JPP JOURNAL OF Pharmacy and Pharmacology
ROYAL PHARMACEUTICAL SOCIETY
JPP Journal of Pharmacy And Pharmacology
Research Paper

Size estimation of chemical space: how big is it?

Kurt L. M. Drew, Hakim Baiman, Prashanna Khwaounjoo, Bo Yu and Jóhannes Reynisson

School of Chemical Sciences, The University of Auckland, Auckland, New Zealand

Keywords
drug characterisation studies; other topics

Correspondence
Jóhannes Reynisson, School of Chemical Sciences, The University of Auckland, Private Bag 92019, Auckland 1142, New Zealand.
E-mail: j.reynisson@auckland.ac.nz

Received August 15, 2011
Accepted November 9, 2011

doi: 10.1111/j.2042-7158.2011.01424.x

Abstract

Objectives To estimate the size of organic chemical space and its sub-regions, i.e. drug-like chemical space and known drug space (KDS).

Methods Analysis of the growth of organic compounds as a function of their carbon atoms based on a power function ($f(x) = A \times B, C = x$) and an exponential function ($f(x) = AeBx$). Also, the statistical distribution of KDS and drug-like chemical space (drugs with good oral-bioavailability) based on their carbon atom count was used to deduce their size.

Key findings The power function ($f(x) = A \times B, C = x$) gives a superior fit to the growth of organic compounds leading to an estimate of 3.4×109 populating chemical space. KDS is predicted to be 2.0×106 molecules and drug-like chemical space is calculated to be 1.1×106 compounds.

Conclusions The values here are much smaller than previously reported. However, the numbers are large but not astronomical. A clear rationale on how we reach these numbers is given, which hopefully will lead to more refined predictions.

Introduction

Computer-based methods are now an integral part of drug-discovery projects and the concept of chemical space is widely used.[1–4] Different areas within chemical space are defined with molecular descriptors and chemical moieties.[5–14] The most commonly used subspace is the drug-like chemical space as defined by Lipinski.[15] Furthermore, other regions such as lead-like and fragment-based chemical space are well defined.[16–18] Recently, known drug space (KDS) was introduced, a concept based on the simple idea that marketed drugs have acceptable pharmacokinetic profiles.[19–21]

The size of the chemical space that is of interest to drug developers is estimated to lie between 1×10^{18} and 1×10^{200} compounds, the usual number given as $\sim 1 \times 10^{60}$.[2,22,23] It is therefore highly desirable to deduce a more precise estimate of the size of chemical space and its subspaces since the current numbers span hundreds of orders of magnitude. The aim of this project is to establish the size of organic chemical space based on the growth of known organic compounds as a function of their carbon atom count. This allows the statistical distribution of the drugs populating KDS and drug-like chemical space according to their carbon atom number to be used to reduce the number of possible compounds, in other words to derive the respective sizes of these subspaces.

Methodology

Organic compounds containing C, N, O, F, P, S, Cl, Br, I and/or H were compiled from the formula indexes of the Merck Index (2312 molecules)[24] and the *CRC Handbook of Chemistry and Physics* (4305 formulae).[25] Each compound formula was entered into the ChemSpider (~26 million entries)[26] and NIST Chemistry WebBook (72618 entries)[27] databases and the corresponding number of registered compounds for each molecular formula were recorded, including isomers but excluding isotopes. The compounds were categorised according to their carbon atom number and a sum generated for each category, in other words the number of organic compounds as a function of their carbon count. The larger value for every compound formula from the different databases – Chem Spider or NIST WebBook – was taken to represent the total number of compounds populating each of the carbon number groups; in other words if NIST had more entries it was used and similarly for ChemSpider. The growth of organic compounds was fitted to a power function $f(x) = \boldsymbol{A}x^{\boldsymbol{B}}$, where the number of carbons is the variable ($\boldsymbol{C} = x$), $\boldsymbol{A}$ is a scaling factor and $\boldsymbol{B}$ is the power. This function was used to predict the number of viable organic compounds for carbon atom numbers from 1 to 100. An exponential function was also used in the same way ($f(x) = \boldsymbol{A}e^{\boldsymbol{B}x}$). Furthermore, 1396 approved small molecule drugs containing fewer than 100 carbon atoms were collected from the DrugBank 3.0 website[28,29] and classified by their number of carbon atoms. The drugs with oral absorption rate higher than 50% or a description of rapid, complete or high absorption were defined as populating drug-like chemical space (517 drugs). The drugs used are given in Tables S1 and S2 in the Supporting information.

Results

Size of organic chemical space

In this study the count of the number of organic molecules is based on compounds that are characterised and have an entry in either of the chemical databases used: Chemspider and/or NIST Chemistry WebBook. This means that the growth curve is built from known organic molecules but not hypothetical ones. When the results are plotted as a function of the molecules' carbon numbers it is apparent that after six carbons the rate of growth of entries is reduced and that the number of entries declines after ten carbons, as shown in Figure 1. It can be stated that chemical space containing organic compounds with relatively few carbon atoms is thoroughly investigated whereas more complicated molecules are less so for the simple reason that there are so many more of them and they are often more challenging to synthesise.

The fitting of the mathematical functions was therefore only based on the first four to seven carbon atoms, as these values are expected to be fairly representative of the true sum of compounds for each carbon group. This is obviously an approximation and can potentially affect the results. The results are shown in Table 1.

It was found that a power function ($f(x) = \boldsymbol{A}x^{\boldsymbol{B}}$) fitted the trend line of the data best as shown in Figure 2 and is reflected in good R^2 values (see Table 1). An exponential function ($f(x) = \boldsymbol{A}e^{\boldsymbol{B}x}$) did *not* fit the data as accurately as the power function, i.e. the R^2 value of the exponential was 0.934 compared to 0.997 for the power function at six carbons. Other mathematical functions fitted the increase of organic compounds less accurately. The results using seven carbons were rejected because for both the power and exponential functions the R^2 values are relatively poor. The power function scaling factor ***A*** is very similar for four, five and six carbons, sitting around 170; the power factor ***B*** is also very similar with a value slightly less than 3. For the exponential fit there is a much greater variability than for the power function. As the number of carbons increases, so does the pre-exponential factor **A**. The exponential factor **B** decreases. A further comparison between the power and exponential functions was

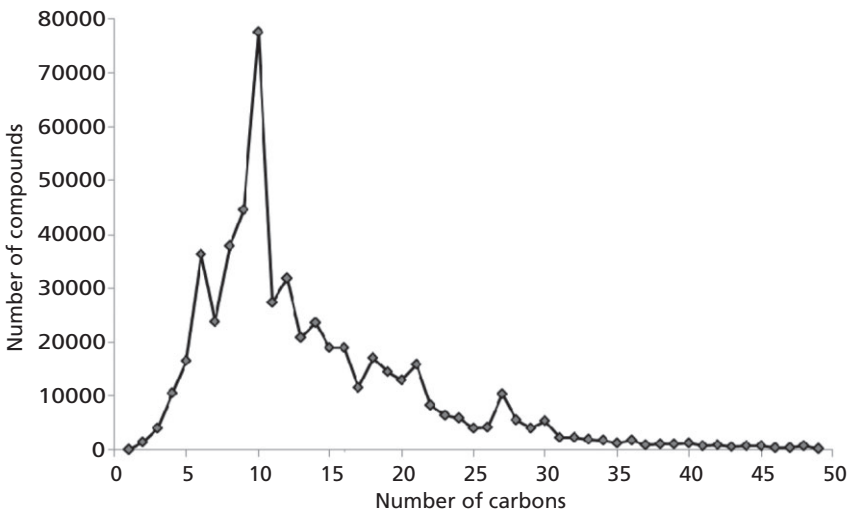

Figure 1 The distribution of organic compounds as a function of their carbon atom content derived from the Chemnetbase database.[30] The number of molecules grows rapidly up to six carbons. A similar distribution shown is seen for ChemSpider and NIST databases.

Table 1 The result of the fitting of power and exponential functions to the growth of known organic molecules as a function of their carbon atom count. The size of chemical space is estimated for ≤100 carbon atoms per molecule

	Power function $f(x) = Ax^B$				Exponential function $f(x) = Ae^{Bx}$			
Carbons	A	B	R^2	Chemical space size	A	B	R^2	Chemical space size
4	167.03	2.97	0.998	3.7×10^9	60.13	1.35	0.955	4.5×10^{60}
5	172.96	2.89	0.997	2.8×10^9	95.18	1.12	0.929	8.3×10^{50}
6	168.06	2.95	0.997	3.4×10^9	124.05	1.01	0.934	1.4×10^{46}
7	192.20	2.74	0.979	1.6×10^9	208.11	0.82	0.860	9.7×10^{37}

performed with two separate Kolmogorov–Smirnov (K–S) tests.[31] This method was employed because it is non-parametric and is well suited for testing the quality of fit of non-linear data. Although the *P* values for both were 1, the exponential function D values were much larger than the power function D values, which reflects the superiority of the power function. Furthermore, the use of the power function also makes intuitive sense because for each carbon group there is only a fixed number of possible compounds depending on all the possible combinations of C, N, O, F, P, S, Cl, Br, I and/or H that can be bonded to each carbon. This is the same principle used to model the number of possible combinations of throws of a die, where the total number of possible combinations is calculated as the number of faces on each die raised to the power of how many dice there are. In this case it is the number of carbons raised to a fixed power.

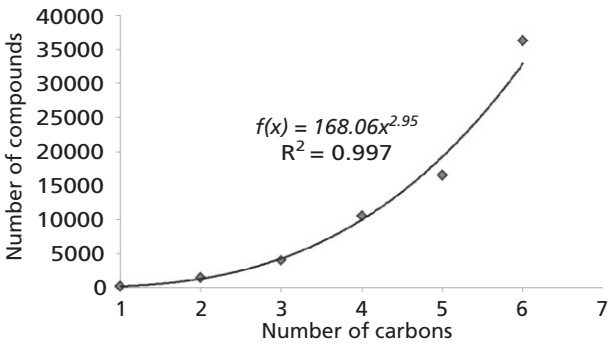

Figure 2 The number of organic molecules as a function of their carbon atoms fitted to a power function. The R^2 value of 0.997 indicates a good fit between the power trend line and growth rate of organic molecules.

In order to deduce an estimate of the size of chemical space, the power function derived from the fitting curve shown in Figure 2 was used ($f(x) = 168.06x^{2.95}$). The estimated value for each carbon atom category was calculated and these were summed up to 100 carbons. The cumulative size estimates are given in Tables 2 and 3. The predicted size of chemical space ≤100 carbon atoms is 3.4×10^9 compounds, which is much lower than the frequently cited value of ~1×10^{60}.[2] The published size estimates of chemical space vary greatly depending on the criteria used and it is difficult to make direct comparisons. For example, the size of synthetically accessible chemical space has been estimated to lie between 1×10^{20} and 1×10^{24} molecules.[32] Geysen *et al.*[33] mention that the size of chemical space may lie between 1×10^{14} and 1×10^{30}. Ogata *et al.*[34] estimated the size of chemical space to be only 1×10^8 to 1×10^{19} molecules. Fink *et al.*[35,36] estimated 26.4 million compounds from 11 atoms with C, N, O and F, and 9.8×10^8 molecules containing 13 atoms. Finally, neurological drug space is calculated to be approximately 1×10^{15} molecules.[37] In general, the size we calculated of 3.4×10^9 is much lower than most of the estimates previously published, with the exception of the Ogata prediction.

Size estimation of drug-like chemical space and known drug space

A total of 1396 small drugs from Drug Bank[28,29] were arranged into a histogram according to their carbon atom counts, as shown in Figure 3. The most common carbon count in drugs is 17, with 82 entries. For the orally

Table 2 Estimated size of chemical space, known drug space (KDS), based on the number of carbon atoms

Carbons	Chemical space size	C17 = 1 total drugs	Ratio between known drugs and known organic compounds	Ratio between known drugs and all drugs
≤20	6.4×10^6	5.3×10^6	2.7×10^4	3.1×10^5
≤40	9.4×10^7	1.8×10^7	3.6×10^5	1.0×10^6
≤60	4.6×10^8	2.8×10^7	3.2×10^6	1.6×10^6
≤80	1.4×10^9	3.2×10^7	1.1×10^7	1.9×10^6
≤100	3.4×10^9	3.4×10^7	2.7×10^7	2.0×10^6

Table 3 Estimated size of chemical space, drug-like chemical space based on the number of carbon atoms

Carbons	Chemical space size	C17 = 1 total drugs	Ratio between known orally bioavailable drugs and known organic compounds	Ratio between drugs and all orally bioavailable drugs
≤20	6.4×10^6	4.9×10^6	1.2×10^4	3.7×10^5
≤40	9.4×10^7	1.1×10^7	1.3×10^5	8.6×10^5
≤60	4.6×10^8	1.3×10^7	8.6×10^5	9.8×10^5
≤80	1.4×10^9	1.5×10^7	2.9×10^6	1.1×10^6
≤100	3.4×10^9	1.5×10^7	6.6×10^6	1.1×10^6

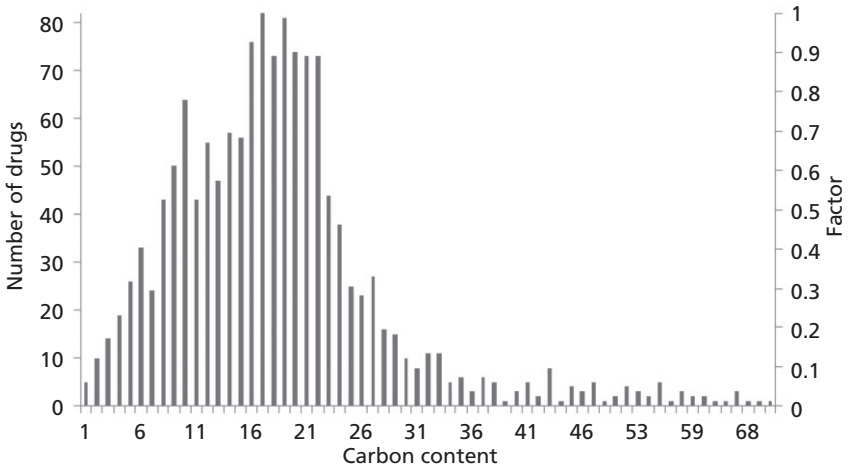

Figure 3 The distribution of 1396 known drugs according to their carbon count. 82 drugs have 17 carbons, which represents the maximum.

bioavailable drugs (drug-like chemical space) the maximum was also found at C17, with 39 drugs. The statistical distributions for KDS and drug-like chemical space are very similar, except the former has a longer tail into the higher carbon values.

Three approaches were used to estimate the size of drug-like chemical space and KDS.

First, it is clear that drugs with 17 carbons are the most prevalent. Here it is assumed that this is the optimal number of carbon atoms in drugs and all molecules in this category are drug candidates. The other carbon categories are calculated as a fraction of carbon 17: the number at C17 was set as 1 and the rest of the carbon groups were calculated as a fraction of this figure. For example, C17 has 82 entries and C9 has 50, giving it a fraction of 0.61. These fractions were then multiplied by their corresponding chemical space values derived from the power function ($f(x) = 168.06x^{2.95}$). The values for the carbon categories were summed and the results are given in Table 2. The number of possible drugs or the size of KDS $\leq$100 carbons is estimated at 3.4×10^7 molecules. The same procedure was performed for the 517 orally bioavailable drugs, giving an estimate for drug-like chemical space of 1.5×10^7 molecules (see Table 3).

In the second method employed, the total number of organic compounds for each carbon count up to 100 was obtained from the online Chemnetbase database (see Figure 1).[30] These values were compared with the results previously collected from Chemspider up to ten carbon counts.[26] Chemspider has ~26 million entries and the current estimate of characterised organic compounds is ~30 million.[38] This means that Chemspider contains the majority of known organic compounds (87%). It was found that on average the Chemspider values were three times greater than that of the Chemnetbase. Therefore, in order to estimate the values of known compounds in each carbon category, the results from Chemnetbase were multiplied by three. The number of known drugs for each carbon group was divided by the value for known compounds. For example, C17 has 82 drug entries and an estimated 11,561 known organic compounds, giving a fraction of 0.007. In some cases this fraction was zero because there are no known drugs with these carbon counts, which is unlikely to be a fair representation. To compensate for this, the average of these fractions was calculated and used to replace every zero in the 1 to 100 range. The fractions thus generated were multiplied by their corresponding values from the power function ($f(x) = 168.06x^{2.95}$) and summed over the carbon categories. This resulted in a value of 2.7×10^7 for KDS and 6.6×10^6 molecules for drug-like chemical space, as shown in Tables 2 and 3.

In the third approach a fraction was calculated from the overall total of known drugs. For example, C17 has 82 entries and there are a total of 1396 small drugs, giving a fraction of 0.0587. These fractions were multiplied by their corresponding chemical space values generated from the power function shown in Figure 2. These values were summed and the results are shown in Tables 2 and 3. According to this method KDS is estimated to have 2.0×10^6 molecules at $\leq$100 carbons and 1.1×10^6 molecules for the orally bioavailable drugs.

When these three methods are considered then the first scenario surely overestimates the size of the subspaces calculated, since it is impossible that all organic compounds containing 17 carbon atoms can be drug molecules. For the second method the number of known organic compounds has to be estimated based on the ratios of compounds found in ChemSpider (it is not possible to retrieve this data directly from Chemspider) and ChemnetBase, and it is difficult to assess the quality of this estimate. The third method uses the ratio between known drugs in each carbon category and the

total number of marketed drugs. This is a simple approach and in the tradition of Ockham's razor should give the most reasonable results.

The simplest way to calculate how many drugs are possible is to set the number of possible organic compounds at 3.4×10^9 and to note that it is established that ~30 million compounds have been characterised so far.[38] From these 30 million compounds 1396 drugs have emerged. Assuming a linear correlation exists between the number of known organic molecules and the number of marketed drugs, the value of 1.6×10^5 possible drugs is reached.

In all three scenarios it is found that KDS is larger than drug-like chemical space. This is in line with the finding that the molecular descriptors used to define chemical space are considerably larger for KDS than for drug-like chemical space. In other words, for KDS, MW $\leq$ 800, log P $\leq$ 6.5, HBD $\leq$ 7, HBA $\leq$15, polar surface area $\leq$ 180 Å^2 and rotatable bond $\leq$ 17, while for drug-like chemical space, MW $\leq$ 500, log P $\leq$ 5, HBD $\leq$ 5, HBA $\leq$ 10, polar surface area $\leq$ 140 Å^2 and rotatable bond $\leq$ 10.[20] In this study, the only molecular descriptor used is the carbon count of the benchmarking drugs, making the fractionation of the power function mathematically feasible. It can be stated that the carbon content of compounds correlates to some degree with increasing MW, log P, number of hydrogen bond donors and acceptors, polar surface area and number of rotatable bonds, and therefore captures their ability to define areas in chemical space. In order to improve the estimates derived in this work, the classical molecular descriptors and undesirable molecular moieties could be used. The challenge is to define chemical space unbiased towards molecules with drug-like properties. This excludes both commercially available compound collections for high throughput screening and more generic collections, due to the extensive impact medicinal chemistry has had on the development of synthetic organic chemistry.

Conclusion

Gaining a robust understating of the nature of chemical space is critical in order to facilitate the discovery and development of drugs. The size of this phenomenon is estimated to be 3.4×10^9 molecules, which is a much lower value than previously reported. The pharmacokinetically benign areas of drug-like chemical space and KDS are deduced to be 2.0×10^6 and 1.1×10^6 molecules, respectively. This means that 7×10^{-2}% of known drugs (1396 drugs) have been discovered and 5×10^{-2}% of all orally bioavailable drugs (517 drugs). The results presented here indicate that we have hardly started to tap into the potential that small molecules represent. This gives us hope that the needs of therapeutic areas such as cancer, cardiovascular, Alzheimer's disease, Parkinson's disease, diabetes and other conditions will be met by the development of small molecule drugs.

Declaration

Conflict of interest

The Author(s) declare(s) that they have no conflicts of interest to disclose.

References

1. Muchmore SW *et al*. Cheminformatic tools for medicinal chemists. *J Med Chem* 2010; 53: 4830–4841.
2. Reymond J-L *et al*. Chemical space as a source for new drugs. *Med Chem Comm* 2010; 1: 30–38.
3. Leeson PD, Springthorpe B. The influence of drug-like concepts on decision making in medicinal chemistry. *Nat Rev Drug Dis* 2007; 6: 881–890.
4. Medina-Franco JL *et al*. Visualization of the chemical space in drug discovery. *Curr Comp Aided Drug Des* 2008; 4: 322–333.
5. Lipinski CA *et al*. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development setting. *Adv Drug Deliv Rev* 1997; 23: 3–25.
6. Lipinski CA. Lead- and drug-like compounds: the rule-of-five revolution. *Drug Discov Today Technol* 2004; 1: 337–341.
7. Palm K *et al*. Polar molecular surface properties predict the intestinal absorption of drugs in humans. *Pharm Res* 1997; 14: 568–571.
8. Veber DF *et al*. Molecular properties that influence the oral bioavailability of drug candidates. *J Med Chem* 2002; 45: 2615–2623.
9. Lu JJ *et al*. Influence of molecular flexibility and polar surface area metrics on oral bioavailability in the rat. *J Med Chem* 2004; 47: 6104–6107.
10. Rishton GM. Reactive compounds and in vitro false positives in HTS. *Drug Discov Today* 1997; 9: 382–384.
11. Rishton GM. Nonleadlikeness and leadlikeness in biochemical screening. *Drug Discov Today* 2003; 8: 86–96.
12. Baell JB, Holloway GA. New substructure filters for removal of pan assay interference compounds (PAINS) from screening libraries and for their exclusion in bioassays. *J Med Chem* 2010; 53: 2719–2740.
13. Mcgovern SL *et al*. A specific mechanism of nonspecific inhibitors. *J Med Chem* 2003; 46: 4265–4272.
14. Jadhav A *et al*. Quantitative analyses of aggregation, autofluorescence, and reactivity artifacts in a screen for

inhibitors of a thiol protease. *J Med Chem* 2010; 53: 37–51.
15. Lipinski CA. Properties and the causes of poor solubility and poor permeability. *J Pharmacol Toxicol Methods* 2000; 44: 235–249.
16. Oprea TI. Current trends in lead discovery: are we looking for the appropriate properties? *Mol Divers* 2000; 5: 199–208.
17. Oprea TI *et al.* Is there a difference between leads and drugs? A historical perspecitve. *J Chem Inf Comput Sci* 2001; 41: 1308–1315.
18. Chen Y, Shoichet BK. Molecular docking and ligand specificity in fragment-based inhibitor discovery. *Nature Chem Biol* 2009; 5: 358–364.
19. Axerio-Cilies P *et al.* Investigation of the incidence of 'undesirable' molecular moieties for high-throughput screening compound libraries in marketed drug compounds. *Eur J Med Chem* 2009; 44: 1128–1134.
20. Bade R *et al.* Characteristics of known drug space. Natural products, their derivatives and synthetic drugs. *Eur J Med Chem* 2010; 45: 5646–5652.
21. Mirza A *et al.* Known drug space as a metric in determining the boundaries of drug-like chemical space. *Eur J Med Chem* 2009; 44: 5006–5011.
22. Bohacek RS *et al.* The art and practice of structure-based drug design: a molecular modeling perspective. *Med Res Rev* 1996; 16: 3–50.
23. Fink T *et al.* Virtual exploration of the small-molecule chemical universe below 160 daltons. *Angew Chem Int Ed Engl* 2005; 44: 1504–1508.
24. Windholz M *et al. The Merck Index*, 9th edn. Rahway: Merck & Co. Inc., 1976.
25. Lide DR. Formula index for organic compounds. In: Lide DR, ed. *Handbook of Chemistry and Physics*. London: CRC Press, 1993: 3-601–633.
26. ChemSpider database. http://www.chemspider.com/.
27. NIST Webbook. http://webbook.nist.gov/chemistry/.
28. Wishart DS *et al.* DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucl Acids Res* 2008; 36: D901–D906.
29. Wishart DS *et al.* DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucl Acids Res* 2006; 34: D668–D672.
30. Chemnetbase database http://www.chemnet.base.com/. *Dictionary of Drugs*. Taylor & Francis Group, 2011.
31. Clauset A *et al.* Power-law distributions in empirical data. *SIAM Rev* 2009; 4: 661–703.
32. Ertl P. Cheminformatics analysis of organic substituents: identification of the most common substituent, calculations of substituents properties, and automatic identification of drug-like bioisoteric groups. *J Chem Inf Comput Sci* 2003; 43: 374–380.
33. Geysen HM *et al.* A guide to drug discovery: combinatorial compound libraries for drug discovery: an ongoing challenge. *Nat Rev Drug Dis* 2003; 2: 222–230.
34. Ogata K *et al.* A quantitative approach to the estimation of chemical space from a given geometry by the combination of atomic species. *QSAR Comb Sci* 2007; 26: 596–607.
35. Fink T, Reymond J-L. Virtual exploration of the chemical universe up to 11 atoms of C, N, O, F: assembly of 26.4 million structures (110.9 million stereoisomers) and analysis for new ring systems, stereochemistry, physicochemical properties, compound classes, and drug discovery. *J Chem Inf Model* 2007; 47: 342–353.
36. Blum LC, Reymond J-L. 970 million drug-like small molecules for virtual screening in the chemical universe database GDB-13. *J Am Chem Soc* 2009; 131: 8732–8733.
37. Weaver DF, Weaver CA. Exploring neurotherapeutic space: how many neurological drugs exist (or could exist)? *J Pharm Pharmacol* 2011; 63: 136–139.
38. Nicolaou KC, Montagnon T. *Molecules That Changed the World*. Weinheim: Wiley-VCH Verlag GmbH, 2008.

Supporting information

Additional supporting information may be found in the online version of this article:

Table S1 All drugs analysed.
Table S2 Drugs with good bioavailability.

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.

