Finding Drug Candidates in Virtual and Lost/Emerging Chemistry

Andy de Laet, Jurriaan J. J. Hehenkamp and Richard L. Wife*

SPECS and BioSPECS b.v., Fleminglaan 16, 2289 CP Rijswijk, The Netherlands

There are many opportunities for chemists to help make the drug discovery process more efficient. The virtual world of chemistry contains many challenges to the heterocyclic chemist but this world is very large and attention must be focused on the best candidates for synthesis and testing. The algorithm set SORT&gen makes it possible to form an opinion about the constitution of large compound collections. It also generates the ring scaffolds of the molecules that are missing in these collections. Its application in the drug discovery process is discussed and the smallest missing molecular scaffolds are presented.

J. Heterocyclic Chem., **37**, 669 (2000).

Drug discovery and development is expensive. Current estimates indicate that the total costs for a marketed drug are $500 million and that only one in three marketed drugs provide a return on investment. Drug development accounts for 60% of the total costs and the remaining 40% is spent on drug discovery.

Drug discovery begins with testing compounds, often by High Throughput Screening (HTS), and ends with candidate selection. The discovery process is illustrated in Figure 1 together with the in-house and out-sourced activities that contribute to a faster drug discovery process. The discovery process is dynamic and seemingly complex.

Despite the creative measures to make the drug discovery process more effective, it has been recognized for some time that the number of New Chemical Entities (NCEs) coming out of discovery is insufficient to sustain the pharmaceutical industry as it stands today [1-3]. The impact of new technologies such as Genomics will certainly have a positive effect on producing better drugs but time is also important. The faster a drug is on the market, and the longer the patent lifetime, the better the investment will be. For a billion dollar per year drug, every month of extended patent life brings an extra $80 million.

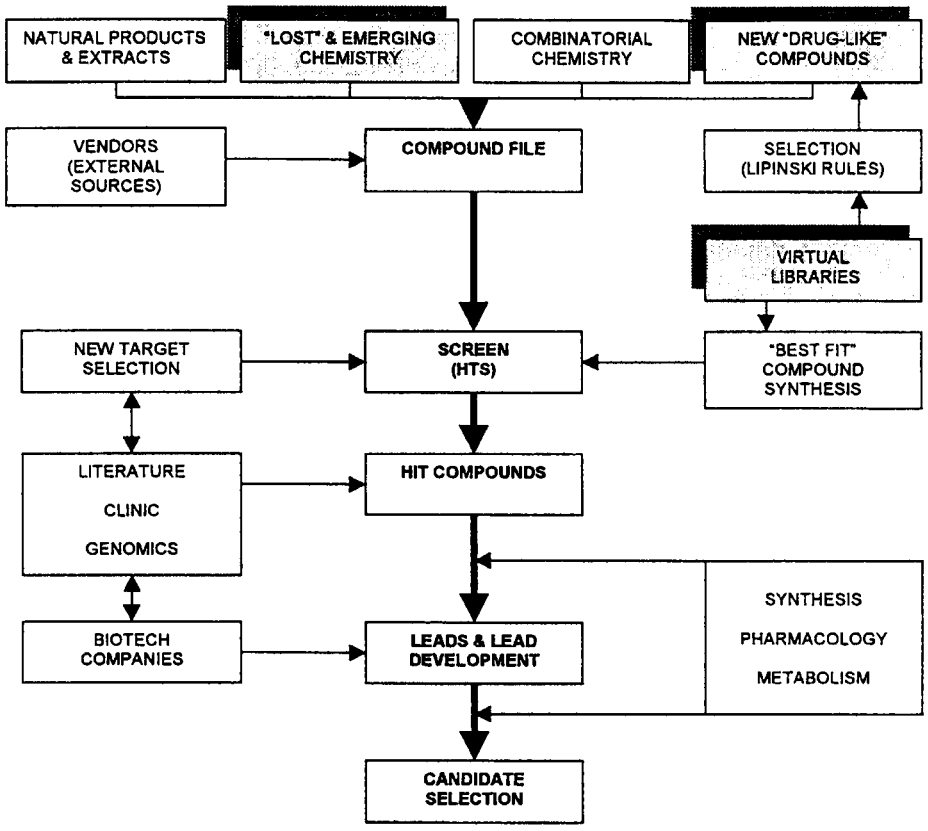


Figure 1. The Drug Discovery Process.

Genomics will result in many more targets and, as a result, there will be many more hits to evaluate. More components to the already complex discovery process will be added to select the more promising hit molecules and to develop them into leads. It is a constantly evolving process with high stakes. The basis in the drug discovery process is the compound file or the collection of compounds that can be tested/screened against various targets. While the focus of attention is on speeding up and making drug discovery more efficient, the quality of the compound file determines how effective this drug discovery will be. This is where the heterocyclic chemist can play a vital and innovative role.

Of the compounds described in Beilstein [4], 53% are heterocyclic. Analysis of the drugs in late development or on the market shows that 68% are heterocyclic [5]. The presence of one or more heteroatoms in a molecular scaffold is a characteristic property for many drugs. A heteroatom in the cycle exerts its influence on the shape and chemistry of the molecule, interrupting the hydrocarbon scaffold with functionality.

After some ten years of HTS, a picture has been made of the compounds most likely to exhibit useful biological activity. The so-called drug like properties are described in Lipinski's Rules of Five [6].

Lipinski's Rules of Five.

Drug-like properties for any molecule are:

- MW less than 500 Daltons
- LogP less than 5
- Number of hydrogen bond donors less than 5
- Total number of oxygens and nitrogens less than 10
- No more than 5 fused rings

This is a very useful and relatively simple set of rules based on the screening of hundreds of thousands of discrete molecules. While there will be exceptions to these rules, they do focus attention on drug-like characteristics which helps in the selection process of which compounds should be screened.

There is another analysis which helps guide drug discovery research and this is based on commonly recurring drug shapes or molecular frameworks [7]. It provides insight into the scaffold systems of known drugs and provides valuable information that can be used in compound selection for HTS. The study set for this analysis is from the Comprehensive Medicinal Chemistry (CMC) database [8] of 5120 compounds. It is a study of scaffolds in successful drug compounds.

In Figure 2, we depict the scaffolds of compounds that are currently the most interesting for screening for new biological activity [9]. It should be noted that these are the bare scaffolds and that functional groups have been removed. These compounds are selected for screening as a result of computational or human experience and analysis. They represent typical examples of what might one day become a successful drug, based on accumulated information from many years of testing and results.

In all of these analyses, the challenge is to find a meaningful link between biological activity and parameters that describe the molecules that are tested. Based on real screening data, such a link would be of tremendous predictive value and would make the discovery process much more effective. Perhaps the biggest challenge is first to find the most effective molecular descriptors!

One tool that has been developed for this purpose is the 2-Dimensional Molecular Diversity Space [10] which enables comparisons to be made between a compound file and an optimally diverse library. Diversity Space highlights compound "redundancies" (similar compounds that are likely to have the same biological activity) as well as "holes" (compounds that are not represented in the compound file). The size of the holes depends on the chosen descriptors and is frequently underestimated. In Diversity Space, the biggest challenge is to deconvolute from the holes to the molecular structures that would fill these holes.

More recent studies address the virtual chemistry world, its size and how intelligent choices can be made to prepare and test certain candidates. The virtual world is very big as is illustrated by the simple products from reacting 1700 different diamines with 26,700 halides. It would take 30,000 years to test these products at a screening rate of 100,000 compounds per day [11]. Such large numbers are commonplace in combinatorial chemistry. This example concerns derivatized diamines which may not be especially exciting compounds. It does nevertheless underline the size of the virtual chemical world and how intelligent methods must be designed to identify which virtual molecules deserve closer attention.

Our assessment of what constitutes a rich compound file begins with the information from the Lipinski Rules, the analysis of molecular frameworks and the sort of molecules that are most chosen for screening (Figure 2). These analyses are based on real compounds and we supplement our assessment by addressing the virtual world and our methods to generate the structures of the compounds that have not been reported. We can use data on real molecules which have proven biological activity to refine this virtual world in a second iteration.

Our approach is to use molecular descriptors that are derived from rings and heteroatoms contained in a compound structure. The ring descriptor (RD) and heteroatom descriptor (HD) are the axes for a 2-Dimensional *Scaffold Space* in which compound files can be visualized and assessed. More importantly, using these descriptors it is now possible to deconvolute from holes in Scaffold Space to the structures of scaffolds that will fill these holes.

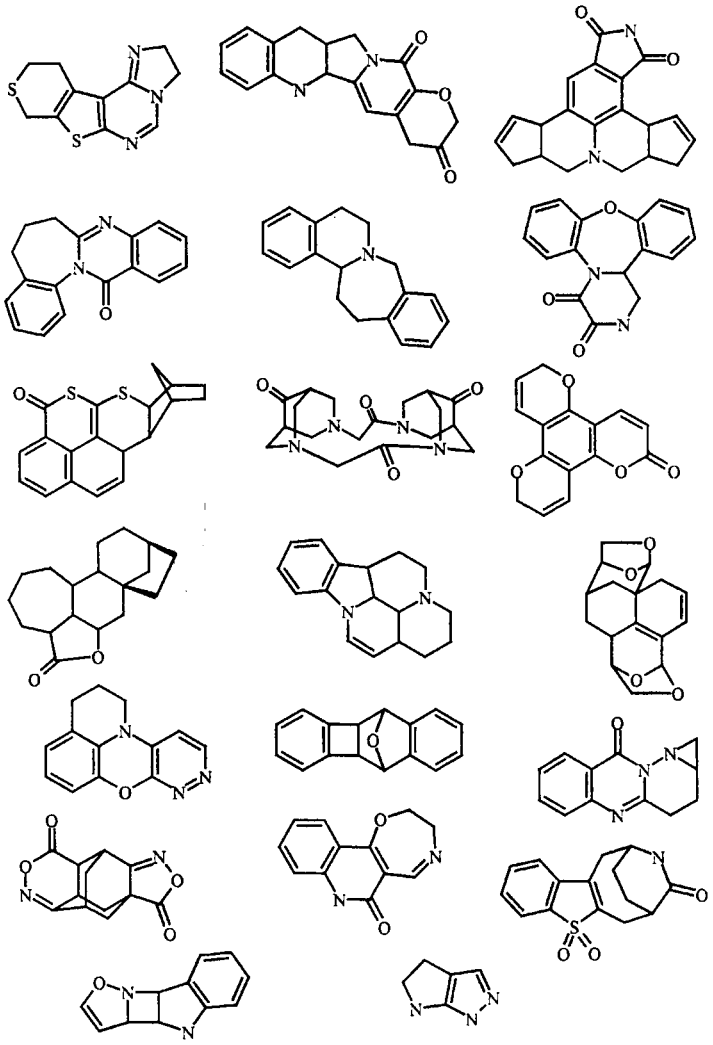
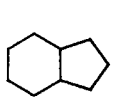
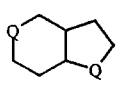


Figure 2. Most Interesting Scaffolds from Recent SPECS and BioSPECS Databases.



Largest ring = 09
Next largest = 06
Next largest = 05

RD = 090605



Q-C-C-C-C-C-C-Q = 08
Q-C-C-C-C-Q = 06
Q-C-C-C-Q = 05
Q = 01

HD = 08060501

Figure 3. Calculation of RD and HD.

The first of the algorithms that we have developed assign RD and HD parameters to molecular structures as illustrated in Figure 3. A procedure then sorts collections of molecular structures using the RD and HD resulting in a 1-dimensional representation from the largest to the smallest molecule. Compounds with the same molecular scaffold are grouped into "bins". For large collections, where it would be tedious to form an opinion by browsing through thousands of structures, a second algorithm produces the 2-Dimensional Scaffold Space where the axes are the RD and HD. The occupied bins represent the compounds that are present and a third algorithm generates the structures of the compounds that are missing — the holes in the compound collection. The complete algorithm set is called "SORT&gen".

Before illustrating the application of SORT&gen, it is worthwhile to make a critical appraisal of the value of using RD and HD as parameters to describe a molecule.

Biological activity results from the correct placement of functionality in a 3-dimensional space such that the interactions between the small drug molecule and the big molecule receptor are optimal. The RD/HD approach is 2-dimensional and addresses the molecular scaffold with little attention to the functional groups that are responsible for biological activity. That the RD/HD approach is meaningful in the classification of molecules is derived from the notion that the *molecular scaffold is a chemical surrogate.* What is meant here is that the scaffold determines where the functionality can be and what the functionality is likely to be. For a particular molecular scaffold, whether it is heterocyclic or not, the presence of functional groups on the scaffold is determined by the starting materials and reagents that are used to synthesize the molecule and subsequent chemistry is controlled by the nature of the scaffold itself.

Not only is the "3-dimensional shape partially encoded in the 2-dimensional graph of a molecule" [7] but the 2-dimensional graph also contains crucial information about the sort of functionality that can be attached to the molecular scaffold. The RD/HD approach is simple yet highly effective and enables useful comparisons to be made between compound collections. Further, the RD/HD approach in SORT&gen makes it possible to deconvolute from empty bins to missing structures.

To illustrate SORT&gen as a means to analyze compound collections and to make comparisons, we choose two sets of compounds, natural products (2,661) and synthetic products (2,000).

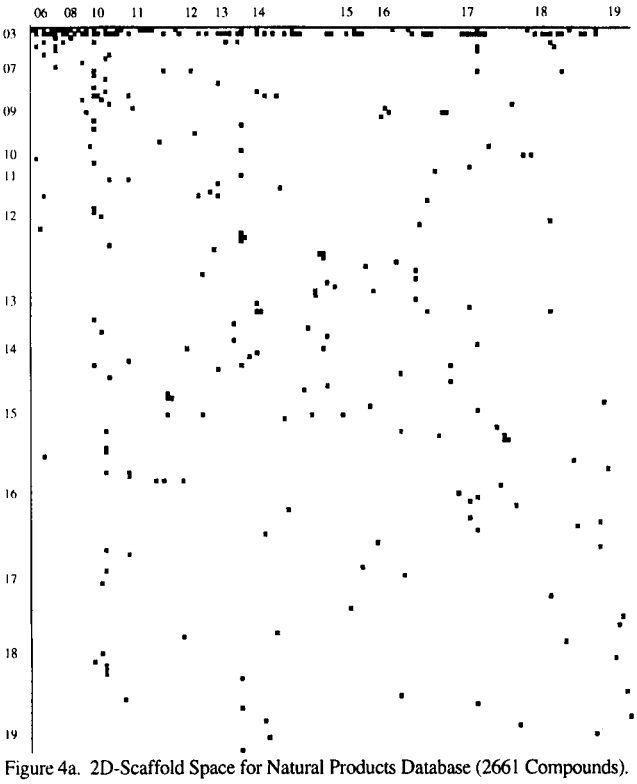


Figure 4a. 2D-Scaffold Space for Natural Products Database (2661 Compounds).

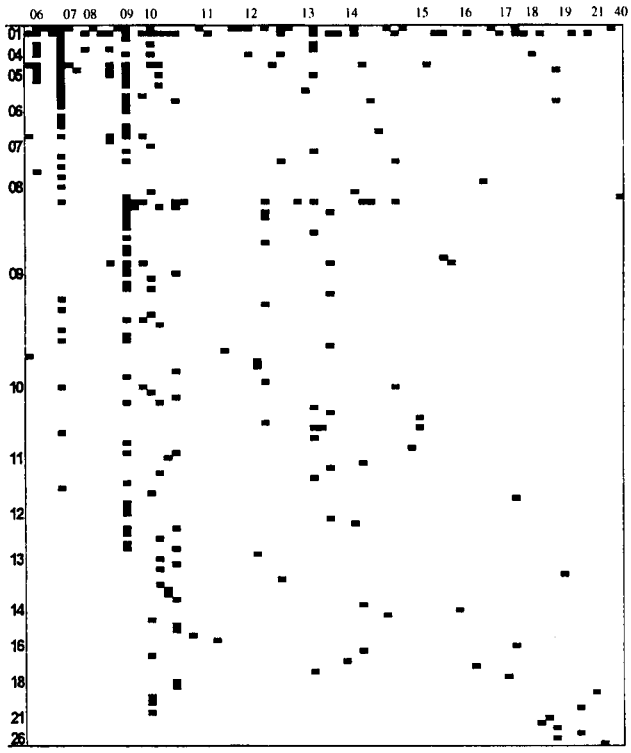


Figure 4b. 2D-Scaffold Space for Synthetic Products Database (2000 Compounds).

Scaffold Space enables meaningful comparisons between large data sets. In this illustration, it is clear that natural products contain more diverse ring systems than synthetic compounds which is to be expected. The more complex synthetic compounds are largely (hetero)aromatic and arrived at by ring annulations while natural products often contain ring-ring bridges in more saturated (hetero)ring systems.

The 2-Dimensional Scaffold Space for both natural and synthetic products contains a large number of bins that are empty. The algorithm that generates the structures of the missing molecules identifies the bins that can never be filled, the so-called "Black Space". The generation algorithm is based on graph theory and calculates all possible connectivities for a given RD. Black space is an impossible connectivity. Of the possible connectivities that are not represented in Scaffold Space, the so-called "White Space", the generation algorithm provides solutions in accordance with energy limitations applied to candidate structures. Setting a real energy limit that excludes compounds with no chance of being stable, SORT&gen displays meaningful White Space.

The question arises as to how much of the total chemistry world have we seen so far? There are some 8 million molecules in Beilstein, so how big is the unexplored virtual world and what does it look like? To gain insight into the size of the virtual world, we have started to examine solutions for (maximum) ring sizes up to 9 atoms. For each ring size, the total number of White Space solutions are calculated and this is graphically represented in Figure 6.

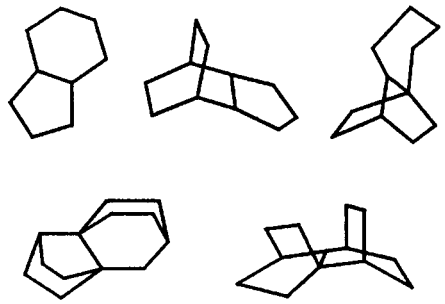


Figure 5a. White Space Solutions for RD = 090605 and 5 Extra Atoms.

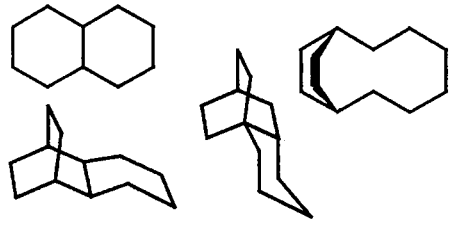


Figure 5b. White Space Solutions for RD = 1006 and 2 Extra Atoms.

Our first attention has been focused on the smallest of the unknown scaffolds. In Figure 7, we depict two solutions derived from this exercise.

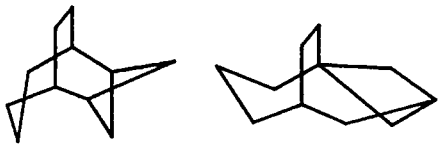


Figure 7. Smallest Missing Scaffolds; RD = 080704 and 3 Extra Atoms.

Both scaffolds contain eleven atoms and have the RD = 080704 and are unknown in the chemical literature. Lower and more strained homologues of the symmetrical scaffold have been reported however [12-15].

In arriving at this first set of missing scaffolds it should be realized that they can contain heteroatoms in any position(s). They are the missing connectivities. It is the challenge to the heterocyclic chemist to devise synthetic methods to bring to reality these virtual molecules. It should also be realized that the known or reported chemistry is only one-fifth of the chemistry that has been done and the compounds that have been made. The missing

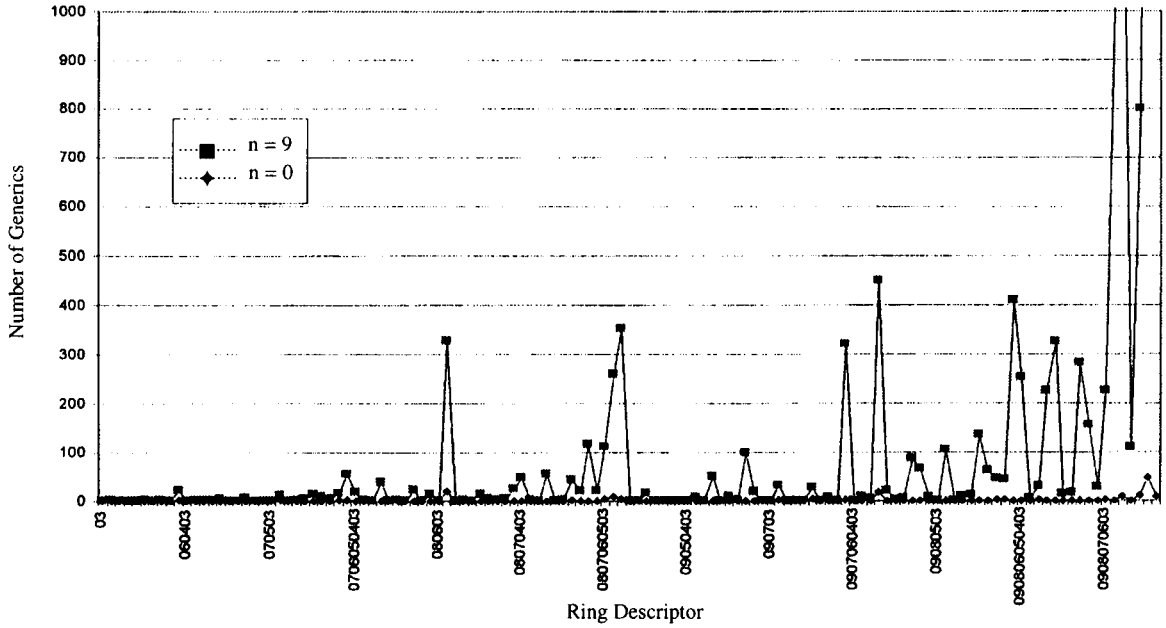


Figure 6. White Space with n = 0 and n = 9 for Ring Sizes 3-9.

By subtracting the known chemistry from the total virtual world, we arrive at the very large collection of compound scaffolds that have never been reported. From this enormous virtual collection, new drug candidates can be selected for synthesis and testing, adding functionality where appropriate.

80% will be in laboratory notebooks somewhere but never appeared in a publication.

Returning to the drug discovery process as illustrated in Figure 1, we highlight areas where chemists can make fundamental contributions that make the process faster and more efficient. The creation of virtual libraries as we

have described in SORT&gen creates a picture of the vast virtual world. Using selection criteria based on drug-like properties and synthetic access, new drug-like compounds can be designed for preparation and testing. At the same time, attention should also be focused on the other missing chemistry, the "lost" and "emerging" chemistry that is not accessible to the pharmaceutical industry. Ways to capture this lost and emerging chemistry will augment the complete drug discovery strategy.

The role of the heterocyclic chemist is clear and crucial in these three opportunity areas. Heterocyclic chemistry is the cornerstone of the pharmaceutical industry and there is a vast amount of new chemistry to be done. SORT&gen will help in the creation and selection of missing molecules and provide many challenges to the synthetic heterocyclic chemist for years to come.

Acknowledgements.

The authors gratefully acknowledge the key contributions from Hans de Bie, Peter Franken, Jiri Krechl and Greg Gardiner in developing SORT&gen and Sheila Ash and Robin Williams from Oxford Molecular Ltd. (UK) for producing the first SORT&gen product for testing.

REFERENCES AND NOTES

[*] E-mail: specs@specs.net

[1] J. Drews, The Impact of Cost Containment on Pharmaceutical Research and Development, Tenth Center for Medicines Research (CMC) Lecture, CMR Publication, 1995, p 1.

[2] J. Drews and S. Ryser, Innovation Deficit in the Pharmaceutical Industry, *Drug Information Journal*, **30**, 97 (1996).

[3] Pharma 2005: an Industrial Revolution in R & D, PriceWaterhouseCoopers, 1998.

[4] Beilstein Information Systems, release 9901.

[5] World Drug Index (WDI) release 1997.

[6] C. A. Lipinski, Advanced Drug Delivery, Fourth International Conference on Drug Absorption, Edinburgh, Scotland, June 1997.

[7] G. W. Bemis and M. A. Murcko, *J. Med. Chem.*, **39**, 2887 (1996).

[8] Available from MDL Information Systems Inc., San Leandro, CA.

[9] The scaffolds of the most ordered compounds from SPECS and BioSPECS B.V., Rijswijk, The Netherlands.

[10] H. Matter and D. Lassen, *Chimica Oggi/Chemistry Today*, 9 (1996).

[11] R. D. Cramer, D. E. Paterson, R. D. Clark, F. Soltanshahi and M. S. Lawless, *J. Chem. Inf. Comput. Sci.*, **38**, 1010 (1998).

[12] G. A. Molander and J. A. McKie, *J. Org. Chem.* **56**, 4112 (1991).

[13] W. Kirmse, D. Monch, K. Muller and K. Gomann, *Chem. Ber.*, **125**, 1297 (1992).

[14] T.-K. Yin and W. T. Borden, *J. Org. Chem.*, **51**, 2285 (1986).

[15] A. Otterbach and H. Musso, *Angew. Chem. Int. Ed. Engl.*, **26**, 554 (1987).