

## Mining a Chemical Database for Fragment Co-occurrence: Discovery of “Chemical Clichés”

Eric-Wubbo Lameijer,<sup>†</sup> Joost N. Kok,<sup>‡</sup> Thomas Bäck,<sup>‡,#</sup> and Ad P. IJzerman<sup>\*,†</sup>

Division of Medicinal Chemistry, Leiden/Amsterdam Center for Drug Research, Leiden University, Einsteinweg 55, 2300 RA Leiden, The Netherlands, Leiden Institute of Advanced Computer Science (LIACS), Leiden University, Niels Bohrweg 1, 2333 CA Leiden, The Netherlands, and NuTech Solutions, Martin-Schmeisser-Weg 15, 44227 Dortmund, Germany

Received September 2, 2005

Nowadays millions of different compounds are known, their structures stored in electronic databases. Analysis of these data could yield valuable insights into the laws of chemistry and the habits of chemists. We have therefore explored the public database of the National Cancer Institute (>250 000 compounds) by pattern searching. We split the molecules of this database into fragments to find out which fragments exist, how frequent they are, and whether the occurrence of one fragment in a molecule is related to the occurrence of another, nonoverlapping fragment. It turns out that some fragments and combinations of fragments are so frequent that they can be called “chemical clichés”. We believe that the fragment data can give insight into the chemical space explored so far by synthesis. The lists of fragments and their (co-)occurrences can help create novel chemical compounds by (i) systematically listing the most popular and therefore most easily used substituents and ring systems for synthesizing new compounds, (ii) being an easily accessible repository for rarer fragments suitable for lead compound optimization, and (iii) pointing out some of the yet unexplored parts of chemical space.

### INTRODUCTION

Over the last two centuries, chemists have synthesized many millions of structures. Two of the largest chemical databases, the Beilstein<sup>1</sup> and the CAS,<sup>2</sup> contain over 8 million and 25 million compounds, respectively. Such large collections of compounds could give much insight into chemistry, both into what kinds of compounds can be made with current chemical technology and into which parts of chemical space have been extensively investigated or, conversely, barely explored.

In most investigations so far, databases have been used for classification of compounds: for example, when is a compound druglike.<sup>3–5</sup> Or which substructures or properties correlate with mutagenicity?<sup>6,7</sup>

Relatively unexplored are the possibilities of general databases: collections of molecules for which no data other than their structure are available. The only investigations known to us which had this purpose did so to extract a catalog of rings<sup>8</sup> respectively substituents<sup>9</sup> for drug designers. However, it seems desirable to get more information out of those general databases than just rings. We could also learn much about chemistry and the habits of chemists by studying which substructures and substructure combinations occur.

In this investigation we want to delve deeper into the knowledge stored in the molecular structures. This would give us not only an extensive catalog of fragments to reuse

in the synthesis of drugs and other compounds but also insight into “chemical habits”. What kinds of compounds are made frequently, and which substructures are relatively rarely found together in a molecule? Some of these rare combinations might indicate barely explored parts of chemical space, potentially interesting for designing new compounds.

In this work we will use the name “chemical clichés” for some of the most-occurring fragments and frequently co-occurring pairs of fragments. The word “cliché” originated in the French printing industry where it denoted a stereotype, a kind of stamp of, for example, a picture that was pressed on the paper to produce the same image many times. Nowadays the word cliché is mainly used to denote a trite expression, such as “missed by a mile” or “top research institute”. However, some classes of chemical compounds also seem to be based on the same “stamp” with only slight variations, such as benzodiazepines and tricyclic antidepressants. Also, single fragments such as the benzene ring can occur extremely often in molecules. We think that the word “cliché” is useful to describe this reuse of ideas, while stressing that the current templates might not be the only ones that are viable.

Then the question remains how to extract knowledge from chemical databases. Knowledge is usually found in occurrences and patterns: what occurs in nature, what does not occur, and which events occur together? What is correlated with what? Looking at the molecular structure as a whole is not very useful, since all chemical structures in a database are unique, so on that level they are incomparable. Splitting the molecules into the chemically smallest fragments, the atoms, will yield no more information than the periodic table. One should therefore look for chemical knowledge at a level

\* Corresponding author phone: +31-(0)71-5274651; e-mail: ijzerman@chem.leidenuniv.nl.

<sup>†</sup> Leiden/Amsterdam Center for Drug Research, Leiden University.

<sup>‡</sup> Leiden Institute of Advanced Computer Science (LIACS), Leiden University.

<sup>#</sup> NuTech Solutions.

between these extremes, and that is the level of the molecular fragments.

In this paper we first discuss our choice of fragmentation method. Then we will describe the method we used to detect whether two different fragments co-occur more or less often than expected, and thereafter we will present the results of the fragment mining and co-occurrence analysis. We will conclude with a discussion of our findings, suggestions for application of the data obtained, and directions for future investigations.

## DATABASE

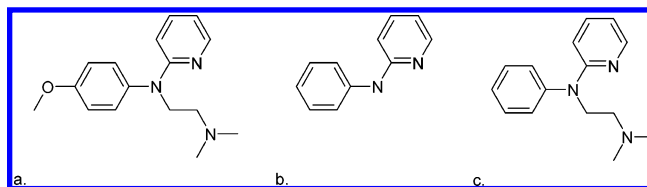
In this investigation we used the public database of the United States National Cancer Institute (NCI). The August 2000 version, which we mined, contains 250 251 structures.<sup>10</sup> The molecules in this database have been selected to be tested against cancer, so were deemed by the database compilers to possibly have biological activity. Since many of the compounds are experimental and have not been tested on bioavailability and safety in humans, the diversity in structures is quite large and should give a decent cross section of the range and preferences of chemical synthesis.

## FRAGMENTATION METHOD

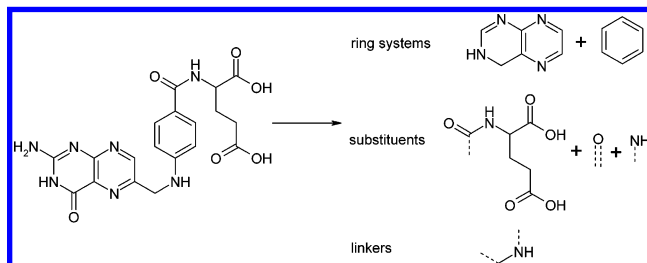
To find patterns in structures, the first step is to break the molecules into fragments. Two categories of fragmentation methods can be distinguished: the “full substructure set”, in which all possible 1, 2, 3 ...*n*-atom sized substructures of a molecule are detected, and “molecule parts”, in which a molecule is divided into a number of nonoverlapping substructures.

While the full substructure set would give all information possible, in practice it yields huge numbers of substructures per molecule (several thousands for even a medium-sized molecule). This makes such kind of data mining computationally very expensive, especially for large collections of compounds. For an exploratory study such as this, a “molecule parts” method would be more suitable, since there are fewer parts, they do not overlap, and they correspond to chemically intuitive units.

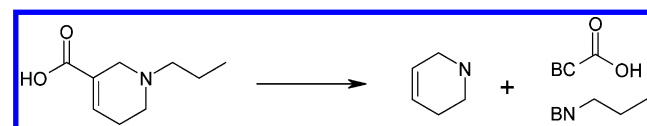
The next question is at which point to “break” the molecules. The two main methods here are graph splitting<sup>11–13</sup> and virtual retrosynthesis.<sup>14–16</sup> Graph splitting breaks molecules at topologically interesting points, such as the bond between a substituent and a ring, while virtual retrosynthesis uses specific rules based on chemical reactions and breaks, for example, ester bonds. Both methods yield manageable sets of substructures (10 000–100 000 for a medium-sized database). From a chemical point of view, the retrosynthesis method seems most logical; however, it does not reflect actual syntheses very well as, for example, Vinkers et al. found out.<sup>14</sup> The reason is that chemical reactivity depends on steric and electronic factors which are for a large part determined outside the three or four atoms of the “breakable bond” and its neighbors. Conversely, synthesis can often create bonds (such as alkane C–C bonds) which are not considered to be cleavable by most retrosynthesis algorithms, because typically only a few dozens of the hundreds of organic reactions are incorporated into the software. A last disadvantage is that different retrosynthetic rules give different fragment sets. In contrast, graph splitting is quite



**Figure 1.** (a) A drug molecule (pyrilamine). (b) Framework of pyrilamine according to Bemis et al.,<sup>11</sup> consisting of only the ring systems and the atoms that directly connect the ring systems. (c) Framework of pyrilamine according to our definition so without substituents attached to the rings. This framework is later split into ring systems and linkers.



**Figure 2.** Our algorithm breaks the bonds between ring systems and the rest of the molecule and thereby splits the molecule (in this example folic acid) into several types of fragments: ring systems, substituents, and linkers.



**Figure 3.** Storage format of ring systems and nonring systems. While ring systems are stored as normal molecules, substituents and linkers include one or more “branching atoms” that encode the symbols of the atoms to which the substituent or linker is attached.

reproducible, easy to implement, and divides structures in chemically intuitive units of “ring systems” and substituents. This is why we chose the graph splitting method.

Deciding to do graph splitting is however not enough, since graph splitting can be done in a number of different ways. Bemis et al.<sup>11</sup> iteratively cut off all 1-connected atoms, so that only “substituents” and frameworks were left, the frameworks being the ring systems with the part of the linkers that directly connect them, see Figure 1. We decided to go one step further and also split up the frameworks into ring systems and linkers. We therefore ended up with splitting molecules into substituents, ring systems, and linkers of different orders (linking two ring systems, three ring systems, etc.) See Figure 2 for an illustration of our fragment classes and the decomposition of an example molecule, folic acid.

The ring structures were stored as normal molecules, only without hydrogen atoms (similar to the format of the NCI database itself). For the substituents and linkers we considered it useful, like Bemis et al.,<sup>12</sup> to encode which atoms of the substituent/linker bind to the ring systems as well as to which atom types they bind. We therefore encoded the ring attachment atoms as special types, the “BX” atoms, where X was the elemental symbol of the ring atom to which the substituent was attached. This encoding is illustrated in Figure 3.

Splitting molecules in this way can already yield useful information, such as which ring systems occur, and which do not (like an N<sub>6</sub>-ring). But we could get even more

information by also recording the frequencies of the substructures, as this would allow us to analyze frequency distributions and to explain why some fragments are more prevalent than others.

As a practical point, we had to find a method to encode the fragments uniquely, so that of each fragment mined we could determine whether it was already in the database or of a new fragment type. For this problem (the so-called "canonicalization problem") various algorithms and notations have been developed, such as Unique SMILES.<sup>17</sup> In this investigation, we implemented a canonical code of which the first part included the number of atoms, the number of rings (in the case of ring systems), and the number of attachment points (in the case of substituents and linkers). The second part contained the atoms, which were sorted first on their number of neighbors and the number of bonds of those neighbors, second on atom type, and finally on hybridization (sp, sp<sup>2</sup>, sp<sup>3</sup>). Since the fragments were relatively small, this simple method worked well.

### CO-OCCURRENCE ANALYSIS

After the entire molecule database was split into fragments we performed a co-occurrence analysis: which fragments were unexpectedly often found together and which seemed to "avoid" each other.

Of course, two fragments that are frequent would occur much more often together than two infrequent fragments; however, that would not necessarily mean that there is a relationship between the two. Therefore we decided to do a stochastic experiment.




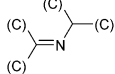
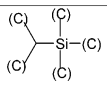
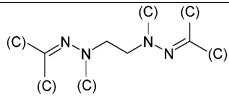
First, we selected those fragments which occurred in more than 20 molecules (in order to obtain statistically significant and chemically useful results). Then we "simulated" an NCI database by randomly dividing the different fragments over as many "molecules" (bins) as they were part of in the real database. So if a certain fragment occurred in 500 molecules in the original database, it was divided over 500 randomly chosen bins of the 250251 available in the simulated database. We counted how often each combination of fragments occurred and repeated the simulation a thousand times. These results were compared with the co-occurrence counts of the NCI database.

For example, if we had taken a database of 1000 molecules, in which 500 phenyl rings occur and 100 methyl groups, there would be on the average 50 co-occurrences of these two groups. An experiment would find that they co-occurred together about 50 times, with a standard deviation (SD) of about 4.7. However, if in the real database they occur 75 times together, which is 5.3 SD from the expected value, this indicates a correlation, which may have synthetic and/or biological reasons.

### RESULTS OF FRAGMENT FINDING

The mined NCI database contained 250 251 compounds. These molecules were split into ring systems, substituents, and several types of linkers, in total 13 509 different ring systems and 52 103 different nonring fragments. Of these nonrings 19 602 were unconnected fragments, mostly anions such as sulfate and molecules without rings. More interesting were the other nonring fragments: 18 015 were substituents, 9675 linked two ring systems, 2531 linked three ring systems,

**Table 1.** Overview of Some of the Fragment Databases We Created by Fragmenting the NCI Database<sup>a</sup>

	Number of unique fragments	Number in database	Example
Ring systems	13509	416867	
Substituents	18374	617722	(C)—CH <sub>3</sub>
2-linkers	9990	101402	
3-linkers	2602	5776	
4-linkers	974	2388	
5-linkers	126	177	
6-linkers	218	280	

<sup>a</sup> For example, the database of substituents (groups attached to only one ring system) contains 18 374 different types of substituents, which together occur 617 722 times in the NCI database. Also given is an example of the particular type of fragment, for the substituents this is the methyl group attached to a carbon atom in a ring system.

and 2280 linked four or more ring systems. The most highly connected linker was attached to 18 ring systems. The number of different fragments in the largest categories as well as the total occurrences in the molecules and some example fragments are shown in Table 1.

Visual inspection of the fragments and their occurrences led to the following observations, some of which were already known qualitatively but which could now be confirmed quantitatively through the data mining:

(1) Many of the ring systems and branches contain metal atoms or metallic atoms as boron. In the case of rings, 2722 out of 13 509 (20%) contained atoms other than C, N, O, and S, such as As, Fe, B, and Si. Of the substituents, 1736 out of 18 015 contained atoms other than C, N, O, S, and the halogens, less than 10%. The two- and three-connected branches had 11% and 24%, respectively, while most linkers with six or more attachment points contained metals or less common heteroatoms (B, P, Si, and such).

(2) In general, the larger the ring or branch, the smaller its frequency seems to be.

(3) Metals and higher-weight nonmetallic elements both occur relatively rarely in fragments and make a ring or branch occur less often. Carbon atoms dominate rings and other fragments, followed by nitrogen and oxygen, which are in turn more prevalent than sulfur, phosphorus, and finally the metals.

(4) In branches, a higher number of attachment points seems to mean that it is less used. Bemis et al.<sup>11</sup> did not find any frequent frameworks with 3-linkers or higher-order linkers in the Comprehensive Medicinal Chemistry (CMC) database.<sup>18</sup> Tables 1 and 2 confirm this observation.

**Table 2.** Overview of the Number of Fragments Linking Seven or More Ring Systems<sup>a</sup>

no. of attachment points	7	8	9	10	11	12	13	14	15	16	17	18
no. of fragment types	22	15	15	1	1	26	0	0	0	0	0	2

<sup>a</sup> For example, there are 15 unique fragments which are attached to nine ring systems simultaneously.

(5) The only exception to the rule that the more attachment points a linker has, the less frequent it is, is going from 5-attached linkers to 6-attached linkers. Inspection of the structures of the 6-linkers shows that most of them are symmetrical and therefore possibly easier to synthesize. However, 6 and multiples of 6-linkers are uncommonly popular (Table 2; compare 12 to 11 and 13, 18 to 17)—probably this is due to metal complexes and the high symmetry possible (both 2- and 3-fold symmetrical). Perhaps investigations of larger databases could confirm whether there really is a “rule of six”.

(6) The ratio of the occurrence of fragments to the number of unique fragments decreases as one goes from substituents to rings to linkers. The ring ratio (13 509 unique rings, together occurring 416 867 times in the database) is 31, for the 18 374 substituents it is 33, for the 2-linkers 10, for the 3-linkers 3.2 times, and for the four-linkers 2.5. It may be that the more unique fragments there are in a category, the more lopsided the distribution will be.

As an illustration of our results, the top 10 fragments of the most common fragment families are shown in Table 3.

## RESULTS OF THE CO-OCCURRENCE ANALYSIS



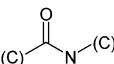



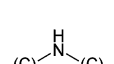



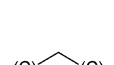



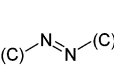



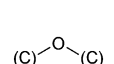



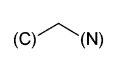



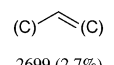



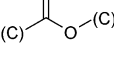



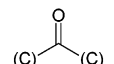



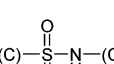

Our investigations found 65 612 fragment types in 250 251 molecules. Correlating all these fragments to each other would have resulted in about 4 billion correlations, but most of these would be meaningless since about 70% of fragments occur only once in the database. To reduce computational cost and find only the co-occurrences in a decently sized set of molecules (we set the threshold somewhat arbitrarily to at least 20 molecules), we only calculated co-occurrences for fragments which occurred in 20 or more molecules—1895 fragments, just 2.9% of the total. Among these fragments were also some metal-containing and therefore less interesting fragments, which we removed. The final set therefore contained 1730 different fragments, 2.6% of the total number.

We created 1000 simulated databases, as described in the methods section, and calculated the expected occurrence of each pair of fragments as well as the standard deviation of these co-occurrences (for among different simulations, the number of co-occurrences of a pair of fragments can vary). Then we compared the expected co-occurrences and the deviation in SDs (the z-values) to the real co-occurrences in the NCI database.

The distribution of z-values of a sample simulated database yielded a Gaussian-like distribution, as expected (Figure 4). The distribution of fragment co-occurrences in the NCI database in the same figure is, however, remarkably different.

The NCI database turns out to possess both a large number of fragment pairs which seem to avoid each other (1110 of z-value < -3) and an even larger number of fragment pairs which seem to group together (2897 of z-value > 3).

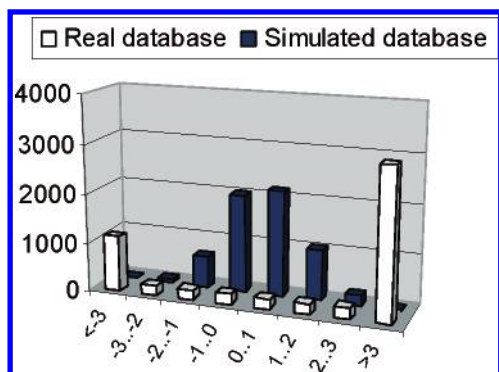
**Table 3.** Most Frequently Occurring Fragments in the NCI Database<sup>a</sup>

	ring systems	substituents	2-linkers	3-linkers
1	 200644 (48%)	 102157 (17%)	 4484 (4.4%)	 415 (5.0%)
2	 11442 (2.7%)	 77907 (13%)	 4266 (4.2%)	 384 (4.6%)
3	 7731 (1.9%)	 62949 (10%)	 3873 (3.8%)	 240 (2.9%)
4	 6991 (1.7%)	 46734 (7.6%)	 3695 (3.6%)	 179 (2.1%)
5	 6185 (1.5%)	 42080 (6.8%)	 3638 (3.6%)	 107 (1.3%)
6	 5814 (1.4%)	 24997 (4.0%)	 2940 (2.9%)	 88 (1.1%)
7	 5352 (1.3%)	 19421 (3.1%)	 2699 (2.7%)	 87 (1.0%)
8	 5120 (1.2%)	 14654 (2.4%)	 2542 (2.5%)	 87 (1.0%)
9	 4526 (1.1%)	 3184 (1.8%)	 2278 (2.2%)	 80 (0.95%)
10	 3991 (0.96%)	 7916 (1.3%)	 1966 (1.9%)	 76 (0.91%)
Total	62%	66%	32%	21%

<sup>a</sup> Listed are the top ten ring systems, substituents, 2-linkers, and 3-linkers. The numbers in each cell represent the number of occurrences of the fragment in the database and what percentage this is of all occurrences of fragments of its type. For example, the keto substituent (C=O) occurs 79 707 times in the database and thereby represents 13% of all substituent occurrences; if one would randomly pick a substituent from a molecule, the chance is 13% that it is a keto group. The numbers in the bottom row are the total percentage of the top ten fragments.

We first sorted our results on statistical significance, reaching z-values of up to 490 and down to -65. However, after viewing the results we realized that a significant





**Figure 4.** Overview of the number of standard deviations that the real co-occurrence of a fragment pair differs from how the co-occurrence would be if the distribution of fragments over the molecules in the database were random. The X axis displays the deviation range, the Y-axis the number of pairs in a certain range. Only pairs which occur over 20 times are counted here. The distribution of the simulated (random) database is Gaussian-like, but the real database has lots of fragments co-occurring much more or much less frequently than expected.

**Table 4.** Some Fragment Pairs Which Occur Much Less Frequently Together in One Molecule than Expected and Therefore Seem To Avoid Each Other<sup>a</sup>

z-value	Fragment 1	Fragment 2	Expected occurrence	Real occurrence	Fraction
-31			534	42	0.08
-36			1209	115	0.10
-67			2653	270	0.10
-19			544	139	0.26
-37			1186	323	0.27
-14			611	281	0.46

<sup>a</sup> For example, the phenyl group and the tetrahydrofuran group (row 3) are both very prevalent fragments and would be expected to occur together in about 2653 molecules of a 250 000 molecule database. However, in the NCI database they are combined in only 270 molecules, giving an “expectation fraction” of 270/2653 or 0.10. This relation is highly significant, being 67 standard deviations under the expected value.

effect does not have to be a very large effect. The most statistically significant negative correlation is that between the benzene and tetrahydrofuran rings, which occur about 10 times less together than expected. The second most significant correlation is between benzene and pyridine, only a 2.4-fold difference. However, *lower* significance is given to higher co-occurrence factors, such as 6.5 for the benzene–tetrahydropyran pair. Therefore we sorted all pairs that had sufficient z-values on their ratio of expected occurrence to actual occurrence. To illustrate our results, we have listed six of the most “avoiding” combinations with z-values < -5 (Table 4). These combinations were expected to occur

**Table 5.** Some Fragment Pairs Which Co-occur Much More Often than Expected<sup>a</sup>

z-value	Fragment 1	Fragment 2	Expected Occurrence	Real Occurrence	Factor
206			45	1396	31
206			122	2292	19
117			97	1171	12
122			371	2677	7
117			2.3	206	88
185			6.1	463	76
108			7.0	286	41
125			21	591	28
425			0.024	65	2708
125			0.26	62	238
122			0.30	66	221
106			0.39	68	173

<sup>a</sup> Shown are four fragment pairs which occur in more than 1000 molecules, four pairs which occur in more than 200 molecules, and four pairs which occur in more than 50 molecules. For example, the tetrahydrofuran group and the -CH<sub>2</sub>OH group would be expected to occur only 122 times together, but the pair appears in 2292 molecules (the explanation is, of course, that these would be ribose-containing molecules). The “gain factor” here is 2292 divided by 122 is 19, and the relation is highly significant, since the found occurrence is over 200 standard deviations from the expected occurrence in a random database (making the chance that their occurrence is independent under 0.0000001%).

at least 500 times in the database but were found much less frequently.

Even more numerous than the avoiding pairs were fragments which seemed to group together. Many of them co-occur so often, that they could be called “chemical clichés”. Often the z-values of the correlations were much bigger for these groups than for the avoiders. In Table 5 we show some of the fragment pairs with the strongest enrichment in occurrence for clichés which occur over a 1000 times, over 200 times, and over 50 times in the NCI database.

## DISCUSSION

In this work we performed fragment mining and co-occurrence analysis on a diverse, medium-sized chemical database. In this section we discuss what we can learn from

the results, compare our work with that of other investigators, consider uses of the fragment data acquired, and hypothesize about the possibilities for extension and improvement of this work.

First, we summarize our conclusions from the results obtained.

Our first observation confirms that of Bemis et al.<sup>11,12</sup> and Xue et al.,<sup>13</sup> namely that chemical fragment distributions are extremely lopsided, with a few frequent fragments and many infrequent fragments. Our investigations of several different categories of fragments (rings, substituents, and linkers) however refine this rule; it seems that the classes of the most prevalent fragments, substituents and ring systems, have the most lopsided distribution. In the less used classes of fragments (such as 3- or 4-connecting linkers) the differences in occurrence between the “top-10” and the “bottom-10” fragments are less pronounced (Table 3). It may be that in a category of fragments which has not yet been used often, strongly preferred substructures cannot arise or have not arisen yet.

One could speculate on what influences the occurrence of a certain fragment. Three factors come into mind. First, synthetic feasibility/availability: how easy is it to synthesize the fragment or is the fragment already incorporated into commercially available starting materials. The phenyl group would be a good example of this. Second, versatility: how easy is it to attach other groups to it. Third, popularity: more and more knowledge is accumulated on a popular fragment which makes it more attractive for use by others, since there is more knowledge available for its manipulation. This could lead to a kind of “winner takes it all” effect, in which relatively small differences in fragment quality may lead to big differences in use. Distinguishing between these possibilities would require further study, for example, to find out how many combinations with other groups a certain fragment has per occurrence.

The fragment co-occurrences give other insights into chemical space. From looking at the structures of the fragments that seemed to avoid each other we could think of different reasons why fragments co-occur less frequently than expected. The first reason may be that there are different classes of compounds, such as natural compounds and “synthetic” compounds. In natural compounds, sugar and nucleobase systems may be more prevalent, while in many industrial chemicals the phenyl group plays a dominant role. An example would be the third pair of Table 4, in which the tetrahydrofuran ring (as part of ribose) would occur in the natural compounds, while the phenyl is more likely to occur in “synthetic” compounds. A second reason may have to do with ease of combination: a keto group cannot be attached directly to a phenyl ring and therefore tends to occur less often with it. It can, of course, be attached to another ring system in the same molecule, but since effectively one part of the molecule has no positions available for it, the overall chance of the keto group occurring in a phenyl-containing molecule will be lower than average. Finally, some combinations may be found by the statistics to be less frequent than expected since one group is used as a replacement for the other group. Thus, bromine and chlorine relatively rarely occur in the same molecule, possibly because they have similar electronic and chemical properties. Likewise, naphthalene and benzene can take similar roles as

molecule cores and “compete” since most molecules have only one or two ring systems.

Let us now turn to the possible reasons for the clichés. Looking at the clichés we found, the first likely reason for their existence is that synthetically the clichés do not represent the smallest building block of a molecule. If moieties such as ribose (instead of unsubstituted tetrahydrofuran) are the real building blocks used to create larger molecules, the co-occurrence of the tetrahydrofuran ring (present in ribose) and  $-OH$  groups is certainly not surprising.

The explanation for a number of other clichés is that they represent specific classes of biologically active compounds. As examples, we found dihydrocholesterol analogues (Table 5, fifth pair), doxorubicin analogues (Table 5, sixth pair), mitomycins (Table 5, ninth pair), and folic acid derivatives (not shown) listed as clichés with 70- to 2700-fold occurrence relative to expectation. These clichés do not so much reflect the choice of building blocks but rather show the active structures nature provided and chemists explored around.

Let us now compare our results with those of others. Fragment mining has been done by several researchers, both as a main research subject<sup>11–13</sup> and as a preparation for virtual synthesis.<sup>14–16</sup> In the Methods section we already touched upon the different types of fragmentation, of which the retrosynthetic fragmentation, though not chemically perfect, has been applied and used by those researchers who want to perform virtual synthesis as a prelude to real synthesis. Studies with graph splitting have mainly focused on exploration of the (druglike) molecule space.

Co-occurrence analysis as reported here has to our knowledge not been done yet, and we therefore will focus our comparison on the diverse types of fragment finding as performed by other investigators and ourselves. The differences between our research in fragment frequency and that of others are caused by three factors: the database mined, the breaking points considered, and the ways in which the fragments are represented and distinguished.

Let us first compare the databases mined. Bemis et al.<sup>11,12</sup> used a rather small database, the Comprehensive Medicinal Chemistry (CMC) database,<sup>18</sup> which was filtered to get an even smaller database containing only drugs and drug candidates (5120 compounds). On the other hand, Lewell et al.<sup>9</sup> used a big database containing several millions of compounds, many more than we used. However, the relatively small size of the NCI is sufficient for an exploratory study such as ours, and our algorithm is fast enough to make mining of databases of 10 million compounds quite doable on a personal computer. Fragment mining, though not as “new” anymore as in 1996,<sup>11</sup> is something that has to be done periodically since the amount of data available is also growing and offering new opportunities. Another aspect is the quality of the data; a druglike database might give more valuable information for the pharmaceutical industry but might give a skewed image of chemical synthesis. A more general database, such as the NCI that we mined, seems more appropriate for getting general information about chemistry but will give fragments which would be unsuitable for drug development.

The investigations also differed in the breaking points considered. Most authors have divided drug molecules into substituents and a kind of framework that contains ring

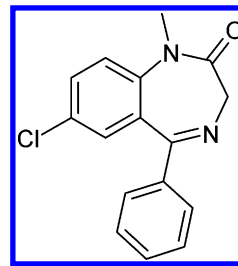
systems and linkers. Bemis et al.<sup>11,12</sup> considered the substituents and the frameworks, which were the ring systems together with parts of the linkers. Lewell et al.<sup>9</sup> concentrated on the rings, Xue et al.<sup>13</sup> on a framework-like part of the molecule called the scaffold. To our knowledge, there has not been a separate investigation of linkers, especially not of the rarer linkers with more than two attachment points. However, linking ring systems together is important for drug development, and a catalog of linkers of varying length could have similar usefulness as the ring system catalog compiled by Lewell et al.<sup>9</sup> So the use of the linker breaking points by our method yields additional useful information.

The last point is substructure representation. The aspects relevant here are the representations of the atoms and bonds in the substructure itself and the encoding of the attachment points. The atoms and bonds can be given a general type (like a wildcard that can represent any atom or any heteroatom), which results in more “general” fragments. These general fragments will necessarily be fewer in number than the original (normal) chemical fragments. A more important issue, from a chemist’s point of view, is that such a general fragment can encode many substructures of possibly vastly differing ease of synthesis. So before one chooses between using more general or more specific fragments, one should consider whether one just needs to know if frameworks with approximately the right size and shape are available, or whether one rather needs a specific substructure with a high chance to be synthesized easily.

Xue et al.<sup>13</sup> treated the substituents as unattached R-groups and did not distinguish between a methyl attached to a carbon ring atom and a methyl attached to a nitrogen ring atom. We and Bemis et al.<sup>12</sup> do distinguish between those options. For the ring systems, Lewell et al.<sup>9</sup> also considered ring systems with different attachment points (for example ortho- and meta-substituted phenyl) as distinct ring systems. Chemically, some positions in rings are easier to modify than others, but it is unclear how important this difference in reactivity is. Would other substitution positions be impossible? It is difficult to estimate the advantage gained by using only known ring substitution patterns against the loss of perfectly viable ring systems which accidentally have not been substituted in that particular pattern yet. For initial exploration of chemical space around a lead molecule, one would prefer substructures which are easy to incorporate in the molecule. For the fine-tuning of structures, however, it would be better to have more candidates available, even if not much is known yet on some of them. We could therefore say that the level of detail of substructure representation can be chosen relatively freely, but different levels of detail will be preferred for different phases of molecule design and by different chemists. Some chemists might choose a maximum amount of attachment information, while others might allow more “wildcards” in the structure. The best answer would therefore be a number of fragment databases, each with its own specificity, from which the most appropriate level of specificity can be chosen.

The next point is how we can make use of the fragment libraries and correlations.

The first use of the fragment libraries would be to give chemists more ideas for lead optimization. While investigators such as Lewell et al.<sup>9</sup> mainly considered ring systems that are sterically and electronically similar to a lead ring



**Figure 5.** Diazepam, a typical benzodiazepine.

system to be useful for chemists, we suggest that using the most common as well as the least common fragments could also be effective. The most common fragments could be used as a kind of checklist in the first and most exploratory phase of lead optimization; these fragments are apparently often easy to incorporate into molecules and thus can lead to fairly diverse exploration at relatively low costs in time and effort.

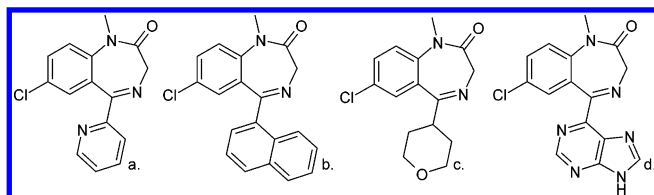
For example, consider benzodiazepines, which are widely used as tranquilizers. A typical benzodiazepine scaffold is shown in Figure 5.

Almost all benzodiazepines use the phenyl ring as group. However, is this biologically necessary or just usage of the well-known phenyl cliché? Using the program SciFinder<sup>19</sup> to search the CAS database, we found over 14 000 phenyl-benzodiazepine compounds. Going down in our ring system list of Table 3, we found that the second and third most popular ring systems, pyridine and cyclohexane, have been tried a few hundred times as phenyl substitutes. The numbers 4–10 of our list have in general only been used a few times up till a few dozen times, but numbers 6 (tetrahydrofuran), 8 (tetrahydropyran), and 10 (purine) never. So while the compounds in parts a and b of Figure 6 have been made, parts c and d are yet unexplored. While it may be possible that these particular ring systems are difficult to incorporate, the search strongly suggests that by just going over the top positions of a list of ring systems or substituents one can easily generate a few dozen variations on a lead compound. Since all of these fragments are quite frequent, many of the suggestions are probably relatively easy to synthesize or incorporate.

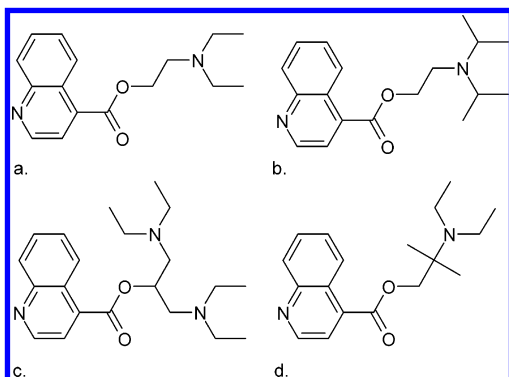
Another example of clichés is the class of local anesthetics, many of which have a phenyl ring (procaine, benzocaine, prilocaine), with as one of the few exceptions cinchocaine, which has a quinoline ring instead of the phenyl. Searching at the typical local anesthetic tail-pattern of  $C(=O)OC_xN$  in the substituent database yields many variants of the standard  $COOC_2H_4N(C_2H_5)_2$  pattern, all of which have been tried with phenyl (which is not surprising, since phenyl is the “golden standard” among rings), but some of the less frequent substituents have never been paired yet with the quinoline group (Figure 7). Some of these might also be worthy of further investigation.

Selecting fragments for rarity, conversely, can also pay off if the current scaffold is patent-protected; rare substructures can be especially helpful in this case as it is unlikely that many experiments have been done on them. Also, industries could deliberately add the more attractive of the rare fragments to their compound libraries and collections. In this way the libraries will become more diverse and thereby increase the coverage of chemical space and the chance of finding a lead compound in the first place.





**Figure 6.** Nonphenyl benzodiazepines. Compounds with the first two types of attached ring systems (pyridine (a) and naphthalene (b)) are known. However, neither the tetrahydropyran (c) nor the purine ring (d) have so far been combined with the benzodiazepine scaffold.





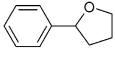


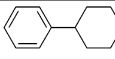


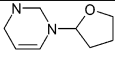
**Figure 7.** Finding cinchocaine derivatives: examples of local anesthetic-like tails which have been used (a, b) and which have not been used yet (c, d) with the quinoline group.

The fragment co-occurrence data can also have some applications. First, co-occurrence analysis can help database analysis by automatically finding clusters of biologically active or well-investigated structures. This can help “summarize” the database for a new user or alert an expert that a certain class of compounds is suddenly becoming popular. Second, co-occurrence analysis can add information to structure searching in databases. Often, when a substructure is entered, a long list of structures, each with its specific combination of R-groups, is returned. While currently there is only little attention paid to the number of times a certain R-group appears, a co-occurrence analysis could direct the chemist’s attention to the fact that a certain combination occurs quite often. So there would be something interesting with that combination.

A third application would be for chemists to find relatively unpopulated places in chemical space. For example, there would not be many good reasons why phenyl and tetrahydrofuran could not be combined into one molecule; the relative absence of the combination (Table 4, pair 3) suggests that a compound combining these groups might be worthwhile to synthesize.

In the accompanying article (Lameijer et al. “The Molecule Evuator. An Evolutionary Algorithm for the Design of Druglike Molecules”) we describe a procedure for generating new druglike molecules. The molecules we discovered were small and relatively simple but often not yet known in the literature as a compound or as a substructure of other compounds. The molecule we show in that article has a phenyl ring attached to a piperidine ring and a  $\text{CH}_2\text{OH}$  group. While the phenyl ring itself is very much a chemical cliché, according to our data it usually avoids both the piperidine ring (avoiding pair rank number 75) and the  $\text{CH}_2\text{OH}$  group (avoiding pair rank number 78). This suggests that we might also be able to use this process the other way around: by

**Table 6.** Avoiding Groups Can Indicate “Holes” in Chemical Space, Where Relatively Little Research Has Been Done<sup>a</sup>

Fragment pair	z-value	Substructure investigated	Occurrence in SciFinder (estimated by the program)
 	-57		10,594
 	-23		57,723
 	+87		111,292

<sup>a</sup> In the first column of this table pairs of fragments are shown. The second column contains the z-value of correlation. Positive z-values mean that the fragments group together; negative z-values mean that they avoid each other. The third column contains a specific substructure which contains both fragments. We have always taken a substructure which contains the fragments directly connected to each other, and when there were multiple coupling possibilities, the substructure which was most frequently found by SciFinder.<sup>19</sup> In the final column the estimated number of known molecules containing the substructure is shown. So, the phenyl and tetrahydrofuran seem to avoid each other rather strongly ( $z = -57$ ), and the substructure with the tetrahydrofuran attached at its 2-position to phenyl indeed occurs in only about 11 000 molecules.

combining fragments that are common but seem to avoid each other a person or computer program could find simple structures that have not been synthesized before.

Would we indeed be able to discover new compounds by considering the negative co-occurrences? To check that, we looked at three different pairs of rings, which were a strongly avoiding pair (phenyl and tetrahydrofuran, rank 7 of the list), the less strongly avoiding pair of phenyl and the somewhat less frequent piperazine ring (rank 75), and the cliché of the  $\text{C}_4\text{N}_2$  ring depicted in the table together with the tetrahydrofuran (Table 6). The most strongly avoiding pair appears in about 10 000 molecules in the CAS database (so at a ratio of about one in 2000 compounds in the database). Assuming independence, one would expect that the phenyl-piperazine pair occurs less often since the piperazine is (at least in the NCI database) rarer than tetrahydrofuran. However, since the groups avoid each other less, this combination occurs approximately 60 000 times, so six times as many. Finally, the  $\text{C}_4\text{N}_2$ -tetrahydrofuran cliché, despite both rings being relatively rare on their own, occurs in over 111 000 compounds, since this combination is the core of uracil and thymidine molecules.

It may also be possible to apply fragment analysis and fragment co-occurrence analysis to other problems such as measuring chemical diversity. While there are several different measures of chemical similarity, the methods most similar to ours are those which make fingerprints of molecules based on the presence and absence of fragments (for example the MDL/MACCS keys and the BCI fingerprints<sup>20</sup>). The MDL and BCI fingerprints usually work with only very small fragments, such as “an atom in a multiple, nonaromatic bond located two bonds away from an atom with at least two heteroatom neighbors”. In contrast, our method uses much larger and more specific fragments and generates only a few fragments per molecule, which will lead to a much larger emphasis on changes in framework (so naphthalene will be significantly different from benzene, though they both have aromatic carbon atoms). Diversity



could be estimated by, for example, dividing the number of unique fragments by the size of the database. For the NCI this would be (not counting the unconnected fragments)  $46\,010/250\,251 = 0.18$  or, as Ertl's research<sup>8</sup> suggested, dividing the log values, which would yield 0.91 in our case. Calculating the entropy of the distribution might also be worthwhile (a database with 999 times fragment 1 and 1 times fragment 2 could be considered less diverse than when the division is 500–500).

The result of using larger fragments instead of small fragments would be that chemical diversity is enhanced; a problem might be ease of synthesis or cost of acquiring such a library. But diversity does perhaps not have to be high, since the research of Bemis et al.<sup>11</sup> has shown that half of all drug molecules have one out of only 32 different frameworks. One apparently does not need very high diversity to get active drugs on very different targets. On the other hand, if the NCI is representative of chemical space, most alternative frameworks have been rarely synthesized and screened and would therefore have a much smaller chance of leading to a drug. Until it is shown that alternative frameworks are really much worse for drug design, the chemical diversity stimulated by our "big fragment" method could be a good start.

The final benefit of fragment mining and co-occurrence analysis might be psychological: thinking of substructures which have or have not been used together might make chemists more conscious of their choices, giving them more knowledge to decide whether to use clichés for ease of synthesis or avoid them to explore novel structural classes.

As the final part of this discussion we would like to reflect on what directions our fragment mining investigations could take.

First of all, there are some possibilities for algorithm improvement. While doing a stochastic simulation of fragment pair expectance is relatively easy and computationally cheap (about 15 min on a 3 GHz PC), the averages and standard deviations can be calculated exactly with a so-called chi-squared distribution with one degree of freedom. This will be especially valuable for larger databases. Stochastics, however, may continue to play a role since they make it easy to add certain restraints (such as a maximum number of fragments per molecule or multiple identical fragments per molecule) that are more difficult to enforce by mathematics. It would also be interesting to mine larger databases, such as ZINC<sup>21</sup> or PubChem.<sup>22</sup> In any case this is likely to add fragments to our databases and perhaps discover new clichés or avoiding groups since more data can lead to more strongly pronounced z-values.

A second direction for further investigation would be to make the relationships of the co-occurrences more detailed. For example, currently we only detect whether two fragments are present in the same molecule. However, our method can be extended by taking into account whether the fragments are directly attached to each other, or whether the attachment point is consistent over many molecules. Additionally, detection of co-occurrences of three fragments or more could be a worthwhile extension.

A third development would be experimenting with different representations of the substructures; some chemists would not care whether a methyl group is attached to a ring-N or ring-C, whereas others would like to be sure that

a certain ring position is suitable for substitution. Atoms and bond types could be converted to wildcards to create a smaller library of general fragments, or conversely the connection points could be extended and classified for more certainty of ease of synthesis. In the end, we would like to have a system that provides the right kind of data for each application.

## CONCLUSIONS

In this investigation we mined the NCI database of 250 251 compounds. This resulted in over 60 000 fragments of different types: ring systems, substituents, and diverse kinds of linkers. Fragment occurrence is very skewed, with 70% of fragments occurring only once, and a few fragments (such as phenyl and methyl) being present in many molecules.

The fragment lists and co-occurrences can be used in different ways. In our examples we have shown how the fragment lists can be used to find new ring substituents for benzodiazepines and local anesthetics. Also, we found that co-occurrence analysis can automatically detect groups of biologically active compounds, such as the doxorubicin and mitomycin analogues in the NCI. Finally, co-occurrence analysis of the avoiding fragments can show "holes" in chemical space where there is room for small, novel compounds which may be biologically active.

Future directions of this work could be investigating either larger or more focused databases, taking information of how fragments are attached to each other into account and experimenting with different levels of substructural detail. Fragment analysis can show us many chemical patterns, but the conversion of pattern knowledge into chemical understanding has only just begun.

SD files containing the fragments mined and their occurrences are available from the authors upon request.

## REFERENCES AND NOTES

- (1) [http://www.mdl.com/products/knowledge/crossfire\\_beilstein/](http://www.mdl.com/products/knowledge/crossfire_beilstein/).
- (2) <http://www.cas.org/chemplus/chemplus1.html>.
- (3) Lipinski, C.; Lombardo, F.; Dominy, W.; Feeney, P. Experimental and Computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Delivery Rev.* **1997**, *23*, 3–25.
- (4) Oprea, T. Property distribution of drug-related chemical databases. *J. Comput.-Aided Mol. Des.* **2000**, *14*, 251–264.
- (5) Xu, J.; Stevenson, J.; Drug-like Index: A New Approach To Measure Drug-like Compounds and Their Diversity. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1177–1187.
- (6) Klopman, G. Multi-CASE: a hierarchical computer automated structure evaluation program. *QSAR* **1992**, *11*, 172–184.
- (7) Kazius, J.; McGuire, R.; Bursi, R. Derivation and validation of toxicophores for mutagenicity prediction. *J. Med. Chem.* **2005**, *48*, 312–320.
- (8) Ertl, P. Cheminformatics Analysis of Organic Substituents: Identification of the Most Common Substituents, Calculation of Substituent Properties, and Automatic Identification of Drug-like Bioisosteric Groups. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 374–380.
- (9) Lewell, X.; Jones, A.; Bruce, C.; Harper, G.; Jones, M.; Mclay I.; Bradshaw, J. Drug Rings Database with Web Interface. A Tool for Identifying Alternative Chemical Rings in Lead Discovery Programs. *J. Med. Chem.* **2003**, *46*, 3257–3274.
- (10) <http://cactus.nci.nih.gov/ncidb2/download.html>.
- (11) Bemis, G.; Murcko, M. The Properties of Known Drugs. 1. Molecular Frameworks. *J. Med. Chem.* **1996**, *39*, 2887–2893.
- (12) Bemis, G.; Murcko, M. Properties of Known Drugs. 2. Side Chains. *J. Med. Chem.* **1999**, *42*, 5095–5099.
- (13) Xue, L.; Bajorath, J. Distribution of Molecular Scaffolds and R-Groups Isolated from Large Compound Databases. *J. Mol. Model.* **1999**, *5*, 97–102.

- (14) 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.* **2003**, *46*, 2765–2773.
- (15) Schneider, G.; Lee, M.; Stahl, M.; Schneider, P. De novo design of molecular architectures by evolutionary assembly of drug-derived building blocks. *J. Comput.-Aided Mol. Des.* **2000**, *14*, 487–494.
- (16) Lewell, X.; Judd, D.; Watson, S.; Hann, M. RECAP—Retrosynthetic Combinatorial Analysis Procedure: A Powerful New Technique for Identifying Privileged Molecular Fragments with Useful Applications in Combinatorial Chemistry. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 511–522.
- (17) Weininger, D.; Weininger, A.; Weininger J. SMILES. 2. Algorithm for Generation of Unique SMILES Notation. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 97–101.
- (18) CMC is currently 2005 marketed by Elsevier MDL, see also the Web site [http://www.mdl.com/products/knowledge/medicinal\\_chem/index.jsp](http://www.mdl.com/products/knowledge/medicinal_chem/index.jsp).
- (19) We used SciFinder Scholar marketed by CAS: <http://www.cas.org/SCIFINDER/SCHOLAR/>.
- (20) Wild, D.; Blankley, C. Comparison of 2D Fingerprint Types and Hierarchy Level Selection Methods for Structural Grouping Using Ward's Clustering. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 155–162.
- (21) Irwin, J.; Shoichet, B. ZINC — A Free Database of Commercially Available Compounds for Virtual Screening. *J. Chem. Inf. Model.* **2005**, *45*, 177–182.
- (22) <http://pubchem.ncbi.nlm.nih.gov/>.

CI050370C