Is There a Difference between Leads and Drugs? A Historical Perspective

Tudor I. Oprea,*,† Andrew M. Davis,‡ Simon J. Teague,‡ and Paul D. Leeson‡

AstraZeneca R&D Mölndal, EST Lead Informatics, S 431 83 Mölndal, Sweden, and AstraZeneca R&D Charnwood, Medicinal Chemistry, Loughborough, Leics., LE11 5RH, U.K.

Received March 2, 2001

To be considered for further development, lead structures should display the following properties: (1) simple chemical features, amenable for chemistry optimization; (2) membership to an established SAR series; (3) favorable patent situation; and (4) good absorption, distribution, metabolism, and excretion (ADME) properties. There are two distinct categories of leads: those that lack any therapeutic use (i.e., "pure" leads), and those that are marketed drugs themselves but have been altered to yield novel drugs. We have previously analyzed the design of leadlike combinatorial libraries starting from 18 lead and drug pairs of structures (S. J. Teague et al. *Angew. Chem., Int. Ed. Engl.* **1999**, *38*, 3743−3748). Here, we report results based on an extended dataset of 96 lead-drug pairs, of which 62 are lead structures that are not marketed as drugs, and 75 are drugs that are not presumably used as leads. We examined the following properties: MW (molecular weight), CMR (the calculated molecular refractivity), RNG (the number of rings), RTB (the number of rotatable bonds), the number of hydrogen bond donors (HDO) and acceptors (HAC), the calculated logarithm of the *n*-octanol/water partition (CLogP), the calculated logarithm of the distribution coefficient at pH 7.4 ($LogD_{74}$), the Daylight-fingerprint druglike score (DFPS), and the property and pharmacophore features score (PPFS). The following differences were observed between the medians of drugs and leads: $\Delta MW = 69$; $\Delta CMR = 1.8$; $\Delta RNG = \Delta HAC = 1$; $\Delta RTB = 2$; $\Delta CLogP = 0.43$; $\Delta LogD_{74} = 0.97$; $\Delta HDO = 0$; $\Delta DFPS = 0.15$; $\Delta PPFS = 0.12$. Lead structures exhibit, on the average, less molecular complexity (less MW, less number of rings and rotatable bonds), are less hydrophobic (lower CLogP and $LogD_{74}$), and less druglike (lower druglike scores). These findings indicate that the process of optimizing a lead into a drug results in more complex structures. This information should be used in the design of novel combinatorial libraries that are aimed at lead discovery.

INTRODUCTION

The chemical structures of lead compounds that were developed into marketed drugs have rarely been documented in the past 100 years. Even though the concept of lead compounds is central to the process of drug discovery, the exact nature of lead structures has not been previously documented. In fact, a collection of lead structures that resulted in marketed drugs is not available in the literature. One such attempt to cover chemical aspects of drug discovery is the book series *Chronicles in Drug Discovery*,[1−3] but those historical accounts, while rich in chemical information, do not focus on lead structures. The only account focused on lead structures, published 15 years ago, discusses less than 15 structures.[4] A recent (January 2001) search in SciFinder[5] regarding the "chemical history of drug discovery" yielded only five hits,[6−10] but none of them deals with a general analysis of lead structures. Another 480 entries contain the "lead structure to drug" keyword, of which 146 were marked as reviews, but these papers are not focused on the nature of lead structures in general.

Lead structures are ligands that typically exhibit suboptimal target binding affinity. Leads should display the following properties, to be considered for further development: (1) relatively simple chemical features, amenable for combinatorial and medicinal chemistry optimization efforts; (2) membership to a well-established SAR (structure−activity relationship) series, wherein compounds with similar (sub)-structures exhibit similar target binding affinity; (3) favorable patent situation; and (4) good ADME (absorption, distribution, metabolism, and excretion) properties.

In a strict sense, the definition of leads requires the presence of at least one marketed drug, derived from that particular lead structure. However, this work includes several lead structures that served as starting points for medicinal chemistry efforts that have not necessarily reached the marketed drug status. This was merely done to illustrate a trend and not to confuse the reader. Lead structures are, after all, initial starting points in medicinal chemistry efforts that may, or may not, have been successfully optimized to reach the market. One needs to distinguish "leadlike" leads from other sources of lead structures, e.g., natural products that are high-affinity compounds (e.g., neuropeptide Y or taxol) or from "druglike" leads that are marketed structures (e.g., propranolol or nifedipine). This concept is illustrated in Figure 1.

Modern drug discovery is much more dependent on high-throughput screening (HTS) than ever before; therefore, the drug discovery paradigm is shifting focus from identifying suitable candidate drugs—which remains an essential but time-consuming goal—to identifying suitable candidate leads

* Corresponding author phone: +46 31 776 2373; fax: +46 31 776 3792; e-mail: Tudor.Oprea@astrazeneca.com.
† AstraZeneca R&D Mölndal, EST Lead Informatics.
‡ AstraZeneca R&D Charnwood.

10.1021/ci010366a CCC: $20.00
Published on Web 07/20/2001

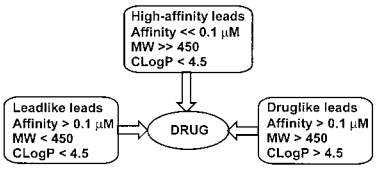

Figure 1. Classification of leads by binding affinity, modified from ref 11.

in order to maximize the cost-effectiveness and speed of the lead optimization process. There is, however, a lack of information regarding the definition of leads. We have previously discussed the design of leadlike combinatorial libraries[11] starting from a qualitative survey based upon 18 lead and drug pairs of structures, while Hann and colleagues have explored[12] the probability of discovering suitable lead structures in relationship to their molecular complexity from a theoretical perspective.

In this paper, we scrutinize further the nature of lead structures from a quantitative standpoint, using an extended dataset of 96 lead-drug pairs. Of these, 62 are leads that are not marketed as drugs (see Figure 2), and 75 are drugs that, to our knowledge, did not serve as chemical leads (see Figure 3). By looking at several properties, we are trying to address the following problem: *Can we provide an objective link between the leadlike chemical space and the druglike chemical space?* The existence of a druglike space has been previously established,[13,14] and the ability to discriminate between drugs and nondrugs has been attempted using several calculated physicochemical properties.[15,16] Therefore, we have used the same properties in an attempt to discriminate between marketed drugs and their established starting points (i.e., leads).

MATERIALS AND METHODS

The following properties, relevant to the druglike chemical space,[15] were examined using SaSA:[17] molecular weight (MW), the calculated molecular refractivity (CMR), the number of rings (RNG), the number of rotatable bonds (RTB), the number of hydrogen bond donors (HDO) and acceptors (HAC), the calculated logarithm of the *n*-octanol/water partition (CLogP), the calculated logarithm of the distribution coefficient at pH 7.4 ($LogD_{74}$), the Daylight-fingerprint druglike score (DFPS) and the property and pharmacophore features score (PPFS).

Property estimates were performed using the Daylight Toolkit[18] to evaluate MW, RNG, RTB, HDO, and HAC. The number of rings (RNG) is evaluated using the SSSR (smallest set of smallest rings) algorithm,[19] as implemented in the Daylight Toolkit. The number of rotatable bonds (RTB) is formulated in eq 1

$$\mathrm{RTB} = N_{nt} + \sum_i (n_i - 4 - \mathrm{RGB}_i - ShB_i) \qquad (1)$$

where N_{nt} is the number of nonterminal freely rotatable bonds (but single bonds observed in groups like, e.g., sulfonamides (N−S) or esters (C−O), are excluded); n_i is the number of single bonds in any nonaromatic ring *i* with 6 or more bonds; RGB_i is the number of rigid bonds in ring *i*; and ShB_i is the number of bonds shared by ring *i* with any other ring.

The number of hydrogen-bond donors and acceptors (HDO, HAC) are based on a look-up table of known fragments that are involved in hydrogen bonding, that includes only nitrogens and oxygens. Other donors, such as thiols, or acceptors, such as halides, are ignored. HDO counts all N−H and O−H fragments. Exceptions are all acids, which are considered deprotonated. Amide and amide-like (e.g., urea, sulfonamide) nitrogens as well as tertiary amines are not considered as H-bond acceptors. Since no pK_a estimator is included in this scheme, protonation states are not considered (e.g., amines are not protonated). However, HDO and HAC are counted separately, meaning that an O−H group can be both a donor and an acceptor.

Druglike scores were calculated starting with the Daylight fingerprints[18] and a number of physicochemical and pharmacophore features, as previously described.[16] Physicochemical parameters were estimated using CLogP[18] for hydrophobicity, CMR[18] for polarizability, and $LogD_{74}$ for the calculated distribution coefficient[20] at pH 7.4.

The Δ symbol indicates differences between "drugs" and "leads".

RESULTS AND DISCUSSION

Leads as Starting Points. Lead structures are often disclosed in SAR series, making it rather difficult to pinpoint a particular compound as being *the* initial chemical starting point. Despite this paucity of information, we have attempted to observe the recorded historical accounts[1−4] and used additional sources of information[21] wherever possible in order to assign a unique lead structure for each drug. While covering less than 10% of the entire therapeutical arsenal, the structures presented in Figures 2 and 3 are diverse and span over 20 clinical indications—therefore being a representative subset of the druglike and leadlike chemical spaces. Since a particular structure was rarely designated as the "lead structure" that inspired the rest of the research, it should be emphasized that these leads are based on literature searches and do not necessarily represent the opinion of the inventors of the corresponding patented drugs. Rather, a stepwise evolution in chemical structure was traceable in most of the accounts. For that reason, a single representative structure was selected, consistent with the initial set of compounds ("leads"). Figure 4 illustrates this point by describing the "pedigree" of cimetidine[22] and ranitidine,[23] starting from burimamide and N-α-guanyl histamine, that were both developed starting from histamine itself. This particular example is more fortunate, since detailed accounts exist—as both cimetidine (Tagamet) and ranitidine (Zantac) were once the largest selling prescription drugs in the world. Such information is, unfortunately, much less prominent for the vast majority of therapeutic agents. To complicate the analysis, a drug can have 1 or more leads, a lead can be a drug, and a lead can be developed into several drugs. For example, mifepristone was discovered[24] starting from progesterone, a natural hormone and marketed drug, and from RU2323, which failed in clinical trials—see Figure 5. In another example, thiazesim—an antidepressant, and oxazepam—a tranquilizer, both marketed drugs, served as lead structures[25] for diltiazem, a calcium-channel antagonist—see Figure 6. Another difficulty encountered when compiling this dataset is related to its dynamic nature. The list of drugs that do not serve as leads and the list of leads that are not marketed as

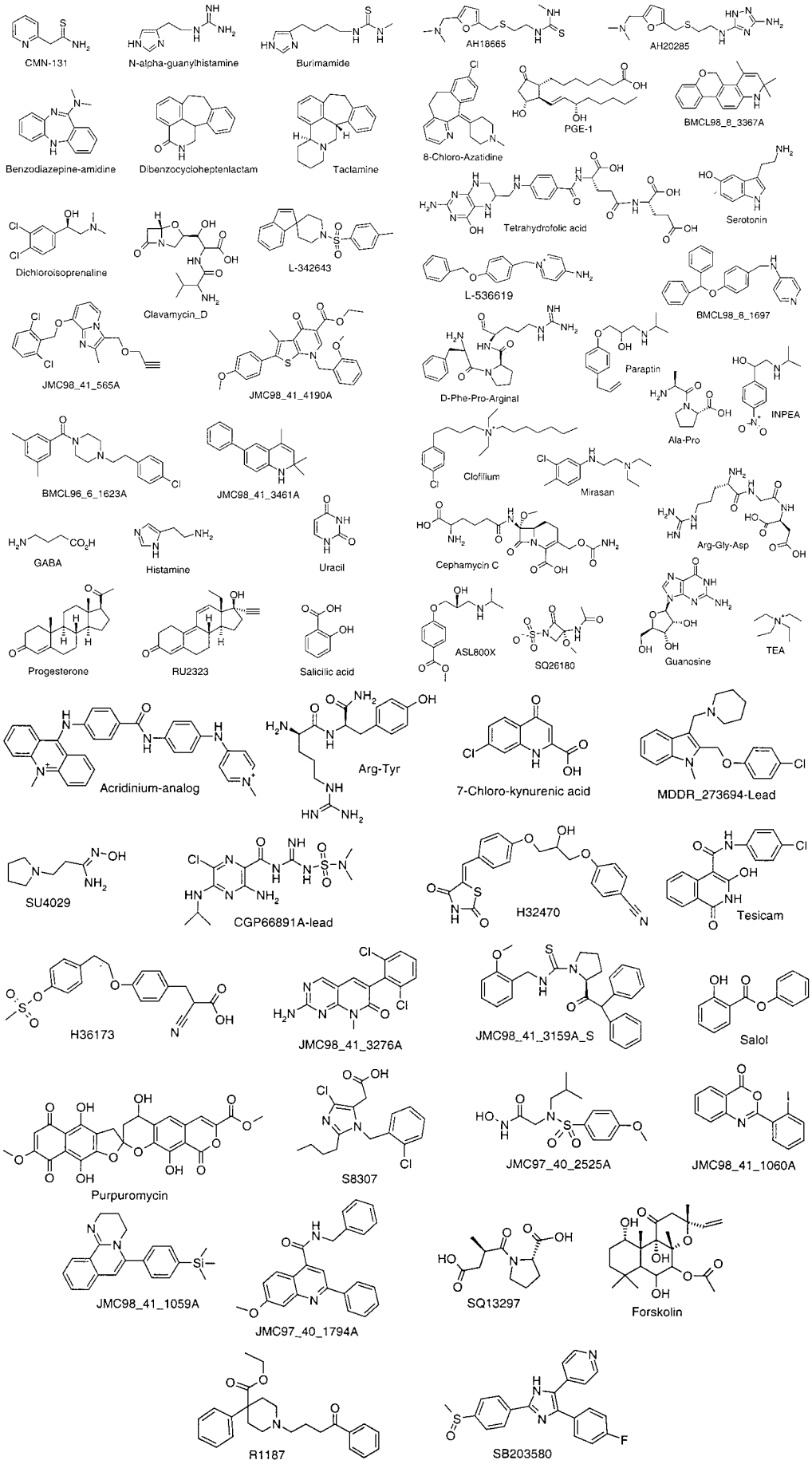

Figure 2. Chemical structures of the 62 leads used in this study.

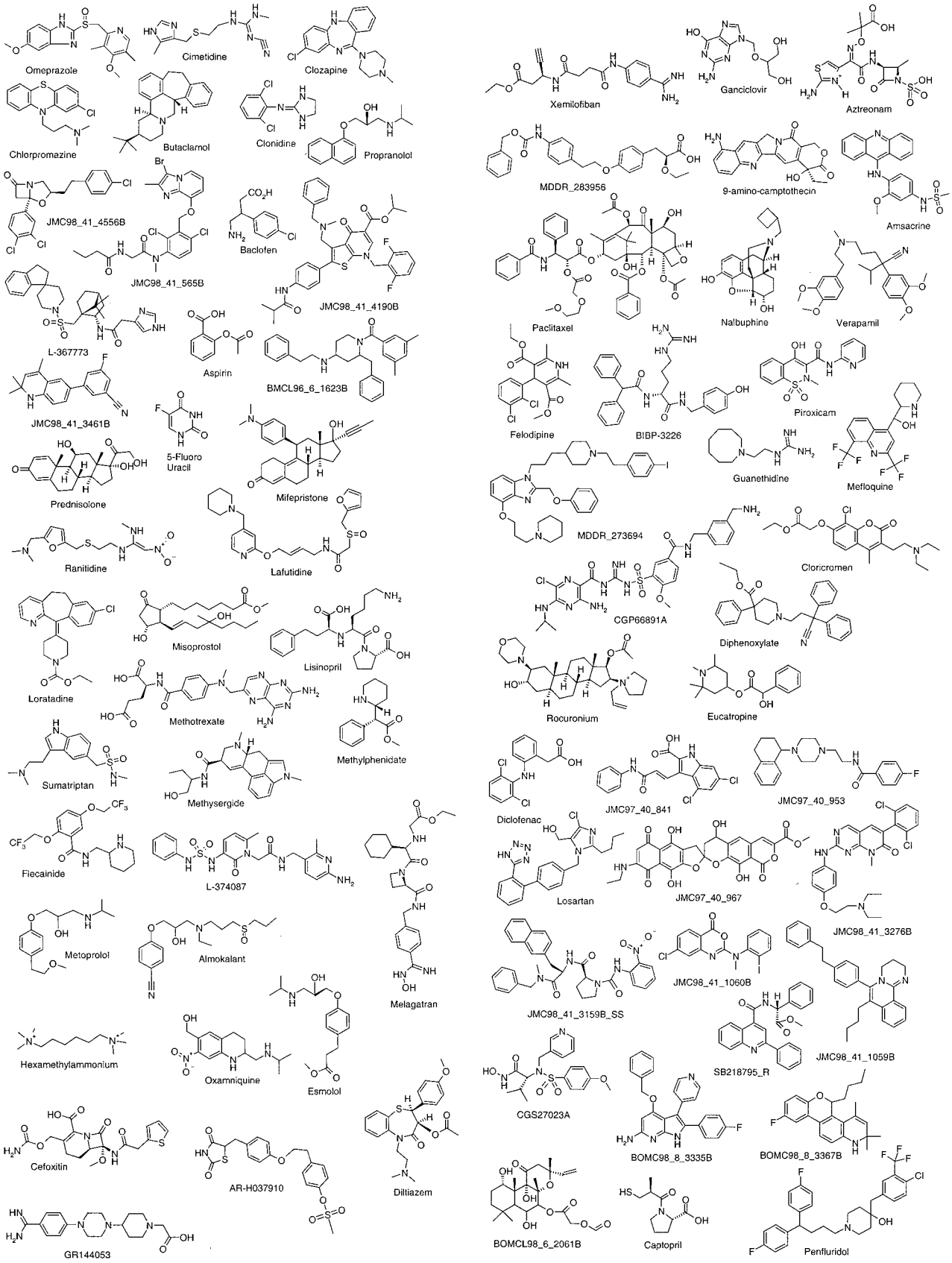

Figure 3. Chemical structures of the 75 drugs used in this study.

drugs may change as new compounds reach marketed drug status or as new patents are disclosed, respectively. For example, omeprazole (Losec) was the lead for esomeprasole (its S enantiomer), which became a marketed drug (Nexium) during the past few months. Technically, this places omeprazole in the "both lead and drug" category, meaning that it should have been removed from the initial list of 75 drugs shown in Figure 3.

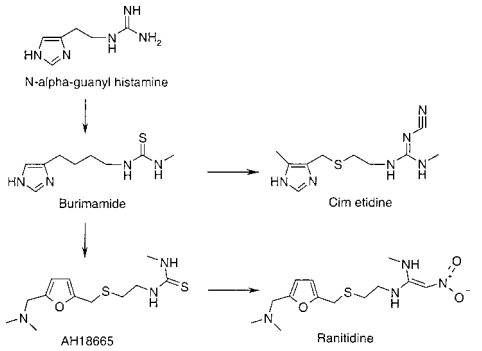

Figure 4. Derivation of cimetidine and ranitidine, starting from lead structures.

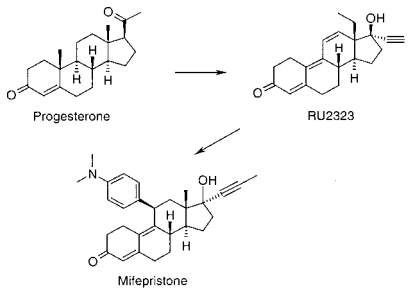

Figure 5. Derivation of mifepristone, starting from lead structures.

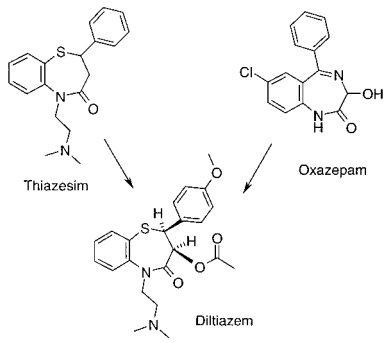

Figure 6. Derivation of diltiazem, starting from lead structures.

Table 1. Property Analysis for 67 Unique Drug-Lead Pairs[a]

	ΔMW	ΔRNG	ΔRTB	ΔHDO	ΔHAC	ΔCLogP
mean	78.97	0.55	1.90	−0.18	0.45	1.25
SE	10.13	0.13	0.42	0.20	0.23	0.31
median	69.88	0.00	2.00	0.00	1.00	0.67
SD	82.95	1.06	3.41	1.62	1.88	2.58
range	506.38	7.00	24.00	8.00	10.00	15.65
minimum	−120.11	−2.00	−12.00	−5.00	−5.00	−5.95
maximum	386.28	5.00	12.00	3.00	5.00	9.70

[a] Property values calculated for the lead structures were substracted from the corresponding values calculated for drugs. For more than 1:1 correspondence, only larger structures were considered.

Leads vs Drugs: Is There a Difference? To address this question, we performed two types of analyses on the same dataset. The first one is restricted to those examples where a 1:1 correspondence between drugs and leads could be accounted for. This served as the basis for the lead-drug paired analysis summarized in Table 1, where property values calculated for the lead structures were subtracted from the corresponding values calculated for drugs. This analysis, however, could not appropriately reflect situations where multiple links between leads and drugs exist, for example, progesterone-RU2323-mifepristone or oxazepam-thiazesim-diltiazem (see Figures 5 and 6). The second analysis, therefore, is a comparative profile analysis of the two major categories: leads that lack any therapeutic use and marketed drugs that have not been recorded as leads for further drug discovery. Property distribution profiles for these two categories are shown in Figure 7.

The 1:1 pairwise comparison summarized in Table 1 indicates that the following property alterations occur, when going from lead to drug: A definite increase in molecular weight (70−79 daltons) and lipophilicity (1.25 log units according to the average, 0.67 log units according to the median), two additional rotatable bonds, a relatively small increase in the number of hydrogen-bond acceptors (1 according to the median, 0.45 according to the average), almost no change in the number of rings (0.55 according to the average, 0 according to the median), and no change in the number of hydrogen-bond donors. Similar property alterations are observed in the property distribution profile analysis (Figure 7). The following differences were observed between the medians of drugs and leads: $\Delta MW = 69$; $\Delta CMR = 1.8$; $\Delta RNG = \Delta HAC = 1$; $\Delta RTB = 2$; $\Delta CLogP = 0.43$; $\Delta LogD_{74} = 0.97$; $\Delta HDO = 0$; $\Delta DFPS = 0.15$; $\Delta PPFS = 0.12$, while the difference between averages indicates a similar trend: $\Delta MW = 89$; $\Delta CMR = 2.3$; $\Delta CLogP = 1.16$; $\Delta LogD_{74} = 0.97$; $\Delta RNG = \Delta HAC = 1$; $\Delta RTB = 2$; $\Delta HDO = 0.2$; $\Delta DFPS = 0.15$; $\Delta PPFS = 0.9$. While these results are not statistically significant (as indicated by the descriptive statistics in Table 1), similar trends were observed by Hann and colleagues[12] in their Table 2 property comparison of Sneader leads to Sneader drugs. The difference between leads and drugs can be, therefore, expressed as follows: Lead structures exhibit, on average, less molecular complexity (less molecular weight, less number of rings and rotatable bonds), are less hydrophobic (lower CLogP and $LogD_{74}$) and have lower polarizability (less CMR), and are, not surprisingly, less druglike (i.e., have lower druglike scores).

Practical Use of the Leadlike Space Concept. The seminal paper by Lipinski and colleagues[26] alerted the drug discovery community about the importance of restricting small molecule synthesis to the druglike space defined by LogP, MW, H-bond donors, and H-bond acceptors. This work emerged as a post-factum analysis of the early (1994) results of HTS and combinatorial chemistry at Pfizer, that were fairly disappointing, since most of the hits were high-MW and high-LogP compounds, not easily suitable for optimization. This paper enhanced the awareness of the medicinal chemistry community regarding the existence of the druglike space and has indirectly spawned this work as well. Many library design programs, based on combinatorial chemistry or compound acquisition, subsequently included filters for the "Lipinski rule of five". Should the aim of such programs be to identify drugs (not leads), then the use of such filters is appropriate. However, most lead discovery projects applied these filters *ad litteram*, i.e., using MW < 500 and CLogP < 5, regardless of the fact that these values had been obtained from analyzing **drugs**, not leads. In our previous[11] work, we highlighted that the outcome of many HTS campaigns were only micromolar hits. Having been filtered according to the drug-based "rule of five", these leads did not prove to be easily amenable to traditional medicinal chemistry optimization, which would take their property profile outside the "rule of five" range.

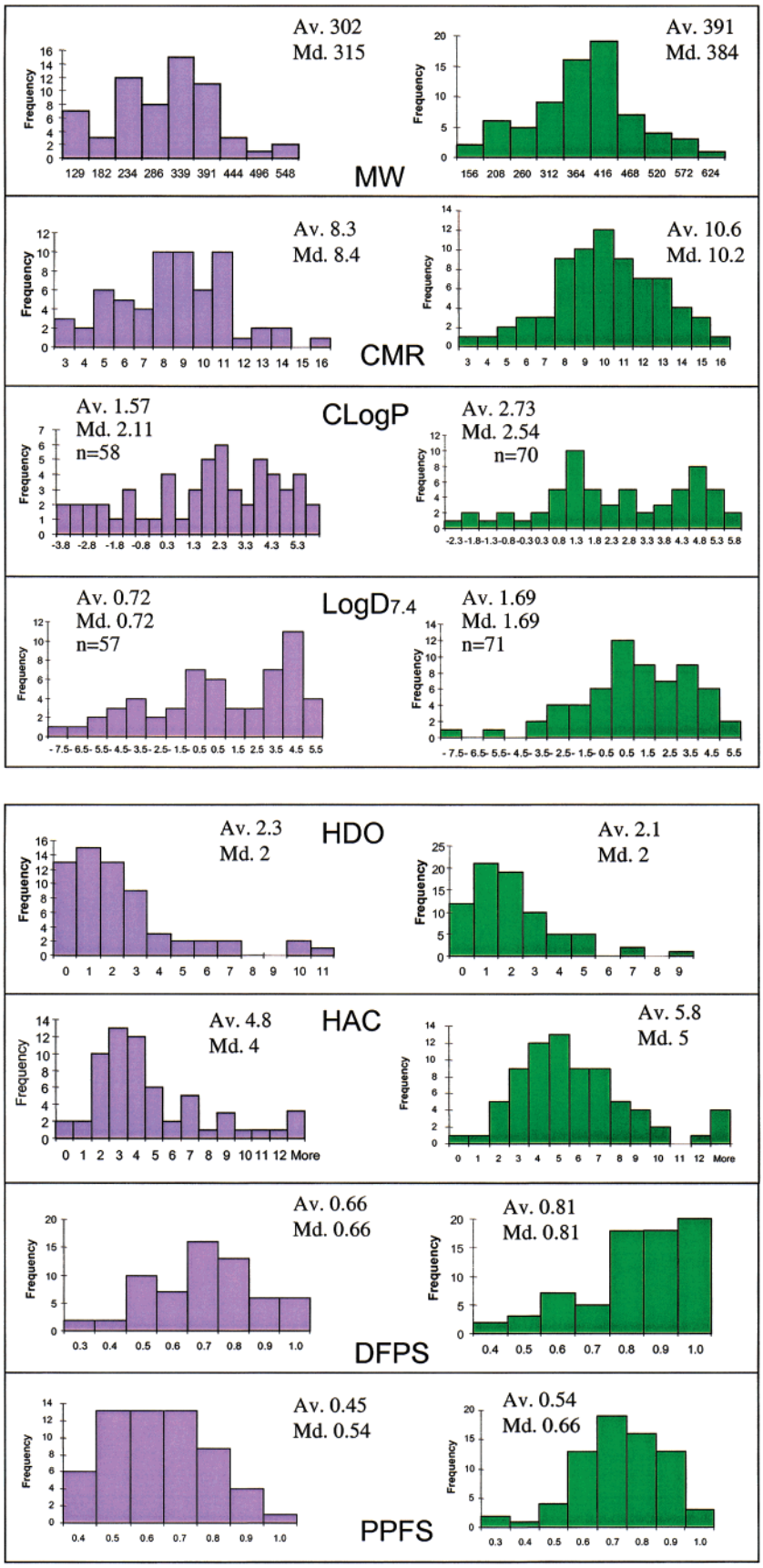

Figure 7. Property distribution profiles for 62 leads (purple, left), compared to 75 drugs (green, right). Abbreviations: Av. is the average, Md. is the median, and n is the sample size whenever it differs from 62 and 75, respectively.

Based on our initial set of 18 lead-drug pairs, we further suggested[11] that leadlike libraries should be designed with lower MW and lower LogP profiles, as opposed to druglike libraries. In this light, we also questioned the role of some combinatorial technologies that simply concatenate several monomers using multicomponent reactions or that use several

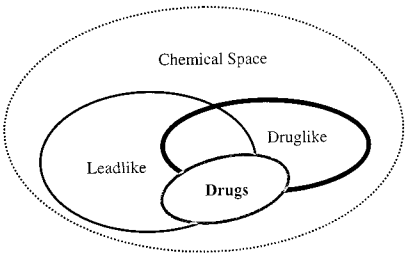

Figure 8. Venn diagram of the medicinal chemistry space related to drug discovery.

steps including split-and-mix protocols. Using our current level of awareness regarding the definition of the leadlike space, we can further suggest the following: When designing leadlike combinatorial libraries, care should be exercised not to exceed the following property values: 450 Dalton in MW, CLogP lower than +4.5 but higher than −3.5 (conversely, $LogD_{74}$ between −4 and +4), no more than 4 rings, no more than 10 nonterminal single bonds, no more than 5 hydrogen-bond donors, and no more than 8 hydrogen-bond acceptors. This information should be used in the design of novel combinatorial libraries aimed at lead identification, keeping in mind that the nature of the target may demand different property profiles. For example, a drug discovery effort aimed at identifying agents active in the central nervous system (that must pass the blood-brain-barrier) would require a different profile in terms of LogP and hydrogen-bond donors and acceptors, compared to a project aimed at identifying antimicrobial agents active in the urinary tract.

CONCLUSIONS

In this paper, we have attempted to provide an objective link between the leadlike chemical space and the druglike chemical space. While this link appears to exist, it is obscured by those drugs that are, in our current understanding, not representative of the "medicinal chemistry druglike space". For example, lithium carbonate, gold (e.g., in aurothhioglucose or gold sodium thiomalate), cis-platin, bis-phosphonates and foscarnet sodium, compounds derived from natural products (taxol, cyclosporin, peptide antibiotics such as vancomycin), and proteins (e.g., interferon-beta and erythropoietin) are currently marketed as drugs but are not representatives for the types of compounds that are expected to emerge from medicinal chemistry efforts. In the same vein, we would not categorize these compounds as being "leadlike" either. The importance of such compounds in the therapeutic arsenal is beyond question, and in fact nine out of the top 20 best-selling drugs are either natural products or semisynthetic products. While the pharmaceutical industry continues to have major efforts aimed at identifying such compounds (e.g., via natural product screening), an increasingly larger effort is aimed at identifying small molecules as drug candidates. These efforts appear to be more suitable from both the R&D perspective (control over the modes of action, ADME properties, formulation, etc.) and from the clinical perspective (drug−drug interactions, interpatient variability, etc.) as well as from the production perspective (cost of synthesis, ecotoxicity aspects, etc.). The chemical space of interest, and its inter-relatedness, is schematically represented in Figure 8.

In this report, we establish the existence of a "medicinal chemistry leadlike space", wherein leadlike structures are, on the average, less intricate in terms of molecular weight, molecular complexity (number of rings, number of rotatable bonds), polarizability, and lipophilicity. This further indicates that the "Lipinski rule of five" needs to be applied in a context-dependent manner, related to the nature of the biological target and to the pharmacokinetic profile imposed by clinical circumstances or by the desired mode of administration. This information should be used in the design of novel combinatorial libraries that are aimed at lead discovery.[27]

ACKNOWLEDGMENT

We thank Drs. Ingemar Nilsson and Magnus Björsne (AstraZeneca R&D Mölndal) and Prof. Bertil Samuelsson (Medivir AB, Stockholm) for invaluable comments. Dr. Mike Hann (Glaxo Wellcome Medicines Research Centre, Stevenage) is gratefully acknowledged for providing the work described in ref 12 prior to publication.

Supporting Information Available: The chemical structures of leads and drugs discussed in this paper. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) *Chronicles of Drug Discovery*; Bindra, J. S., Lednicer, D., Eds.; Wiley-Interscience: 1982; Vol. 1, ISBN 0-471-06516-1.
(2) *Chronicles of Drug Discovery*; Bindra, J. S., Lednicer, D., Eds.; Wiley-Interscience: 1983; Vol. 2, ISBN 0-471-89135-5.
(3) *Chronicles of Drug Discovery*; Lednicer, D., Ed.; ACS, Washington, 1993; Vol. 3, ISBN 0-8412-2733-0.
(4) DeStevens, G. Serendipity and structured research in drug discovery. *Prog. Drug. Res.* **1986**, *30*, 189−203.
(5) *Scifinder 5.0*, Copyright 1999; American Chemical Society: http://www.acs.org.
(6) Drews, J. Drug discovery: A historical perspective. *Science* **2000**, *287*, 1960−1964.
(7) Furka, A. Combinatorial Chemistry. *Chem. Intell.* **1999**, *5*, 22−27.
(8) Hirsch, P. F.; Anderson, J. J. B. Calcitonin. *Calcium Phosphorus Health Dis.* **1996**, 197−206.
(9) Ahn, S. H. What is medicinal chemistry. *Hwahak Sekye* **1994**, *34*, 109−111.
(10) Ramsden, C. A. Medicinal chemistry: A multidisciplinary science. *Sch. Sci. Rev.* **1994**, *75*, 49−58.
(11) Teague, S. J.; Davis, A. M.; Leeson, P. D.; Oprea, T. I. The design of leadlike combinatorial libraries. *Angew. Chem., Int. Ed. Engl.* **1999**, *38*, 3743−3748; German version: *Angew. Chem.* **1999**, *111*, 3962−3967.
(12) Hann, M. M.; Leach, A. R.; Harper, G. Molecular Complexity and its impact on the probability of finding leads for drug discovery. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 856−864.
(13) Sadowski, J.; Kubinyi, H. A scoring scheme for discriminating between drugs and nondrugs. *J. Med. Chem.* **1998**, *41*, 3325−3329.
(14) Ajay; Walters, W. P.; Murcko, M. A. Can we learn to distinguish between "drug-like" and "nondrug-like" molecules? *J. Med. Chem.* **1998**, *41*, 3314−3324.
(15) Oprea, T. I. Property distribution of drug-related chemical databases. *J. Comput-Aided Mol. Design* **2000**, *14*, 251−264.
(16) Oprea, T. I.; Gottfries, J.; Sherbuhin, V.; Svensson, P.; Kühler, T. C. Chemical information management in drug discovery: Optimizing the computational and combinatorial chemistry interfaces. *J. Mol. Graphics Modelling* **2000**, *18*, 512−524.
(17) Olsson, T.; Sherbukhin, V. *Synthesis and Structure Administration (SaSA)*; 1997−2001; AstraZeneca: http://www.astrazeneca.com.
(18) The Daylight Toolkit and the physicochemical property calculations CMR and CLogP are available from Daylight Chemical Information Systems, Santa Fe, NM. http://www.daylight.com.
(19) Downs, G. M. Ring Perception. In *Encyclopedia of Computational Chemistry*; von Ragué Schleyer, P., Ed.; Wiley: New York, 1998; Vol. 4, pp 2509−2515.

(20) ACDLogP is available from ACD Labs, Toronto, Canada, http://www.acdlabs.com/.

(21) Several structures in Figures 2 and 3 were derived from the literature and are marked as such: For example, the compound marked JMC98_41_4556B in Figure 3 is from Aicher T. D., et al. *J. Med. Chem.* **1998**, *41*, 4556−4566; in another example, the compound marked BMCL96_6_1623A is from Ofner S., et al., *Bioorg. Med. Chem. Lett.* **1996**, *6*, 1623−1628; another example is MDDR_273694, which represents the identifier of that structure in the MDDR (MDL Drug Data Report) database, a collection of compounds patented for pharmaceutical use available from MDL (http://www.mdli.com).

(22) Ganellin, C. R. Cimetidine. In *Chronicles of Drug Discovery*; Bindra, J. S., Lednicer, D., Eds.; Wiley-Interscience: 1982, Vol. 1, pp 1−38.

(23) Bradshaw, J. Ranitidine. In *Chronicles of Drug Discovery*; Lednicer, D., Ed.; ACS: Washington DC, 1993; Vol. 3, pp 45−81.

(24) Teutsch, G.; Deraedt, R.; Philibert, D. Mifepristone. In *Chronicles of Drug Discovery*; Lednicer, D., Ed.; ACS: Washington DC, 1993; Vol. 3, pp 1−43.

(25) Inoue, H.; Nagao, T. Diltiazem. In *Chronicles of Drug Discovery*; Lednicer, D., Ed.; ACS: Washington DC, 1993; Vol. 3, pp 207−238.

(26) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Deliv. Reviews* **1997**, *23*, 3−25.

(27) Oprea, T. I.; Davis, A. M.; Teague, S. J.; Leeson, P. D. Is there a difference between leads and drugs? A historical perspective. 13th European QSAR Symposium, Düsseldorf, Germany, August 2000, poster.

CI010366A