Journal of Computer-Aided Molecular Design, **16:** 325–334, 2002.
KLUWER/ESCOM

Current trends in lead discovery: Are we looking for the appropriate properties?

Tudor I. Oprea*
Dept. Biochemistry and Molecular Biology and Office of Biocomputing, UNM School of Medicine, 915 Camino de Salud NE, Albuquerque, NM 87131-5166, USA

Key words: database filtering, drug-likeness, drug research, hydrophobicity, lead-likeness, property distribution, 'rule of 5' test, solubility

Summary

The new drug discovery paradigm is based on high-throughput technologies, both with respect to synthesis and screening. The progression HTS hits → lead series → candidate drug → marketed drug appears to indicate that the probability of reaching launched status is one in a million. This has shifted the focus from good quality candidate drugs to good quality leads. We examined the current trends in lead discovery by comparing MW (molecular weight), LogP (octanol/water partition coefficient, estimated by Kowwin [17]) and LogSw (intrinsic water solubility, estimated by Wskowwin [18]) for the following categories: 62 leads and 75 drugs [11]; compounds in the development phase (I, II, III and launched), as indexed in MDDR; and compounds indexed in medicinal chemistry journals [ref. 20], categorized according to their biological activity. Comparing the distribution of the above properties, the 62 lead structures show the lowest median with respect to MW (smaller) and LogP (less hydrophobic), and the highest median with respect to LogSw (more soluble). By contrast, over 50% of the medicinal chemistry compounds with activities above 1 nanomolar have MW $>$ 425, LogP $>$ 4.25 and LogSw $<$ -4.75, indicating that the reported active compounds are larger, more hydrophobic and less soluble when compared to time-tested quality leads. In the MDDR set, a progressive constraint to reduce MW and LogP, and to increase LogSw, can be observed when examining trends in the developmental sequence: phase I, II, III and launched drugs. These trends indicate that other properties besides binding affinity, e.g., solubility and hydrophobicity, need to be considered when choosing the appropriate leads.

The role of the lead structure in the current drug discovery paradigm

The 'megabrand' marketing concept advocates the focus on single products that yield significant income, e.g., in excess of 500 million USD [1, 2], rather than marketing a variety of products to smaller markets. This concept, based on marketing research [3], has a stop/go influence on the type of compounds that are submitted for drug approval to regulatory agencies, and is part of the decision making process [4] in most of the major pharmaceutical companies. The task of producing good quality compounds has to face two additional constraints: Reduced R&D spending, and shorter time from idea to marketed drug. To support the current drug discovery paradigm, combinatorial chemistry, also termed high-throughput organic synthesis or multiple parallel synthesis (MPS), has emerged [5] in the early '90s as the key methodology to provide large numbers (well above 10^5) of novel, perhaps diverse [6], chemicals. In the same time, a number of the screening techniques, amenable to automation (e.g., fluorescence-based, or scintillation-based, assays), collectively designated as high throughput screening (HTS), have become available to test the increasing number of chemicals [7]. The pharmaceutical industry as a whole embraced this novel paradigm for drug discovery, based on different flavors of MPS/HTS, as solutions to the time- and cost-cutting problem, even though no guarantees

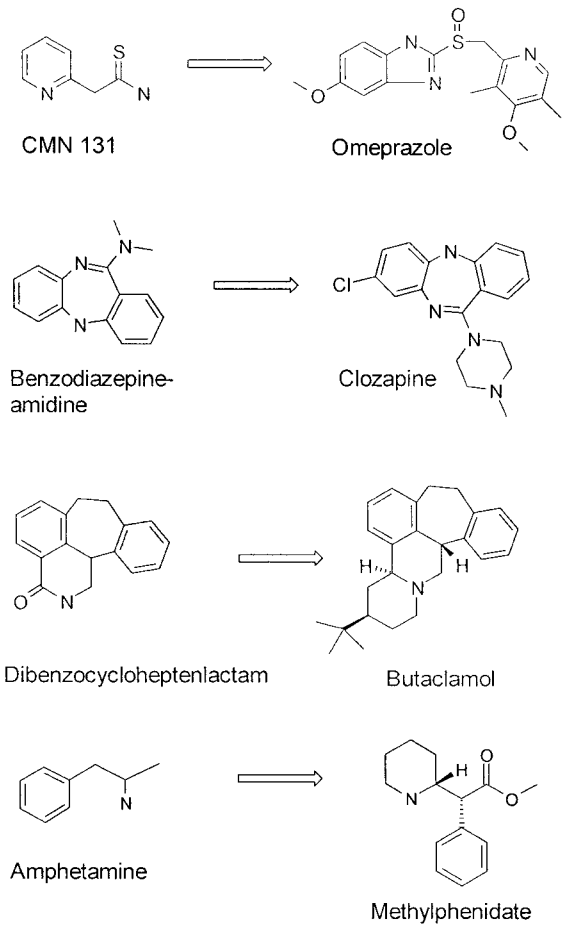


Figure 1. Derivation of Omeprazole, Clozapine, Butaclamol and Methylphenidate, starting from lead structures.

for success were provided [3]. While the impact of MPS/HTS has yet to be established, the drug discovery rate did not improve in the past decade. In fact, the number of new chemical entities (NCEs) has remained rather constant – an average of 37 NCEs per year – between 1990 and 2000, even though the number of compounds being initially screened has increased by several orders of magnitude [8].

The current drug discovery paradigm is based on the following sequence:

- *HTS hits* designate compounds that show activity during HTS that are retested to confirm activity and structure. Single-dose HTS has a success rate below 0.1% – in other words, less than 1 in 1000 compounds from a randomly selected library show activity on any given target. The low success rate is, to a certain extent, due to false positives [9] that interfere with the biological assay (e.g., fluorescent compounds, dyes, or chemically reactive species).
- When multiple hits are identified in the same chemical family, this forms the basis of identifying '*lead series*', i.e., structures that are amenable for further chemistry optimization and exhibit favorable patent situation [10, 11]. Lead compounds are further optimized in order to derive the appropriate combination of potency, pharmacokinetic properties, toxicity etc., as well as good activity in animal models.
- When all criteria are met, the compound reaches the *candidate drug* status. Typically, one in 10,000 screened compounds reaches this stage [12]. After approval from regulatory agencies, candidate drugs are tested in humans. If all three phases of clinical trials prove successful – which only one in 10 candidate drugs [13] does – the compound is approved and becomes a *marketed drug*.

In summary, approximately one in a million compounds (or more), from those initially tested via HTS, has the probability to reach the market – assuming no target-bias was available at the onset of the project. Thus, the progression HTS hits → lead series → candidate drug → marketed drug, has shifted the focus from good quality candidate drugs, to good quality leads.

Therefore, choosing the lead structures with appropriate qualities has become an important focus in preclinical drug research activities. The Lipinski rule of five, derived from marketed drugs [14], sets the upper (ninety percentile) limits for molecular weight (500 daltons), CLogP, the calculated logarithm of the octanol/water partition coefficient (5), the number of hydrogen bond donors (5) and the sum of hydrogen bond acceptors (10), respectively. We recently suggested these values to be more restrictive when applied to leads [15], as leads tend to exhibit lower molecular complexity [16]. Our own analysis [11] indicates that, on the average, leads have 100 daltons, 1 ring, 2 flexible bonds and 1 LogP/LogD unit less, when compared to drugs. Thus, the chemical space for leads is more restrictive than the chemical space for drugs.

Properties and datasets

To examine the current trends in lead discovery, we compared the available known leads and drugs to compounds of interest for pharmaceutical development, as well as to compounds resulting from medicinal

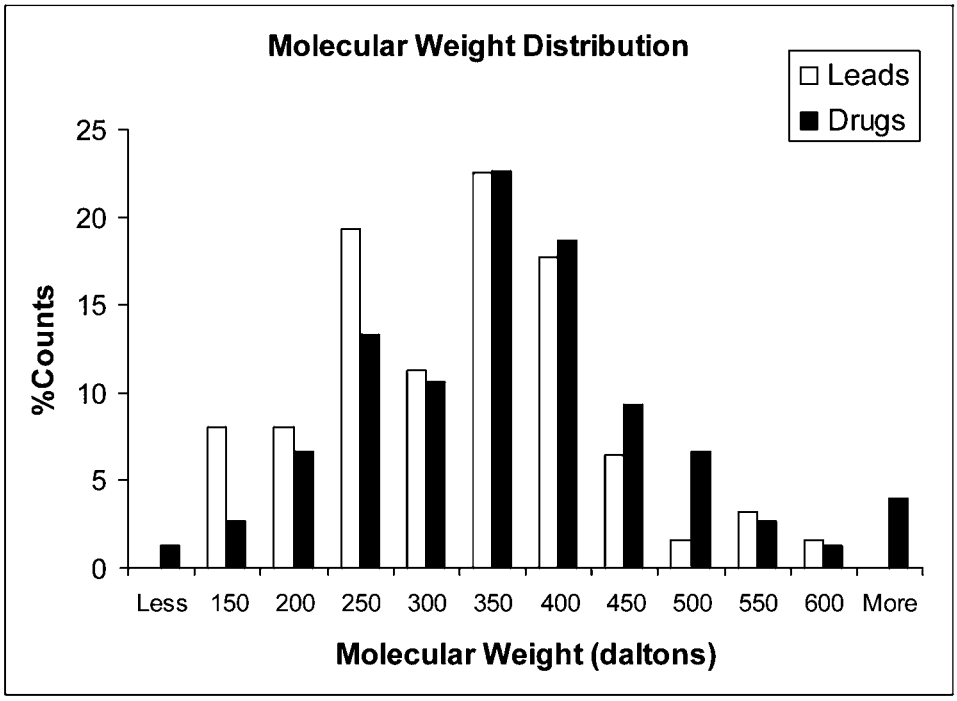
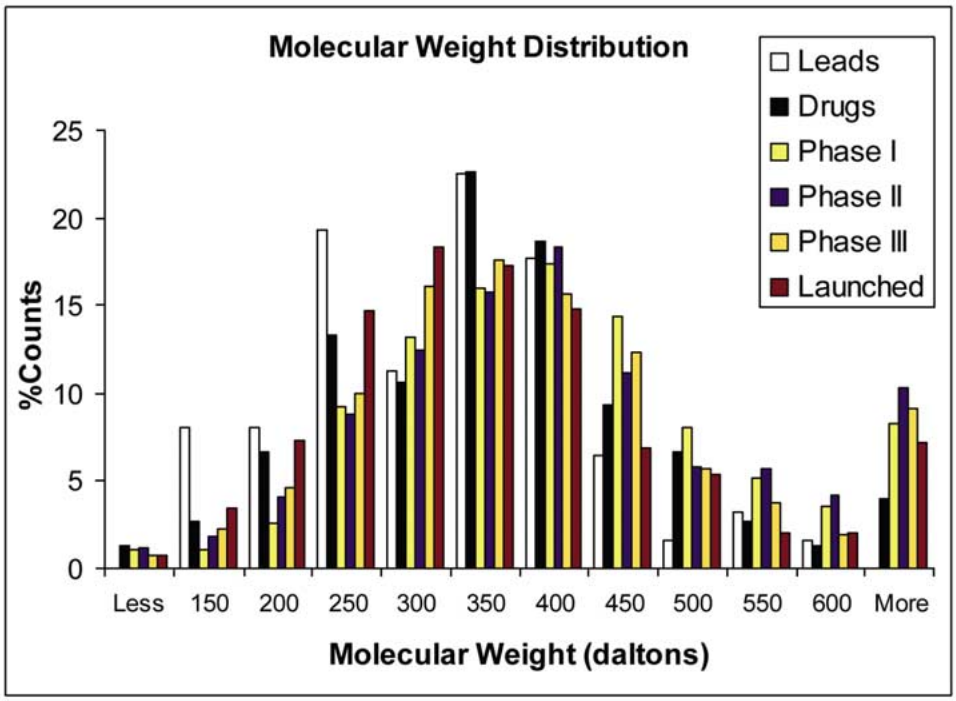


Figure 2. Molecular weight distribution for leads and drugs (a), compared to compounds in phase I-III clinical trials and launched drugs from MDDR (b), and with low-activity and high-activity compounds from JMC (c).

chemistry efforts, following the distribution of MW (molecular weight), LogP – as estimated by Kowwin [17] and LogSw, the logarithm of the intrinsic water solubility – as estimated with Wskowwin [18]. The following datasets were investigated:

- The 62 leads and 75 drugs reported in our previous work [11]
- Compounds in pharmaceutical development, extracted from the MDDR (MACCS Drug Data Report) database [19], categorized according to their clinical testing phase, in the following manner:
  - 985 launched drugs;
  - 260 compounds in phase III clinical trials;
  - 826 compounds in phase II clinical trials;
  - 564 compounds in phase I clinical trials;
- Compounds recently reported in medicinal chemistry studies, extracted from papers published in mainstream journals [20], categorized according to their disclosed biological activity (IC_{50}, K_i, ED_{50}, etc.), in the following manner:
  - 2990 compounds for which the biological activity is above 1 μM, or below 6 units on the –log (activity) scale, on all of the documented assays ('below 6 only');
  - 2266 compounds for which the biological activity is above 1 μM, or below 6 units on the –log (activity) scale, in one of the documented assays ('below 6 mul'); these did not include the 2990 compounds above;
  - 403 compounds for which the biological activity is below 1 nM, or above 9 units on the –log (activity) scale, on all of the documented assays ('above 9 only');
  - 605 compounds for which the biological activity is below 1 nM, or above 9 units on the –log (activity) scale, in one of the documented assays ('above 9 only'); these did not include the 403 compounds above.

Among the 11965 structures surveyed [20], 6435 were tested as receptor antagonists (or blockers), and 4304 as enzyme inhibitors; 1085 antagonists and 1625 inhibitors are in the 'below 6 only' category, while 1484 antagonists and 727 inhibitors are in the 'below 6 mul' category. The 'above 9 only' category includes 166 enzyme inhibitors and 227 receptor antagonists, whereas 74 inhibitors and 538 antagonists are in the 'above 9 mul' category. Of the 5256 compounds with micromolar activity in at least one assay, 44.75% were tested as enzyme inhibitors and 48.88% as re-



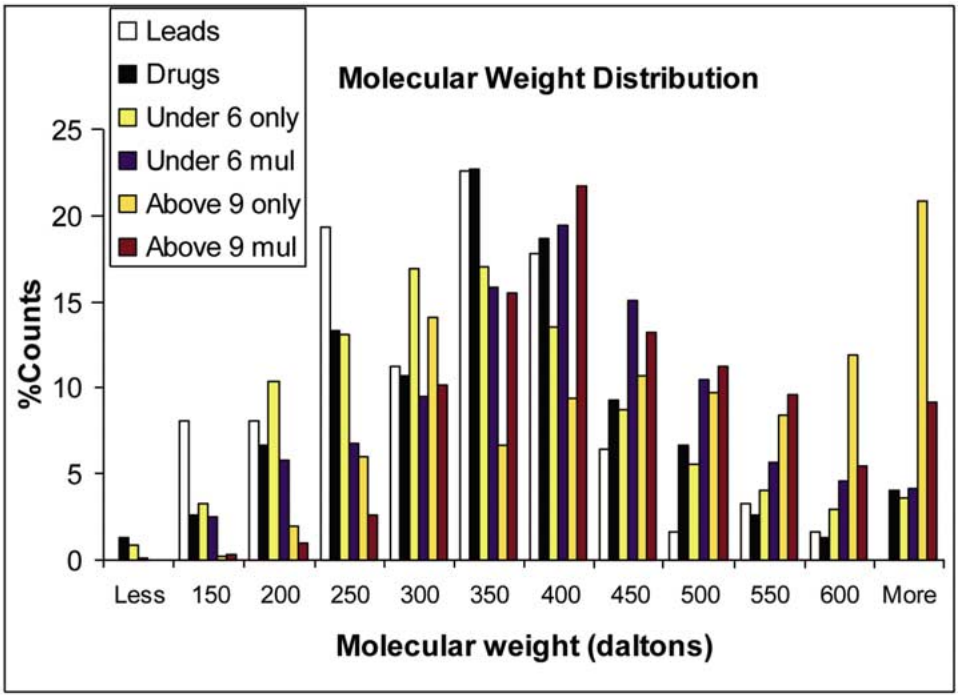


Figure 2. Continued.

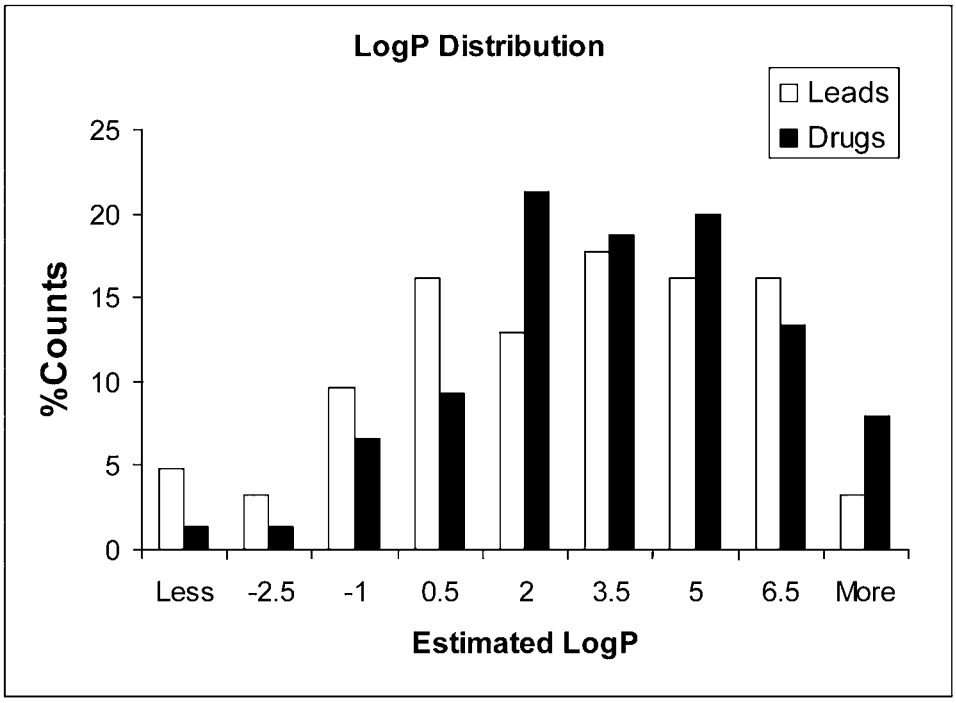


Figure 3. LogP distribution for leads and drugs (a), compared to compounds in phase I-III clinical trials and launched drugs from MDDR (b), and with low-activity and high-activity compounds from JMC (c).

ceptor antagonists. Of the 1008 compounds with tested nanomolar activity, 23.8% were enzyme inhibitors and 75.9% were receptor antagonists.

Current trends in lead discovery

Of the three investigated properties, LogSw is related to MW and LogP by the following equation [18]:

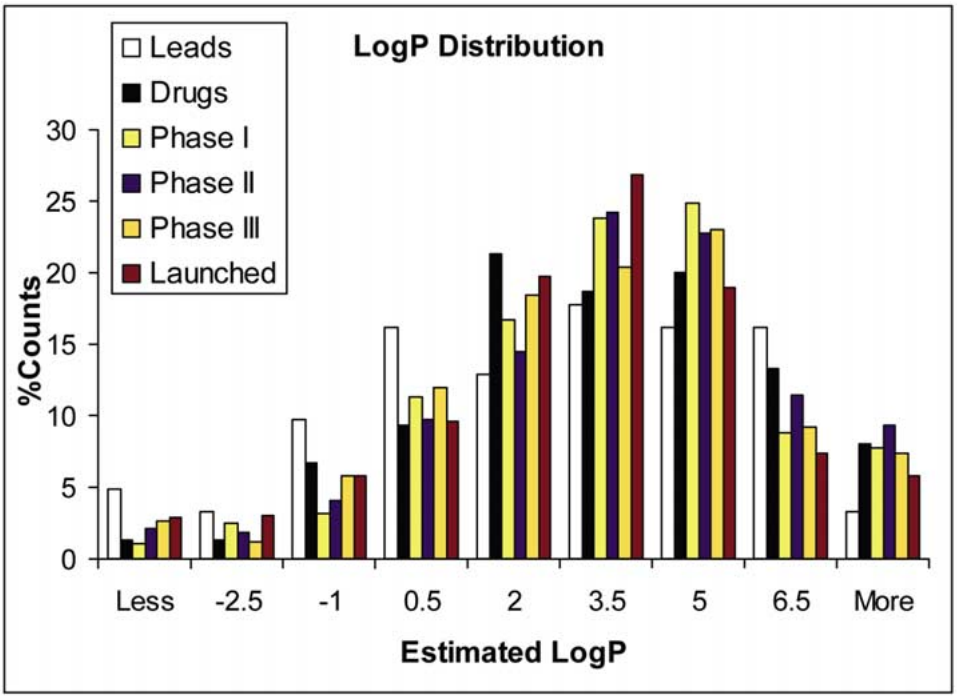
$$\text{LogSw (mol/L)} = 0.796 - 0.854\,\text{LogP} - \\ -0.00728\,\text{MW} + \text{Corrections}$$

The correction terms are applied to 15 structure types (e.g. alcohols, acids, selected phenols, nitros, amines, alkyl pyridines, amino acids, multi-nitrogen types, etc) – depending on the available melting points – as described by Meylan, Howard and Boethling [18].

Lead structures are, as previously noted, smaller and usually more polar than the final structures of the marketed drugs – concept which is illustrated in Figure 1. This trend is observed for our lead:drug dataset in terms of MW, LogP and LogSw (see Figures 2.a., 3.a. and 4.a., respectively). From these figures, it can be concluded that leads (the white bars) appear to be smaller, less hydrophobic and more soluble than their corresponding drugs (the black bars). For the purpose of comparison, we plotted the distribution of these three properties for both leads and drugs, against the set of compounds in pharmaceutical development (Figures 2.b., 3.b. and 4.b.), as well as against the compounds recently reported in medicinal chemistry studies (Figures 2.c., 3.c. and 4.c.), respectively.

Lead structures exhibit lower molecular weight, lower LogP and higher solubility, not only when compared with their corresponding drugs, but also when compared to phase I-III or launched compounds. The 75 drugs derived from the above leads exhibit, in fact, similar trends to the launched category, whereas the phase I-III compounds appear to have somewhat higher MW (in particular in the 450 daltons bin); otherwise, the distribution for LogP and LogSw follows that of launched drugs.

High-activity compounds (above 1 nM) exhibit higher MW, higher LogP and lower LogSw when compared to the lead:drug pairs and to the MDDR compounds. In fact, 61.6% of the 'above 9 only' compounds and 48.7% of the 'above 9 mul' compounds have MW above 425 daltons, compared to 24.8% of the 'below 6 only' and 40% of the 'below 6 mul' compounds, respectively. In the same category fall



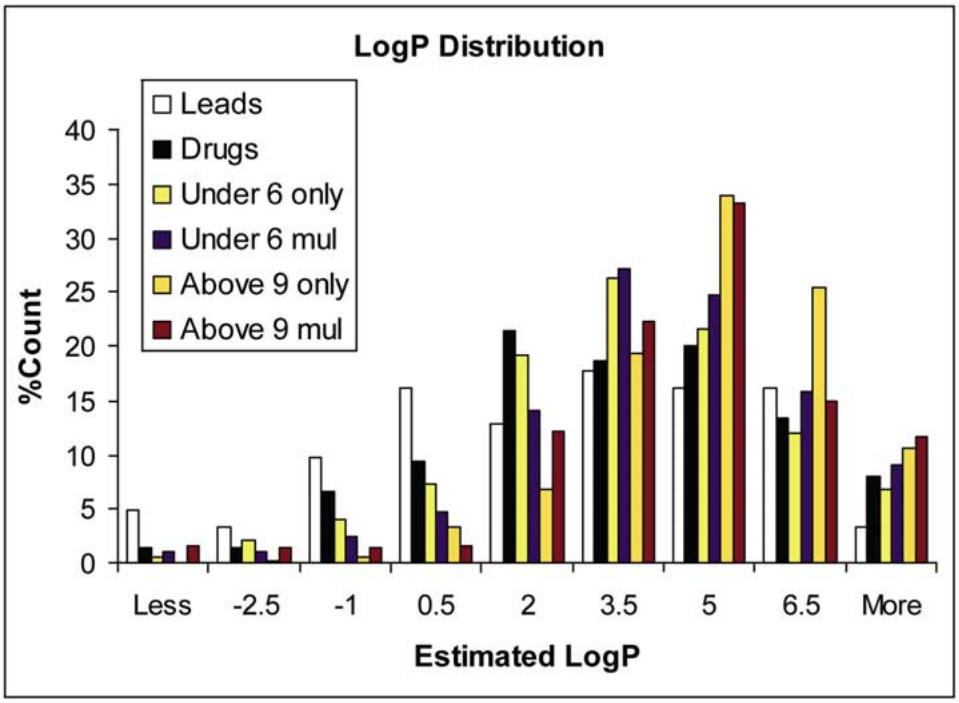


Figure 3. Continued.

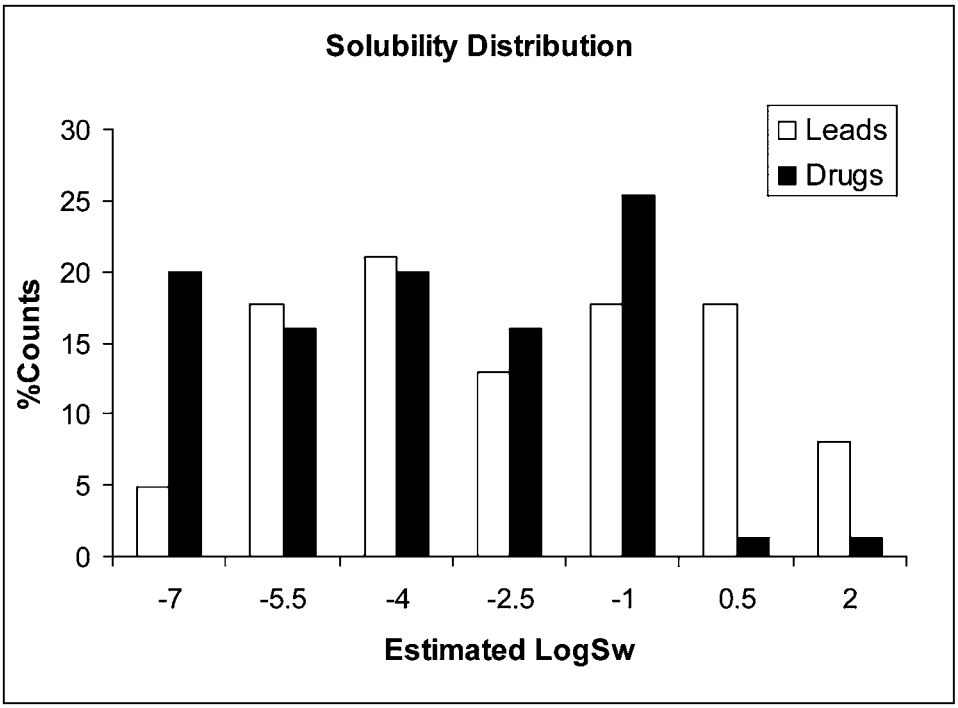
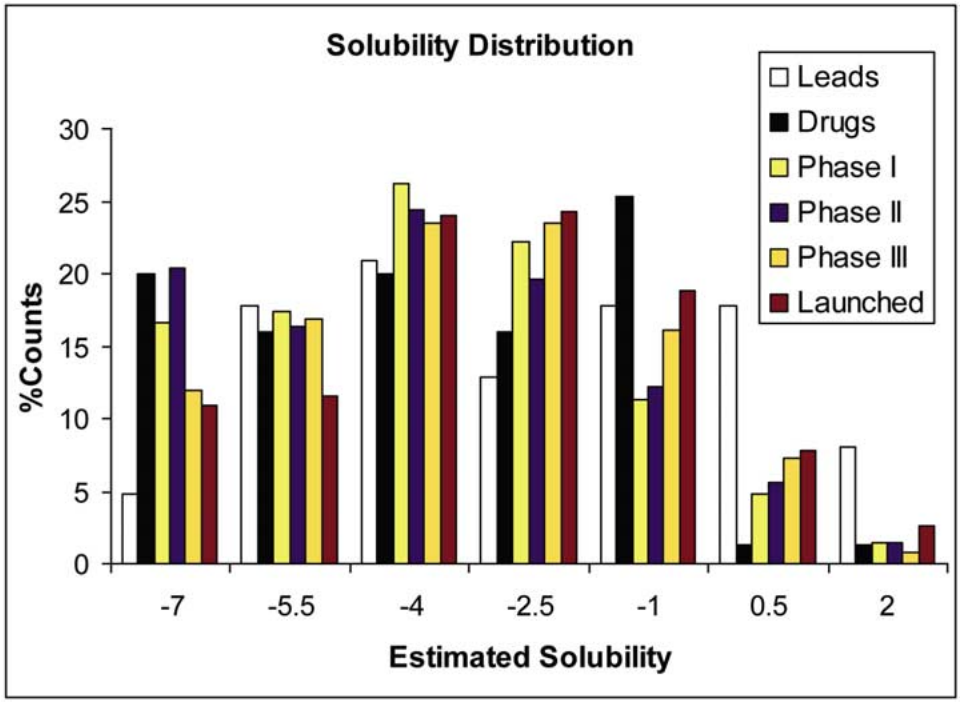


Figure 4. Solubility distribution for leads and drugs (a), compared to compounds in phase I-III clinical trials and launched drugs from MDDR (b), and with low-activity and high-activity compounds from JMC (c).

only 12.9% of the leads, 24% of the corresponding drugs, and 23.5% of the launched drugs from MDDR. However, 39.4% of phase I, 37.2% of phase II and 32.8% of phase III compounds, respectively, are in the same category. Thus, there appears to be a definite increase in MW as one steps backwards from launched drugs to phase III, II and I clinical trials, then on to high-activity compounds. This trend observed in current lead discovery is at variance with the properties observed in the lead:drug dataset.

In addition to high MW, the high-activity compounds also exhibit higher LogP values: 70% of the 'above9 only' compounds and 59.8% of the 'above 9 mul' compounds have LogP above 4.25, compared to 40.5% of the 'below 6 only' and 49.6% of the 'below 6 mul' compounds, respectively. Among MDDR compounds, 41.5% in phase I, 43.6% in phase II, 39.6% in phase III and 32.1% in launched phase have LogP above 4.25, in comparison to 35.5% of the leads and 41.3% of their corresponding drugs, respectively. It can be concluded that, whereas 35 to 45 percent of the MDDR compounds and lead structures have an estimated LogP above 4.25, a significantly higher proportion of the high-activity compounds (60–70%) fall in the same category. There appears to be a clear tendency towards synthesizing more hydrophobic structures in current medicinal chemistry reports.

As previously noted, LogSw – an intrinsic solubility in neutral state that is indicative of a compound's solubility – is estimated from MW and LogP; it can also serve as an indicator of how MW and LogP are distributed when observed simultaneously. High-activity compounds exhibit lower LogSw values: 64.5% of the 'above 9 only' compounds and 49.1% of the 'above 9 mul' compounds have LogSw below −4.75, compared to 29.4% of the 'below 6 only' and 40.4% of the 'below 6 mul' compounds, respectively. In the same category (LogSw below −4.25) rail 22.6% of the leads and 36% of the corresponding drugs. Among the MDDR compounds, 34% of phase I, 36.7% of phase II, 28.9% of phase III compounds and 22.4% of the launched drugs are also in the same category. Thus, there appears to be a definite decrease in solubility as one progresses backwards from launched drugs to phase III, II and I clinical trials, and then to high-activity compounds – trend which is not observed in the lead:drug dataset.



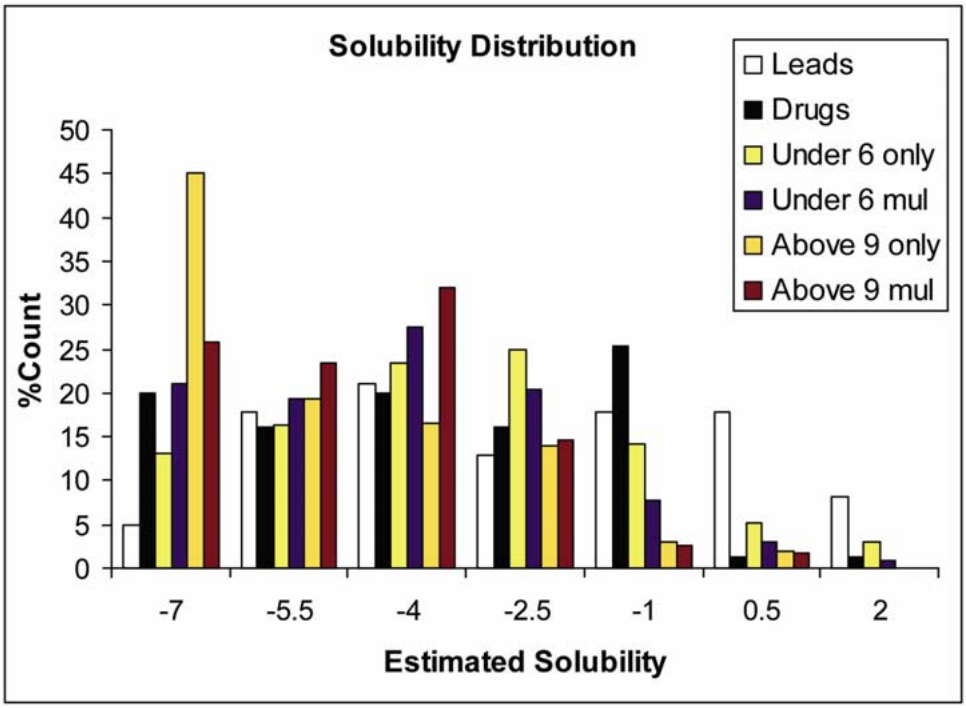


Figure 4. Continued.

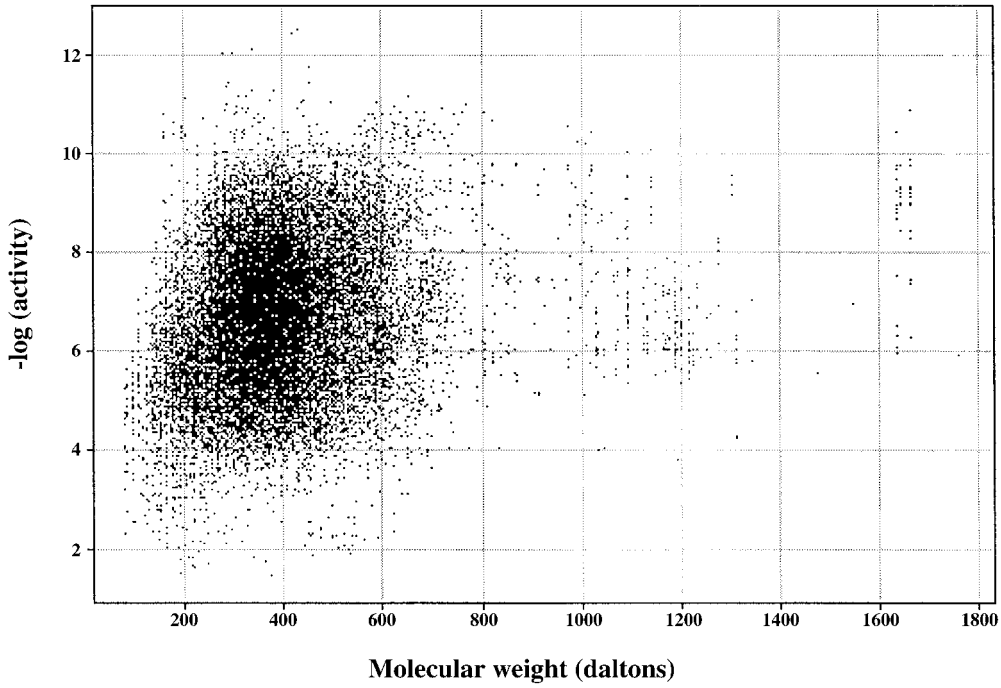


Figure 5. The relationship between molecular weight and biological activity, for 11965 structures and 22763 biological activities indexed from the literature [20].

Conclusions

Significant differences are observed when looking at the property distributions of 62 high-quality leads and their corresponding 75 drugs, as well as 985 launched drugs, compared to compounds tested in early (preclinical) drug discovery – as represented by reports published in the Journal of Medicinal Chemistry between 1997 and 1998 (mostly). More than 50% of the high-activity compounds examined here exhibit molecular weight above 425, LogP above 4.25 and LogSw below −4.75, compared to 35%, or less, of the lead compounds. A significant proportion of these high-activity compounds are receptor antagonists (75.9%) and enzyme inhibitors (23.8%), which typically constitute the major targets in pharmaceutical research. Similar trends in MW, LogP and LogSw were recently discussed for candidate drugs that emerged from Merck and Pfizer [21], indicating that poor solubility and poor permeability are among the main causes for failure during drug development. The above results show that this trend is not restricted to Merck and Pfizer. Rather, it appears to be general.

However, as pointed out by Kuntz et al. [22], higher molecular weight does not warrant higher activity – as depicted in Figure 5. Therefore, reaching nanomolar activity should not be the only goal in lead discovery. A progressive constraint can be observed when moving from preclinical research ('high-activity compounds') to clinical trials (phase I, II, III), and on to launched drugs: molecular weight and LogP appear to decrease, whereas solubility (LogSw) appears to increase in the same direction. Thus, one should seek to simultaneously optimize pharmacokinetic properties and the target property [23], when involved in lead discovery.

References

1. The top 50 pharmaceutical companies have spent, on the average, 750 millions USD for each of the 21 truly novel drugs launched during the past decade [2].
2. Drews, J., Drug Discov. Today, 3 (1998) 491–494.
3. Horrobin, D.F., J. R. Soc. Med., 93 (2000) 341–345.
4. Olsson, T. and Oprea, T.I. Curr. Op Drug Discov. Dev., 4 (2001) 308–313.
5. Lebl, M., J. Comb. Chem., 1 (1999) 3–24.
6. Martin, Y.C., J. Comb. Chem. 3 (2001) 231–250.
7. Fox, S., Farr-Jones, S. and Yund, M.A., J. Biomol. Screening, 4 (1999) 183–186.
8. Oprea, T.I., Curr. Op. Chem. Biol., 6 (2002) 384–389.
9. Rishton, G.M., Drug Discov. Today, 2 (1997) 382–384.
10. DeStevens, G., Prog. Drug. Res., 30 (1986) 189–203.

11. Oprea, T.I., Davis, A.M., Teague, S.J. and Leeson, P.D., J. Chem. Inf. Comput. Sci., 41 (2001) 1308–1315.
12. Boyd, D.B., In: Liljefors, T., Jorgensen, F.S., Krogsgaard-Larsen, P. (eds.) Rational Molecular Design in Drug Research, Munksgaard, Copenhagen, (1998) 15–29.
13. Kennedy, T., Drug Discov. Today, 2 (1997) 436–444.
14. Lipinski, C.A., Lombardo, F., Dominy, B.W. and Feeney, P.J., Adv. Drug. Deliv. Rev., 23 (1997) 3–25.
15. Teague, S.J., Davis, A.M., Leeson, P.D. and Oprea, T.I., Angew. Chem. Int. Ed., 38 (1999) 3743–3748. German version: Angew. Chem., 111 (1999) 3962-3967.
16. Hann, M.M., Leach, A.R. and Harper, G., J. Chem. Inf. Comput. Sci., 41 (2001) 856–864.
17. Meylan, W.M. and Howard, P.H., J. Pharm. Sci., 84 (1995) 83–92. KOWWIN v1.6 is available from US EPA, http://www.epa.gov/oppt/exposure/docs/episuited1.htm
18. Meylan, W.M., Howard, P.H. and Boethling, R.S., Environ. Toxicol. Chem., 15 (1996) 100–106. WSKOWIN 1.40 is available from US EPA, http://www.epa.gov/oppt/exposure/docs/episuited1.htm
19. Available from MDL Information Systems, http://www.mdli.com/dats/pharmdb.html. The MDDR database is developed in cooperation with Prous Science Publishers, http://www.prous.com/index.html. The search was conducted using the 2001.2 release.
20. The following publications were included in this survey: Journal of Medicinal Chemistry volumes 40 (1997) and 41 (1998) – forming over 90% of the structures analysed; Quantitative Structure Activity Relationships volumes 17 (1998), 18 (1999), and 19 (2000), as well as European Journal of Medicinal Chemistry volume 36 (2001).
21. Lipinski, C.A., J. Pharmacol Toxicol Meth., 44 (2000) 235–249.
22. Kuntz, I.D., Chen, K., Sharp, K.A. and Kollman, P.A., Proc. Natl. Acad. Sci. USA, 96 (1999) 9997–10002.
23. Oprea, T.I., Molecules, 7 (2002) 55-64.