Bioorganic & Medicinal Chemistry Letters 19 (2009) 6943–6947

Contents lists available at ScienceDirect

Bioorganic & Medicinal Chemistry Letters

journal homepage: www.elsevier.com/locate/bmcl

Physicochemical property profiles of marketed drugs, clinical candidates and bioactive compounds

Christian Tyrchan, Niklas Blomberg, Ola Engkvist, Thierry Kogej, Sorel Muresan *

DECS Global Compound Sciences, AstraZeneca R&D Mölndal, SE-431 83 Mölndal, Sweden

ARTICLE INFO

Article history:
Received 16 August 2009
Revised 13 October 2009
Accepted 14 October 2009
Available online 20 October 2009

Keywords:
Physicochemical properties
Marketed drugs
Clinical candidates
Bioactive compounds
SAR databases

ABSTRACT

We performed a comparison of several simple physicochemical properties between marketed drugs, clinical candidates and bioactive compounds using commercially available databases (GVKBIO, Hyderabad, India). In contrast to previous studies this comparison was performed at the individual target level. Confirming earlier studies this shows that marketed drugs have, on average and taken as a single set, lower physicochemical property values than the corresponding clinical candidates and bioactive compounds but that there is considerable variation between drug targets. This work complements earlier studies by using a much larger annotated dataset and confirms that there is a shift in physicochemical properties for targets with launched drugs and clinical candidates compared to bioactive compounds.

Inspired by the introduction of the rule-of-five[1], pharmaceutical research scientists have examined a range of simple molecular properties for medicinally relevant compounds.[2–5] The aim has been to identify trends between compounds in different stages of the drug discovery process. These trends should help to define pragmatic boundaries that would increase the quality of the starting points in lead optimisation and consequently would reduce attrition in clinical trials.[6]

Proprietary or published bioactive compounds, hits and leads, clinical candidates and marketed drugs are usually included in such analyses. Structural and associated biological information of various collections of such compounds is available from several commercial and public sources.[7]

The GVKBIO Medicinal Chemistry and Target Class databases capture explicit relationships between published documents, compounds, assay results and Entrez Gene IDs (EGIDs).[8] GVKBIO uses expert curators to populate databases with these relationships extracted from medicinal chemistry journals and patents. At AstraZeneca all this information is merged into one application, IBEX,[9] to enable complex queries on more than 3 million records representing over 2 million unique structures and 10 million structure-activity relationship (SAR) data points. A selection of this large collection represents the bioactive compound set in this work. In addition, the Drug database (3211 records) and the Clinical Candidate database (10,715 records), also from GVKBIO, have been used in this analysis.[8]

This work complements earlier studies by reporting comparisons of physicochemical properties at the individual target level for large collections of medicinally relevant compounds.

All data were extracted from GVKBIO Medicinal Chemistry and Target Class databases, GVKBIO Clinical Candidate and GVKBIO Drug databases. We selected to focus our analysis on orally available compounds with human targets. Furthermore, we restricted the disease areas within the analysis to exclude compounds such as antiseptics, antivirals, disinfectants, vitamins. Compounds with no associated EGID, no defined molecular structure or belonging to one of 29 excluded activity classes were removed from the analysis. The physicochemical properties for the remaining compounds were calculated and large molecules (molecular weight (MW) above 1000 or polar surface area (PSA) above 300) were removed as extreme outliers. Finally, compounds only associated with a set of 31 targets such as histone, cytochrome P450s, multi-drug resistance protein and follicle-stimulating receptor were removed (see Supplementary data for a complete list). The remaining set comprised 2088 drug SAR data points (corresponding to 976 unique compounds and 221 unique EGIDs) and 6607 clinical candidate SAR data points (corresponding to 3957 unique compounds and 473 unique EGIDs). We extracted 898 human targets that had more than 25 bioactive compounds assigned from Medicinal Chemistry and Target Class databases. The final dataset contained 3,006,464 SAR data points, corresponding to 1,184,611 unique compounds, from which 4927 are unique drugs/clinical candidate compounds, and 898 unique EGIDs. Subsequently we mapped the 4927 structural unique drugs and clinical candidates to 504 IBEX targets.

* Corresponding author.
E-mail address: sorel.muresan@astrazeneca.com (S. Muresan).

0960-894X/$ - see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.bmcl.2009.10.068

To exemplify our findings we selected all compounds extracted from patents in IBEX associated to CCR5 (EGID 1234, 266 patents with 14,829 unique compounds), SCL6A3 (EGID 6531, 7336 unique compounds), CNR1 (EGID 1268, 20,766 unique compounds), REN (EGID 5942, 8856 unique compounds) and DPP4 (EGID 1803, 4540 unique compounds).

The physicochemical properties used in this study are cLog *P*,[10] MW, number of heavy atoms (HA, non-hydrogen atoms), PSA,[11] number of rotatable bonds (RotBond), number of H-bond acceptors (HBA, number of O+N) and number of H-bond donors (HBD, number of OH+NH).

Rotatable bonds are calculated using the following equation:

$$RotBond = Nrot, ac + \sum_{rings}(Size - 4 - Nrig, cyc - Nfused)$$

The first term (*Nrot,ac*) is the number of flexible acyclic bonds, whilst the second takes into account ring flexibility. *Size* is the size of the ring, *Nrig,cyc* is the number of rigid bonds within the ring and *Nfused* is the number of bonds which belong to more than one ring system. Rings are defined according to the *SSSR* algorithm.[12]

The data preparation and analysis were done with PipelinePilot,[13] and the statistical analyses was performed with JMP.[14]

For each target we analyzed the difference in physicochemical properties between drugs, clinical candidates and bioactive compounds. For the 504 targets with at least one drug or clinical candidate assigned we calculated the property profile of the three datasets (drugs, clinical candidates and bioactive) by binning and subsequently we determined the cumulative percentage compounds for each bin. Figure 1 shows the result of this analysis for cLog *P*.

We observe lower physiochemical property values for drugs than the corresponding, at the target level, clinical candidates or bioactive compounds (resulting in a left-shift of the drug dataset in the cumulative plot). For instance, more than 74% of the drugs have a cLog *P* lower than 4.0 whereas this is the case for only 62% of the clinical candidates and 50% of the bioactive compounds. Furthermore, 22% of the clinical candidates and 30% of the bioactive compounds have a cLog *P* of higher than 5.0, this is seen for only 11% of the drugs. Only 82% of the clinical candidates and 72% of the bioactive compounds have MW lower than 500 compared to 95% of the drugs (figure not shown).

Table 1 shows the mean, standard deviation and median of the physicochemical properties for the drug, clinical candidate and bioactive datasets. The ANOVA analysis, Student's *t*-test and the Wilcoxon test show that all the differences observed are statistical significant (using a *p*-value significance level of 0.05).[14]

Table 1
The mean, standard deviation (stdev) and median of the physicochemical properties for the drug, clinical candidate and bioactive datasets

Property	Mean	Stdev	Median
Drugs (*n* = 976)			
cLog *P*	2.74	2.22	2.83
MW	335.5	109.2	318.5
PSA	64.7	39.7	59.1
RotBond	5.6	3.6	5
HBD	1.5	1.5	1
HBA	3.9	2	4
Clinical candidates (*n* = 6607)			
cLog *P*	3.39	2.31	3.47
MW	415.1	126.9	402.5
PSA	86.7	49.1	77.7
RotBond	7.5	4.5	7
HBD	1.9	1.7	2
HBA	5.3	2.3	5
Bioactive compounds (*n* = 1,184,611)			
cLog *P*	4.04	2.17	4.09
MW	455.0	111.4	450.5
PSA	87.5	44.4	81.5
RotBond	8.2	3.9	8
HBD	1.9	1.7	2
HBA	5.4	2.1	5

Similar results, without taking target information into account and for a smaller dataset, were reported by Wenlock et al.[2] comparing sets of marketed oral drugs and oral drugs in different phases of clinical development. To extend this common compound centric analysis, we performed a target-centred analysis of compound properties.

For each target, we calculated the average cLog *P* and MW of the drugs, clinical candidates and bioactive molecules reported. The resulting distribution of per-target property averages was binned and the cumulative percentage of targets was subsequently calculated for each bin, as shown in Figures 2 and 3.

For most drug targets, the cLog *P* for the corresponding drugs is considerably lower than the average cLog *P* for associated clinical candidates and bioactive molecules. In Figures 2 and 3 this is seen as a distinct left-shift in the cumulative distribution for drugs compared to those of clinical candidates and bioactive compounds. For

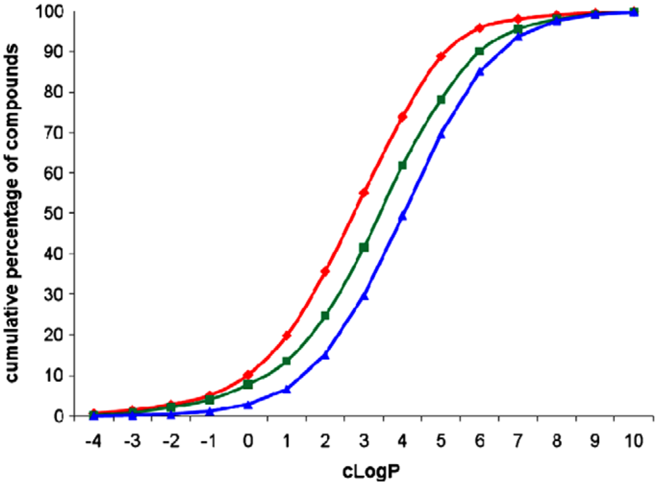

Figure 1. The cumulative percentage of compounds (for all targets containing at least one drug or one clinical candidate) versus cLog *P* (red line: drugs, green line: clinical candidates, blue line: bioactive compounds).

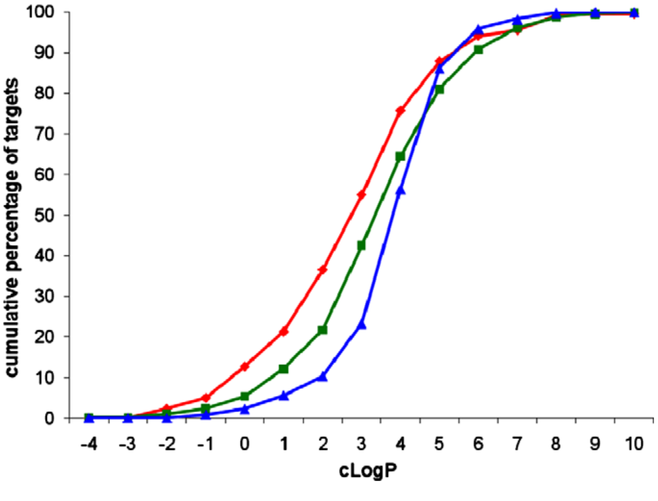

Figure 2. The cumulative percentage of targets (containing at least one drug or one clinical candidate) calculated from the binned averages of cLog *P* for each target versus cLog *P* (red line: drugs, green line: clinical candidates, blue line: bioactive compounds).

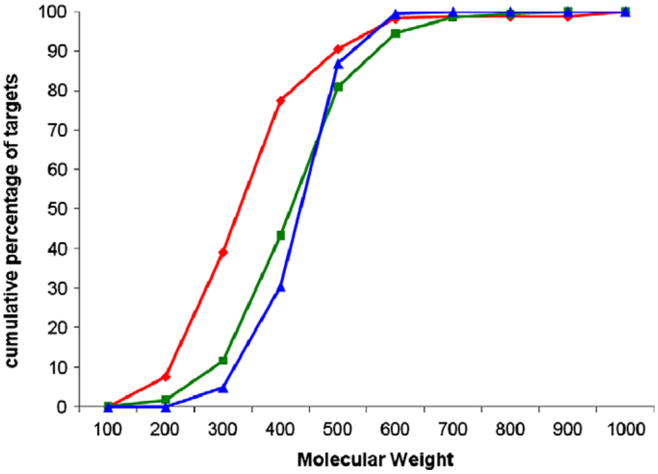

Figure 3. The cumulative percentage of targets (containing at least one drug or one clinical candidate) calculated from the binned averages of MW for each target versus MW (red line: drugs, green line: clinical candidates, blue line: bioactive compounds).

76% of the drug targets the average cLog *P* for the corresponding drugs is below 4.0 whereas only 65% of the same targets have clinical candidates with a cLog *P* of less than 4.0. Looking at assigned bioactive compounds from the patent corpus and medicinal chemistry literature only 58% of these targets have an average cLog *P* of less than 4.0.

The same trend is seen for MW (Fig. 3): 77% of the targets have an average MW of the corresponding drugs of less than 400. In contrast only 37% of the targets have an average MW of their reported bioactive compounds below 400.

A few targets do not follow the observed general trend of lower MW and lower cLog *P* for the drugs compared to the clinical candidates and the bioactive compounds. In the following we will select two targets to illustrate both this general trend as well as the exceptions.

The Dopamine Transporter (SLC6A3, EGID 6531) is an example of a target that follows the general trend. We could assign 10 drugs, 19 clinical candidates and 7307 bioactive compounds to SLC6A3. All drugs have cLog *P* values less than 4.5, whereas only 12 (63%) of the clinical candidates and 4690 (64%) of the bioactive compounds have a cLog *P* lower than 4.5 (Fig. 4).

An exception from the general trend is the Cannabinoid receptor 1 (CNR1, EGID 1268). The endogenous ligand 9-tetrahydrocannabinol of CNR1 has a high cLog *P* of 7.2. The compound is marketed as Dronabinol®. Levonantrapol® (cLog *P* 5.0) and Nabilone® (cLog *P* 6.7) are close analogues to the endogenous ligand. The other two marketed drugs Rimonabant® (a CNR1 antagonist recently withdrawn from the market) and Pravadoline® (CNR1 agonist) have cLog *P* values of 6.5 and 4.4, respectively (see Fig. 5).

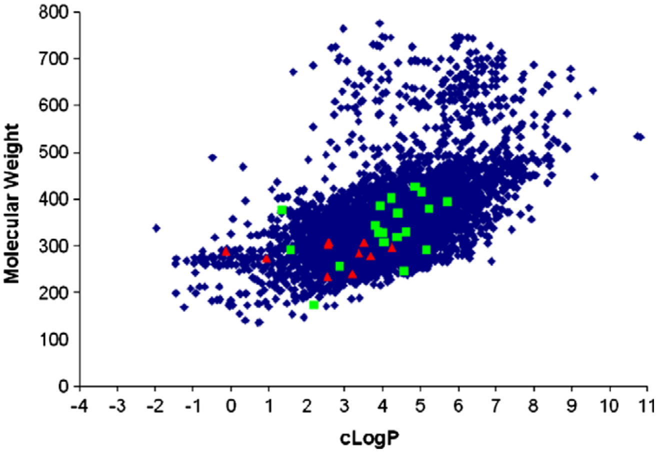

Figure 4. MW versus cLog *P* for drugs (red), clinical candidates (green) and bioactive compounds (blue) for Dopamine Transporter SLC6A3.

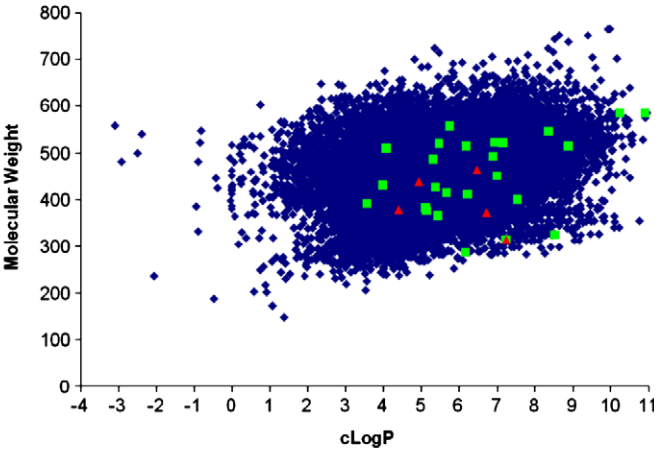

Figure 5. MW versus cLog *P* for drugs (red), clinical candidates (green) and bioactive compounds (blue) for Cannabinoid receptor CNR1.

Following the recent analysis by Leeson and Springthorpe,[15] we further investigated the relation between physicochemical properties of compounds from different pharmaceutical companies.

First we repeated the analysis for CCR5 (EGID 1234). Again, a lower cLog *P* for the marketed drug (Maraviroc®, Pfizer, cLog *P* 3.3) compared to the bioactive compounds even those published by Pfizer (see Fig. 6) was observed.

When comparing the cLog *P* for the bioactive CCR5 compounds disclosed by seven different companies, a shift towards lower cLog *P* values can be observed for compounds from Pfizer and AstraZeneca. More than 90% of the compounds from Pfizer have a cLog *P* lower than 4.0. Interestingly, these lower cLog *P* values are not accompanied by a significant shift in the number of heavy atoms (see Fig. 7).

For other targets like REN (EGID 5942) and DPP4 (EGID 1803) we could not see the same large differences in cLog *P* for bioactive compounds (see Figs. 8 and 9). Nevertheless, we still observe the

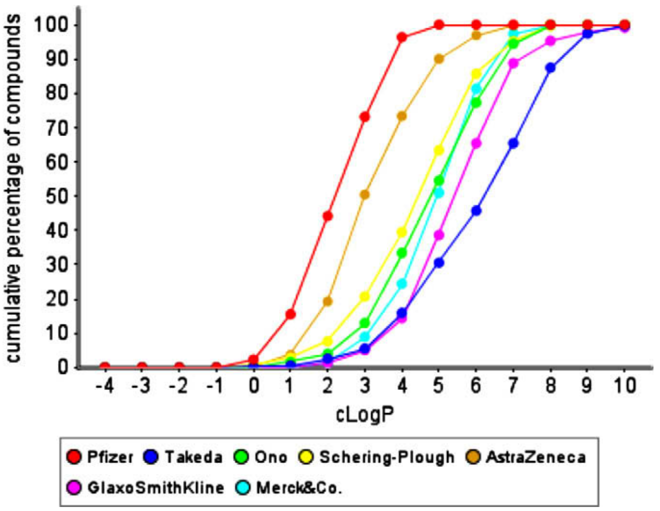

Figure 6. The cumulative percentage of CCR5 bioactive compounds versus cLog *P* for several companies.

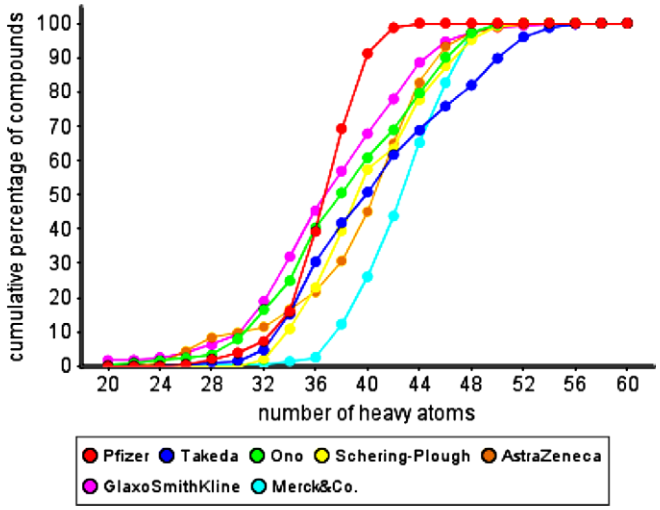

Figure 7. The cumulative percentage of CCR5 bioactive compounds versus number of heavy atoms for several companies.

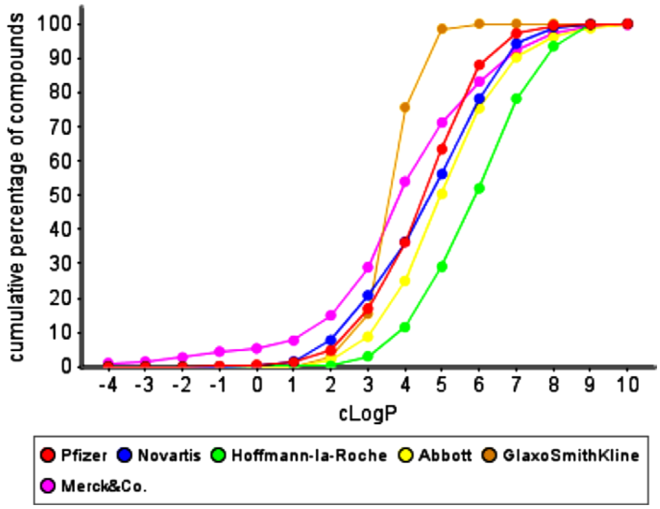

Figure 9. The cumulative percentage of REN bioactive compounds versus cLog *P* for several companies.

trend to lower cLog *P* and MW for marketed drugs, compared to bioactive compounds.

For DPP4 (EGID 1803), both marketed drugs Vildagliptin® (cLog *P* 0.7, Novartis) and Sitagliptin® (cLog *P* 0.7, Merck&Co.) have a significantly lower cLog *P* than the majority of the bioactive compounds from both companies. When we consider all bioactive compounds for DPP4 more than 78% have a cLog *P* value greater than 1.0 and 45% have a cLog *P* value greater than 2.0 (see Fig. 8).

For REN, the marketed drug Aliskiren® from Novartis has a cLog *P* of 3.5. In IBEX there are 1752 compounds associated with both Novartis and REN and more than 60% (1114 compounds) of these bioactive compounds have a cLog *P* higher than 4.0 (see Fig. 9). This also holds true when looking at all 8856 bioactive compounds in IBEX associated with REN, of which 5425 compounds (61%) have a cLog *P* higher than 4.0.

In contrast to DPP4 and CCR5, we observe for REN a significant shift comparing the number of heavy atoms for bioactive compounds between different companies (figure not shown).

Based on over 3 million SAR data points and over 1 million unique structures our analysis at the target level supports the general findings in the literature that marketed drugs have, on average, lower physiochemical property values compared to clinical candidates and bioactive compounds. Two specific examples have been discussed, one (SLC6A3) that is in agreement with the general trend and one (CNR1) where there is no difference between the marketed drugs, clinical candidates and bioactive compounds. We also note the marked difference in physicochemical properties of the natural ligands of these targets (dopamine vs 9-tetrahydrocannabinol) and, while it is tempting to speculate that part of the differences seen in properties for drugs versus clinical candidates is due to target bias, the target-centric analysis clearly shows a shift in distributions within the same target-set. This trend is observed even for targets with a distinct preference for lipophilic compounds, that is, successful molecules have physicochemical properties within specified limits.

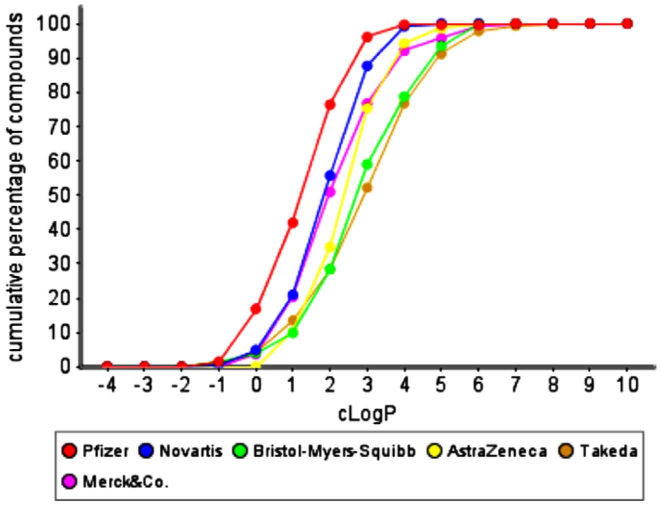

Figure 8. The cumulative percentage of DPP4 bioactive compounds versus cLog *P* for several companies.

For three different targets (CCR5, DPP4 and REN) the properties of bioactive compounds coming from several companies have been analysed. While for CCR5 we notice significant differences in cLog *P* between different companies, this was not the case for the other two examples.

The analysis of per-target physicochemical property distributions (Figs. 2 and 3) showed that for about 90% of the targets associated with marketed drugs and/or clinical candidates the drugs have, on average, lower molecular weight and cLog *P* than the corresponding clinical candidates and bioactive compounds. Even for targets where we do not observe this general trend, like CNR1, we find drugs in the lower physiochemical property value range of the compound set. The large variations of the physicochemical properties seen within a single reported bioactivity suggest that for most targets one can effectively separate bioactivity SAR from physicochemical properties. The clinical candidates reported for CNR1 are a case in point; there is a spread more than five log-units in cLog *P* for the optimised compounds.

We believe that this analysis highlights the findings in the literature of the importance of monitoring physicochemical properties to reduce the attrition in clinical trials and are in agreement with earlier studies based on smaller datasets.[2,15,16] Furthermore, the different trajectories of physicochemical properties seen for different companies working on the same target underscores the importance of aggressively selecting leads for drug projects to effectively control and design compound properties.

Acknowledgements

We thank Dr. Jin Li, Dr. Tim Perkins and Dr. Claire Gavaghan for helpful discussions and comments.

Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.bmcl.2009.10.068.

References and notes

1. Lipinski, C. A. *Drug Discovery Today: Technol.* **2004**, *1*, 337.
2. Wenlock, M. C.; Austin, R. P.; Barton, P.; Davis, A. M.; Leeson, P. D. *J. Med. Chem.* **2003**, *46*, 1250.
3. Vieth, M.; Siegel, M. G.; Higgs, R. E.; Watson, I. A.; Robertson, D. H.; Savin, K. A.; Durst, G. L.; Hipskind, P. A. *J. Med. Chem.* **2004**, *47*, 224.
4. Proudfoot, J. R. *Bioorg. Med. Chem. Lett.* **2005**, *15*, 1087.
5. Vieth, M.; Sutherland, J. J. *J. Med. Chem.* **2006**, *49*, 3451.
6. Gleeson, M. P. *J. Med. Chem.* **2008**, *51*, 817.
7. Southan, C.; Varkonyi, P.; Muresan, S. *Curr. Top. Med. Chem.* **2007**, *7*, 1502.
8. Descriptions of the various databases used in this study are available from www.gvkbio.com/informatics.html (accessed Feb 23, 2009).
9. Varkonyi, P.; Hoppe, C.; Muresan, S. ChemAxon European User Group Meeting May 7–8, 2008, Visegrad, Hungary www.chemaxon.com/forum/download4183.pdf (accessed Feb 23, 2009).
10. BioByte www.biobyte.com CLog *P* Version 4.3 (accessed Feb 23, 2009).
11. Labute, P. *J. Mol. Graph. Model.* **2000**, *18*, 464.
12. Downs, G. M. Ring Perception *in* Encyclopedia of Computational Chemistry; Wiley: Chichester, 1998; p 2509.
13. Accelrys Software Inc. www.accelrys.com/products/scitegic (accessed Feb 23, 2009).
14. JMP SAS Institute. www.jmp.com (accessed Feb 23, 2009).
15. Leeson, P. D.; Springthorpe, B. *Nat. Rev. Drug Disc.* **2007**, *6*, 881.
16. Hughes, J. D.; Blagg, J.; Price, D. A.; Bailey, S.; DeCrescenzo, G. A.; Devraj, R. V.; Ellsworth, E.; Fobian, Y. M.; Gibbs, M. E.; Gilles, R. W.; Greene, N.; Huang, E.; Krieger-Burke, T.; Loesel, J.; Wager, T.; Whiteley, L.; Zhang, Y. *Bioorg. Med. Chem. Lett.* **2008**, *18*, 4872.