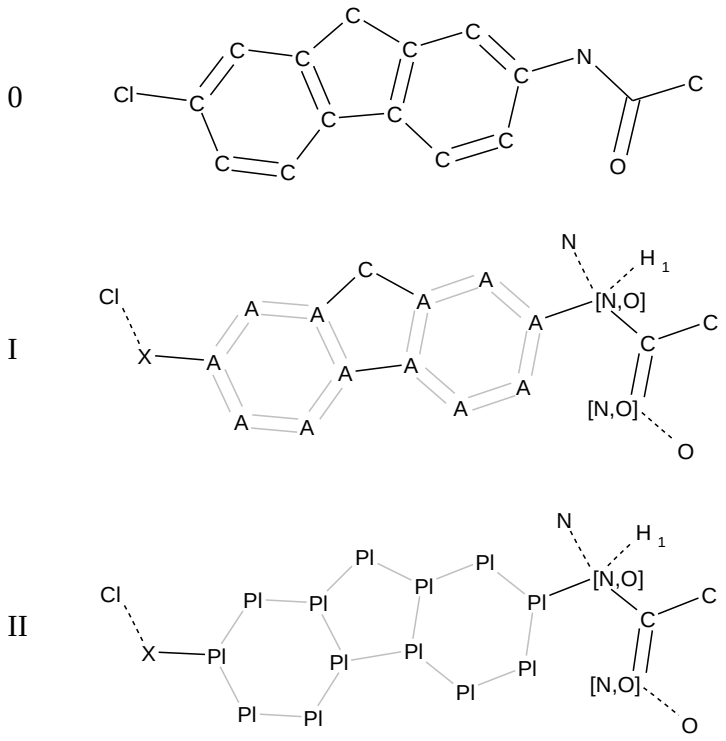


Figure 9. A typical compound (PubChem[1] CID78776) in standard chemical notation (**0**), and two types of elaborate chemical representation, *viz.* the "aromatic" setting (**I**) and the "planar" setting (**II**). Bonds are either single, double, aromatic (grey double bonds in **I**) or planar (grey single bonds in **II**). Additional information is attached using the dashed bonds. Atom labels are carbon (C), nitrogen (N), oxygen (O), small heteroatom ([N,O]), halogen (X), chorine (Cl), aromatic atom (A), planar atom (Pl), and number of implicit hydrogens (H_1).

Fragments from all methods were used together to find nonredundant substructures that are discriminative for mutagenicity. Only those substructures that occurred in more than 70 mutagens were considered. A decision list was constructed (Figure 10) by using the fragment with lowest p-value to split the set into two subsets (one that contained the fragment and one that did not). The p-value of a fragment was defined as the probability to find a statistical association with mutagenicity based on chance alone. It was calculated from the amount of mutagens versus non-mutagens that are detected using that fragment. For the subset that did not contain the fragment, p-values were recomputed and the next most mutagenic fragment was used to split this set. In case of multiple fragments with the lowest p-value, the largest fragment was used. The process was repeated as long as the new set had more than 60% mutagenic compounds. If the best-selected fragment had a p-value of more than 10^{-20}, no further splits were made. From all methods, the use of elaborate chemical representation combined with detection of nonlinear fragments proved best: mutagens were detected with a sensitivity of 84%. The resulting decision list (Figure 10) consisted of six non-redundant discriminating substructures, starting with a polycyclic planar system that described at least three rings, and consisted of 11 planar atoms connected by planar bonds. The next most discriminating fragment was a nitrogen atom double-bonded to a nitrogen or oxygen, followed by a 3-membered heterocycle (aliphatic epoxides and aziridines), and then an aliphatic halogen (chlorine, bromine, and iodine). The second-last fragment was an aromatic primary amine and the list ended with a heteroatom-bonded to a heteroatom fragment. Some of these substructures proved to be very similar to the general toxicophores derived previously by the authors in a laborious approach.[42] These results emphasize the benefit of elaborate chemical representation. For instance, the most discriminative fragment for mutagenicity would not have been detected by other methods, since the planar atom notation proved essential. Moreover, the importance of wildcards is underlined by their presence in all six substructures. Since the list contained two branched and one cyclic substructure, all possible graphs must be considered in substructure mining.

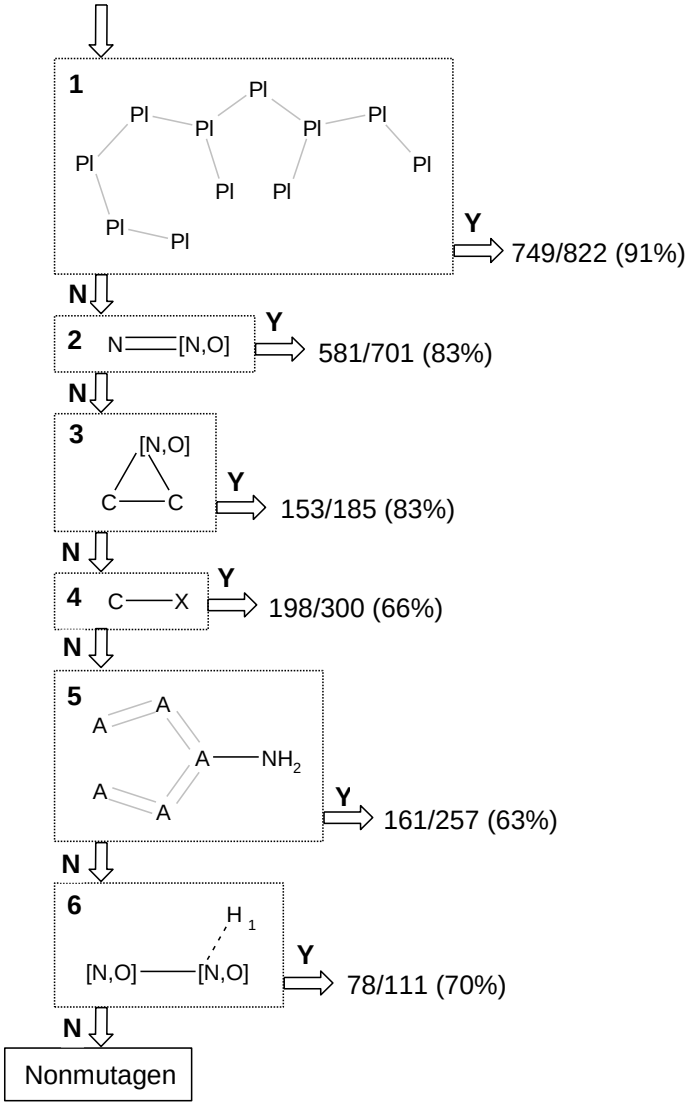


Figure 10. Decision list derived from mutagenicity data.[40] Arrows indicate the direction to follow if a substructure is (**Y**) or is not (**N**) present in a compound. The number of mutagens, the total number of compounds, and the percentage of mutagens is indicated for each subset (right).

2.3.5 Ligand Design

The ring-linker frameworks approach described by Bemis *et al*.[28] was used to design new scaffold classes based on experimental structural information, and to guide the optimization of modestly active ligands.[43] A set of 119 kinase inhibitors for at least 18 different targets, was fragmented into ring systems and linkers, and frequencies of occurrence were analyzed. Since bi- and tricyclic ring systems were relatively rare in the fragmented set, only monocyclic rings were considered. The authors found that the four rings benzene, pyridine, pyrimidine, and pyrrole, comprise almost 90% of monocyclic ring occurrences in the fragmented data set. In addition, eight of the most abundant linkers were responsible for 90% of all linker occurrences in the set. From the four rings and eight linkers, a virtual library of kinase inhibitor scaffolds was constructed. Fragments known to form a critical interaction with the binding site of a kinase, served as a starting anchor. New scaffolds were generated by linking one of the rings to the anchor fragment, using one of the linkers. This was repeated for all ring-linker combinations, and for each attachment point on the rings and anchor fragment. The newly designed scaffolds were docked against their targets, using the placement of the anchor fragment as constraint. A fit-based score was calculated, and the highest scoring scaffolds were clustered according to the connection point at the anchor fragment. Using this method, the authors were able to reproduce the predominant structural motifs for known kinase inhibitors. In addition, they were able to suggest a number of alternative variations for these ligand cores.

Lameijer developed a software tool to design drug-like molecules, the “Molecule Evoluator”.[44] In this tool both atom- and fragment-based evolutionary approaches were implemented. Fragments were taken from the analysis of the NCI database (ref. 27 and reviewed in section 2.3.1). Through *interactive* evolution, a new principle in which the user acts as a fitness function, the authors suggested a number of simple yet novel molecules, eight of which were subsequently synthesized. Four compounds showed affinity for biogenic amine targets (receptor, ion channel, and transport protein).[45]

2.4 Conclusion

In this review, we have compiled a number of computational strategies to dissect molecules into sets of constituting atoms, leading to fragments of different nature. Such fragments may also consist of elaborate atom representations, including wild cards. The reason for doing these, often computationally intensive, operations is found in the wealth of information that can be gleaned from such analyses. Virtual and real-world compound libraries can be mined for their diversity and/or similarity. In addition, the 'synthetic habits' of medicinal chemists can be explored. Furthermore, occurrence and co-occurrence of fragments may suggest new directions into chemical space. Fragments that appear linked to side effects, via either multiple activities or straight toxicity, have been identified. This may help the medicinal chemist in designing safer or more selective lead compounds. Conversely, desired activities can be linked to fragments, and such information may be a decisive factor in a successful medicinal chemistry program. With both the large number of HTS campaigns being performed and the resulting data increasingly being made available in the public domain, it is anticipated that steadily more dedicated datasets will become available for fragment mining. Rule- and knowledge-based design efforts will certainly benefit from this.

2.5 References

(1) PubChem database, available at: pubchem.ncbi.nlm.nih.gov.

(2) eMolecules, www.emolecules.com.

(3) Oprea, T. I.; Blaney, J. M. Cheminformatics Approaches to Fragment-based Lead Discovery. In *Fragment-based Approaches in Drug Discovery*; Jahnke, W.; Erlanson, D. A., Eds.; Methods and Principles in Medicinal Chemistry 34; Wiley-VCH: Weinheim, Germany, 2006.

(4) Cramer, R. D.; Jilek, R. J.; Andrews, K. M. Dbtop: topomer similarity searching of conventional structure databases. *J. Mol. Graph. Model.* **2002**, *20*, 447-462.

(5) Hansen, P. J.; Jurs, P. C. Chemical Applications of Graph Theory. *J. Chem. Ed.* **1988**, *65*, 574-580.

(6) Agrawal, R.; Imielinski, T.; Swami, A. Mining Association Rules between Sets of Items in Large Databases. In *Proceedings of the International Conference on Management of Data*; ACM Press: New York, NY, USA, 1993.

(7) Borgelt, C.; Berthold, M. R. Mining Molecular Fragments: Finding Relevant Substructures of Molecules. In *Proceedings of the International Conference on Data Mining (ICDM)*, 2002; pp 51-58.

(8) Yan, X.; Han, J. gSpan: Graph-Based Substructure Pattern Mining. In *Proceedings of the International Conference on Data Mining (ICDM)*; Maebashi City, Japan, 2002.

(9) Huang, J.; Wang, W.; Prins, J. Efficient mining of frequent subgraphs in the presence of isomorphism. In *Proceedings of the 3th IEEE Intl. Conf. on Data Mining (ICDM)*, Piscataway, NJ, USA; IEEE Press, 2004.

(10) Nijssen, S.; Kok, J. N. A quickstart in frequent structure mining can make a difference. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*; Kohavi, R.; Gehrke, J.; DuMouchel, W.; Ghosh, J., Eds.; Conference on Knowledge Discovery in Data 2004; ACM Press: New York, 2004.

(11) Wörlein, M.; Meinl, T.; Fischer, I.; Philippsen, M. A quantitative comparison of the subgraph miners MoFa, gSpan, FFSM, and Gaston. In *Proceedings of the 3th European Conference of Principles of Knowledge Discovery and Data Mining (PKDD)*, 2005.

(12) Institute of Scientific Information, Inc. (ISI): Index chemicus – subset from 1993.

(13) HIV assay, see: dtp.nci.nih.gov/docs/aids/aids_data.html.

(14) National Cancer Institute database (NCI 3D), available for download at: dtp.nci.nih.gov (Developmental Therapeutics Program NCI/NIH).

(15) Hofer, H.; Borgelt, C.; Berthold, M. R. Large Scale Mining of Molecular Fragments with Wildcards. *Intelligent Data Analysis* **2004**, *8*, 495-504.

(16) Meinl, T.; Borgelt, C.; Berthold, M. R. Mining fragments with fuzzy chains in molecular databases. In *Proceedings of the Workshop W7 on Mining Graphs, Trees, and Sequences (MGTS '04)*; Kok, J. N.; Washio, T., Eds.; Pisa, Italy, 2004.

(17) Meinl, T.; Wörlein, M.; Urzova, O.; Fischer, I.; Philippsen, M.; The ParMol package for frequent subgraph mining. In *Proceedings of the 3th International Workshop on Graph Based Tools*; Margaria-Steffen, T.; Padberg, J.; Taentzer, G., Eds.; Electronic Communications of EASST 1, European Association of Software Science and Technology, Berlin, 2006; ParMol is available for download at: http://www2.informatik.uni-erlangen.de/Forschung/Projekte/ParMol/?language=en.

(18) Helma, C.; Kramer, S.; De Raedt, L. The Molecular Feature Miner MOLFEA. In *Proceedings of the Beilstein-Institut Workshop*; Hicks, M. G.; Kettner, C., Eds.; Molecular Informatics: Confronting Complexity, Logos Verlag, Berlin, 2003.

(19) King, R. D.; Srinivasan, A.; Dehaspe, L. Warmr: a data mining tool for chemical data. *J. Comput.-Aided Mol. Des.* **2001**, *15*, 173-181.

(20) Dehaspe, L.; Toivonen, H. Discovery of frequent DATALOG patterns. *Data Min. Knowl. Discov.* **1999**, *3*, 7-36.

(21) Sheridan, R. P.; Miller, M. D. A method for visualizing recurrent topological substructures in sets of active molecules. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 915-924.

(22) Sheridan, R. P. The most common chemical replacements in drug-like compounds. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 103-108.

(23) Sheridan, R. P. Finding multiactivity substructures by mining databases of drug-like compounds. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1037-1059.

(24) Vinkers, M.; De Jonge, M.; Daeyaert, F.; Heeres, J.; Koymans, L.; Van Lenthe, J.; Lewi, P.; Timmerman, H.; Van Aken, K.; Janssen, P. SYNOPSIS: SYNthesize and OPtimize System in Silico. *J. Med. Chem.* **1998**, *46*, 2765-2773.

(25) Todd, M. H. Computer-aided organic synthesis. *Chem. Soc. Rev.* **2005**, *34*, 247-266.

(26) Vieth, M.; Siegel, M. Structural Fragments in Marketed Oral Drugs. In *Fragment-based Approaches in Drug Discovery*; Jahnke, W.; Erlanson, D. A., Eds.; Methods and Principles in Medicinal Chemistry 34; Wiley-VCH, Weinheim: Germany, 2006.

(27) Lameijer, E.-W.; Kok, J. N.; Bäck, T.; IJzerman, A. P. Mining a Chemical Database for Fragment Co-occurrence: Discovery of "Chemical Clichés". *J. Chem. Inf. Model.* **2006**, *46*, 553-562.

(28) Bemis, G. W.; Murcko, M. A. The Properties of Known Drugs. 1. Molecular Frameworks. *J. Med. Chem.* **1996**, *39*, 2887-2893.

(29) Comprehensive Medicinal Chemistry (CMC-3D) Release 94.1 is available from MDL Information Systems Inc., San Leandro, CA, U.S.A.

(30) Bemis, G. W.; Murcko, M. A. Properties of Known Drugs. 2. Side Chains. *J. Med. Chem.* **1999**, *42*, 5095-5099.

(31) Xue, L.; Bajorath, J. Distribution of Molecular Scaffolds and R-Groups Isolated from Large Compound Databases. *J. Mol. Model.* **1999**, *5*, 97-102.

(32) Garr, C. D.; Peterson, J. R.; Schultz, L.; Oliver, A. R.; Underiner, T. L.; Cramer, R. D.; Ferguson, A. M.; Lawless, M. S.; Patterson, D. E. Solution Phase Synthesis of Chemical Libraries for Lead Discovery. *J. Biomol. Screen.* **1996**, *1*, 179-186.

(33) Maybridge is available from Maybridge Chemical Company LTD, Trevillet, Cornwall, UK.
(34) Xu, J. A new approach to finding natural chemical substructure classes. *J. Med. Chem.* **2002**, *45*, 5311-5320.
(35) Available Chemicals Directory, available from MDL Information Systems Inc., San Leandro, CA, U.S.A.
(36) MDL Drug Data Report, version 99.1 is available from MDL Information Systems Inc., San Leandro, CA, U.S.A.
(37) Kho, R.; Hodges, J. A.; Hansen, M. R.; Villar, H. O. Ring Systems in Mutagenicity Databases. *J. Med. Chem.* **2005**, *48*, 6671-6678.
(38) Ames-test data set, available at: www.altoris.com.
(39) SARvisonPlus 1.5, available from ChemApps, San Diego, CA, (www.chemapps.com).
(40) Kazius, J.; Nijssen, S.; Kok, J.; Bäck, T.; IJzerman, A. P. Substructure Mining Using Elaborate Chemical Representation. *J. Chem. Inf. Model.* **2006**, *46*, 597-605.
(41) Chemical Carcinogenesis Research Information System, available through TOXNET at: http://toxnet.nlm.nih.gov.
(42) Kazius, J.; McGuire, R.; Bursi, R. Derivation and validation of toxicophores for mutagenicity prediction. *J. Med. Chem.* **2005**, *48*, 312-320.
(43) Aronov, A. M.; Bemis, G. W. A minimalist approach to fragment-based ligand design using common rings and linkers: application to kinase inhibitors. *Proteins* **2004**, *57*, 36-50.
(44) Lameijer, E.-W.; Kok, J. N.; Bäck, T.; IJzerman, A. P. The Molecule Evoluator. An Interactive Evolutionary Algorithm for the Design of Drug-Like Molecules. *J. Chem. Inf. Model.* **2006**, *46*, 545-552.
(45) Lameijer, E.-W.; Tromp, R. A.; Spanjersberg, R. F.; Brussee, J.; IJzerman, A. P. Designing Active Template Molecules by Combining Computational de Novo Design and Human Chemist's Expertise. *J. Med. Chem.* **2007**, *50*, 1925-1932.