Meta-analysis of molecular property patterns and filtering of public datasets of antimalarial “hits” and drugs

Sean Ekins^{a,b,c,d,} and Antony J. Williams^e*

^a Collaborative Drug Discovery, 1633 Bayshore Highway, Suite 342, Burlingame, CA

94010. Tel: 215-687-1320; E-mail: sekins@collaborativedrug.com,

ekinssean@yahoo.com

^b Collaborations in Chemistry, 601 Runnymede Avenue, Jenkintown, PA 19046, USA;

^c Department of Pharmaceutical Sciences, University of Maryland, Baltimore, MD, USA.

^d Department of Pharmacology, Robert Wood Johnson Medical School, University of Medicine & Dentistry of New Jersey, Piscataway, New Jersey 08854, USA;

^e Royal Society of Chemistry, 904 Tamaras Circle, Wake Forest, NC-27587.

Summary

Neglected infectious diseases such as tuberculosis (TB) and malaria kill millions of people annually and the oral drugs used are subject to resistance requiring the urgent development of new therapeutics. Several groups, including pharmaceutical companies, have made large sets of antimalarial screening hit compounds and the associated bioassay data available for the community to learn from and potentially optimize. We have examined both intrinsic and predicted molecular properties across these datasets and compared them with large libraries of compounds screened against *Mycobacterium tuberculosis* in order to identify any obvious patterns, trends or relationships. One set of antimalarial hits provided by GlaxoSmithKline appears less optimal for lead optimization compared with two other sets of screening hits we examined. Active compounds against both diseases were identified to have larger molecular weight (~350-400) and logP values of ~4.0, values that are, in general, distinct from the less active compounds. The antimalarial hits were also filtered with computational rules to identify potentially undesirable substructures. We were surprised that approximately 75-85% of these compounds failed one of the sets of filters that we applied during this work. The level of filter failure was much higher than for FDA approved drugs or a subset of antimalarial drugs. Both antimalarial and antituberculosis drug discovery should likely use simple available approaches to ensure that the hits derived from large scale screening are worth optimizing and do not clearly represent reactive compounds with a higher probability of toxicity *in vivo*.

Introduction

Neglected infectious diseases such as tuberculosis (TB) and malaria kill over two million people annually ¹ while estimates suggest that over 2 billion individuals are infected with *Mycobacterium tuberculosis* (Mtb) alone ². These statistics represent both enormous economic and healthcare challenges for the countries and governments affected while these diseases are generally not the focus for large pharmaceutical companies.

Subsequently, research on these neglected diseases in general, and malaria in particular, is largely comprised of the disjointed efforts of many academic and other non-profit laboratories distributed across the globe. These many independent efforts, while providing significant contributions, often lack the project management, data handling, and pipeline integration functions that are critical to efficiently discovering, developing and bringing new drugs to market. These are generally integrated functions found in the pharmaceutical industry, alongside many researchers experienced in drug development. In recent years non-profit organizations have stepped into the void to manage, coordinate and fund such efforts. Such organizations include the Medicines for Malaria Venture (<http://www.mmv.org/>), the TB Alliance (<http://www.tballiance.org/home/home.php>) and the drugs for neglected diseases initiative (<http://www.dndi.org/>). Pharmaceutical company contributions to these efforts, while not necessarily negligible, are rarely shared publicly until development issues halt project development. We are however seeing more partnering with non-profits to take clinical candidates into large clinical trials and share the associated burden of costs. There have been recent developments in providing the neglected disease community with both collaborative tools and databases to integrate

drug discovery efforts together into effective virtual pharmaceutical organizations that can efficiently deliver drug candidates for further development ³⁻⁵. The urgency to develop new drugs is obvious as antimalarial resistance has led to a re-emergence of the disease in areas once controlled. Of particular concern are the chloroquine resistant (CQR) Plasmodium strains, which has resulted in an increase in malaria mortality ¹. Even the artemisinins are subject to resistance as noted on the Thai-Cambodia border and has lead to new World Health Organization guidelines ⁶.

The efforts around screening for neglected diseases like malaria and TB have, in recent years, significantly increased to the point that very large datasets from hundreds of thousands to over a million compounds in some cases are now routinely tested ⁷⁻¹⁰. These datasets have led to the assessment of what molecular properties may be used to parameterize hits or lead compounds in the case of TB ^{5, 11}. For example, in a previous study we have compared actives and inactives against Mtb in a dataset containing over 200,000 compounds ⁵. The mean molecular weight (357 ± 85), logP (3.6 ± 1.4) and rule of 5 alerts (0.2 ± 0.5) were statistically significantly (based on *t*-test) higher in the most active compounds, while the mean PSA (83.5 ± 34.3) was slightly lower compared to the inactive compounds for the single point screening data ⁵. To date we have assessed 15 different datasets for TB extracted from publications, obtained from screening groups or generated through our own manual annotation of the scientific literature and patents ¹¹. These compounds include known drugs against Mtb as well as screening hits and leads. Our most recent analysis for TB used a dataset consisting of 102,633 molecules screened by the same laboratory against Mtb ¹¹. We were able to analyze the molecular properties, differentiate the actives from the inactives and show that the actives had statistically

significantly (based on *t*-test) higher values for the mean logP (4.0 ± 1.0) and rule of 5 alerts (0.2 ± 0.4), while also having lower HBD count (1.0 ± 0.8), atom count (41.9 ± 9.4) and lower PSA (70.3 ± 29.5) than the inactives ¹¹. While two recent landmark studies ^{9, 10} have provided large datasets of antimalarial compounds that were broadly described as drug-like this can have a broad definition ¹²⁻¹⁸ and in one case the drug-like compounds were suggested to be larger and more hydrophobic than the starting screening collection (an average molecular weight of 446 and logP of about 5.0 ⁹). As fundamentally obvious as this would appear to anyone from the pharmaceutical industry, we are not aware of any similar comprehensive analyses of physicochemical properties across multiple datasets performed on compounds screened for activity against *Plasmodium falciparum* or other *plasmodium* species. This type of meta-analysis is likely to be more revealing than analysis of a single dataset. Knowing the optimum physical properties would at least allow academic researchers to focus their efforts on screening compounds as close as possible to the desired values using calculations that can be readily performed. However it is important to note that as with any rules, guidelines or filters there may be compounds that break them that are still of interest, e.g. large antibacterials, prodrugs, active metabolites etc ^{19, 20}.

We have also applied chemical rules as filters to the hit molecules against Mtb which are widely used by pharmaceutical companies to enable removal or flagging of undesirable molecules, false positives and frequent hitters from HTS screening libraries as well as select compounds from commercial vendors ²¹. Examples of such widely used substructure filters include REOS from Vertex ¹⁵, filters from GSK ²², BMS ²³ and Abbott ²⁴⁻²⁶. These filters in particular pick up a range of undesirable chemical substructures such

as thiol traps and redox-active compounds, epoxides, anhydrides, and Michael acceptors. Reactivity can be defined as the ability to covalently modify a cysteine moiety in a surrogate protein ²⁴⁻²⁶. One group has recently developed a series of over 400 substructural features for removal of Pan Assay INterference compoundS (PAINS) from screening libraries ²⁷. While such filters are widely available to the pharmaceutical industry researchers to readily screen 100,000's of compounds there is no capability for academics to access all these rule sets and screen large libraries. Even the recently available Smartsfilter website resource (<http://pasilla.health.unm.edu/tomcat/biocomp/smartsfilter>) used in this study, allows a maximum of 5000 compounds. With the recent publication and open availability of several sets of malaria hits ^{9, 10} in ChEMBL²⁸, PubChem ²⁹ and CDD ⁴ it was decided to analyze them based on available filters and molecular descriptors to evaluate whether there were any common features. In addition we compared the malaria hits and datasets screened against Mtb ¹¹, to potentially develop a further understanding of the influence of physicochemical properties on compounds with activity against these neglected diseases.

Experimental Methods

CDD Database

The development of the CDD database (Collaborative Drug Discovery Inc. Burlingame, CA) has been described previously in detail with applications for collaborative malaria research⁴.

Datasets

Screening datasets were collected and uploaded in CDD TB from sdf files and mapped to custom protocols (Table 1) (see: <http://www.collaborativedrug.com/register>)¹¹. The malaria data were obtained as previously described^{9,30}. We have also used the Microsource US Drugs database (<http://www.msdiscovery.com/>).

Descriptors

The various datasets were compared using simple calculated molecular properties including logP, hydrogen bond donor, hydrogen bond acceptor, Lipinski rule of 5 alerts, polar surface area, molecular weight, rotatable bonds, and atom counts, calculated using the Marvin plugin (ChemAxon, Budapest, Hungary) within the CDD database. Datasets with molecular properties were readily exported from the CDD database to sdf files and excel files for use with other statistical or modeling software (see below).

SMARTS Filters

The Abbott ALARM²⁴, Glaxo²² and Pfizer LINT SMARTS (also called the Blake filters, <http://pasilla.health.unm.edu/tomcat/biocomp/smartsfilter?help=TRUE>)³¹) filter calculations were performed through the Smartsfilter web application, kindly provided by the Division of Biocomputing, Dept. of Biochemistry & Molecular Biology, University of New Mexico, Albuquerque, NM, (<http://pasilla.health.unm.edu/tomcat/biocomp/smartsfilter>). This software identifies the number of compounds that pass or fail any of the filters implemented. Each filter was evaluated individually with each set of compounds.

Results

Three datasets of antimalarial screening hits were evaluated with both simple molecular properties calculated in CDD using Chemaxon ⁴ (Table 1) and also multiple filters for undesirable features using the Smartsfilter website (Table 2) incorporating rules widely used by at least 3 pharmaceutical companies. Additional datasets of drugs and antimalarial compounds were used as comparators.

Molecular property analysis of antimalarial datasets

The GSK antimalarial hits dataset ⁹ stands out from the other datasets in terms of physicochemical properties (Table 1). The mean molecular weight, logP and number of rotatable bonds are much higher than in the St. Jude ¹⁰ and Novartis datasets of antimalarial compounds ²⁸. The GSK dataset is much closer to the mean property values for the subset of 165 FDA drugs from the Johns Hopkins University set of compounds screened against several drug targets ³²⁻³⁵ that were more active against malaria (although the standard deviations around these properties are very large compared to the other datasets). The St Jude and Novartis antimalarial compound datasets have almost identical mean molecular properties which are much closer to the widely accepted values for “lead-like” compounds (MW < 350, logP < 3) ^{36, 37} compared with the GSK data.

Filtering the antimalarial datasets for undesirable compounds.

The GSK, St Jude and Novartis datasets have very high failure rates with the Abbott Alerts ^{24, 26} (75- 85%) and Pfizer Lint filters (40-57%) (Table 2). The failures with the GSK filters ²² are generally lower as seen previously (< 7.5%) ^{11, 30}. The subset of 165

active antimalarial compounds in the Johns Hopkins dataset has an enrichment of filter failures compared to the total Johns Hopkins dataset of drugs. What stands out for the Johns Hopkins set is the much lower percentage of failures with the Abbott filters (63%) which is close to the total drug dataset or the Microsource drugs dataset (Table 2). It would appear that a general trend for those compounds active against malaria across all datasets is the high level of failures relative to the various pharmaceutical company filters and, in particular, the Abbott filters. This may be antimalarial mechanism related or a limitation in the starting libraries used. The latter is more likely as there are 3 independent datasets as well as the set of compounds that includes many FDA drugs from Johns Hopkins.

Surprisingly, a set of 14 FDA approved widely used antimalarial drugs (amodiaquine, amopyroquine, artesunate, atovaquone, proguanil, chloroquine, halofantrine, hydroxychloroquine, mefloquine, pentaquine, primaquine, pyrimethamine, quinacrine and quinine) has properties much closer to the St Jude and Novartis hits (Table 1). These compounds had fewer failures with the Abbott filters when compared to the GSK, Novartis and St. Jude datasets. This suggests that the mean molecular descriptor values and filter failure profiles for at least 2 out of the 3 large malaria active compound datasets are close to known drugs, and these may be focused on as more desirable in future screening campaigns and for lead optimization.

Discussion

We have previously analyzed the GSK dataset of antimalarial compounds alone (Table 1)^{11, 30} and highlighted the high percentage that fail the Pfizer and Abbott filters and

compared it with a set of US FDA drugs from the Microsource database (Table 1), the Mtb active compounds and other literature examples ³⁸. Many companies avoid compounds that have reactive groups prior to screening and the availability and use of such filters is common. This is not however the case in academia (where the research in neglected diseases is predominantly performed) unless you have access to core cheminformatics resources. Similarly, academic groups rarely analyze the calculated physicochemical properties of the libraries of compounds tested which would allow them to focus on a narrower range and improve their chances of finding active compounds that are better optimization starting points (with a lower probability of failure). The GSK screening hits are described as large and very hydrophobic ⁹ which others would suggest as presenting a significant solubility and absorption challenge ¹⁷. These mean molecular properties were not “lead-like” but were closer to “natural product lead-like” rules ³⁹ which is in marked contrast to the GSK paper ⁹ which describes the compounds as “drug-like”. We suggested that these GSK antimalarial hits are also vastly different to the mean molecular properties of compounds that have shown activity against Mtb ¹¹, which are generally of lower molecular weight, less hydrophobic and have fewer rotatable bonds ⁵.

Our further analysis using two additional large datasets of antimalarial compounds and FDA approved drugs tested for antimalarial activity, as well as known FDA approved drugs, suggests that the GSK data may represent a more difficult starting points for lead optimization. For example, the GSK dataset ⁹ has mean molecular weight, logP and number of rotatable bond values that far higher than those in the St. Jude ¹⁰ and Novartis datasets of antimalarial compounds ²⁸ evaluated in this study. Interestingly the St Jude and Novartis datasets have almost identical mean molecular properties that are

closer to desirable “lead-like” characteristics ^{36, 37}. While all the antimalarial datasets (GSK, St Jude and Novartis) have very high failures with the Abbott Alerts (Table 2), this is perhaps a point of concern when compared to the FDA approved drugs or FDA approved antimalarials, as it indicates that all of these datasets of recently screened compounds have a high percentage of potentially thiol reactive compounds. A recent analysis by us suggests that compounds known to cause drug induced liver injury also have a relationship with these types of filters such that they can be used as a partial predictor for this toxicity (data not shown). Compounds failing the Abbott alerts may have a high probability of failure and toxicity. As stated earlier, the antimalarial mechanism of action may require such reactive compounds, however historically, out of 14 FDA approved widely used antimalarial drugs much lower numbers of filter failures were seen. This suggests that it is possible to develop antimalarials that pass the filters. Out of the 3 openly available datasets, the St Jude and Novartis hits are closer to the ideal starting points for lead optimization as defined by others. One suggestion from this combined work is that such reactivity filters or rules should be more widely instituted for groups working in neglected diseases before they embark on large library screening so that they may be alerted to potential false positives beforehand. The data we have provided on pharmaceutical rule failures are currently not available at any of the website repositories which host these 3 antimalarial datasets, however in one case we have suggested how they might be added into the CDD database ¹¹, but an alternative may be via linkage to the Smartsfilter website. One deficit we have noticed is the Smartsfilter website does not identify which substructures failed, instead just a pass or fail score is associated with a molecule. Undoubtedly knowing why a compound failed would be

instructive. As the neglected disease screening datasets are further evaluated, it is likely that such filtering results will be useful for others and should ideally be stored alongside the screening data.

Conclusion

Within a short space of time three large screening datasets of antimalarial hits have become openly available and hosted in three well known databases and we are also seeing deposition in other databases like ChemSpider. This offers the availability of further calculated properties and links to other information that are unavailable at any of the other databases. Two of these antimalarial datasets have been provided by pharmaceutical companies (GSK and Novartis) and this represents something of a breakthrough in releasing data to the neglected disease research community. To our knowledge there has been no collective analysis of these data from either a molecular properties or undesirable features perspective. This is important before further resources are put into optimization of any of the resulting hits. We, and others, have already described how important it is not only to ensure the quality of any data made available to the research community including chemical structure verification ^{21, 30}, but also the chemical properties that can identify potentially undesirable problems with molecules whether this be poor solubility or toxicity etc. While others have identified problems in other sets of compounds caused by aggregation ⁴⁰, false positives ⁴¹⁻⁴⁷ or artifacts ⁴⁸ in screening libraries these can be pre-filtered and it is not appropriate that the screeners should remain ignorant of such liabilities any longer. The weight of evidence from the datasets we have evaluated suggests that although FDA approved drugs are not ideal, the most conservative filter in the form of the Abbott alerts used in this study routinely fails a

larger percentage of the compounds in the antimalarial hit datasets than in known drugs or antimalarial compounds and this should be of concern. We have also seen a similar pattern with hits against Mtb also failing a very high percentage of these alerts (81-92%) compared to known Mtb drugs (54%)¹¹. While the approximately 13,500 GSK compounds⁹ have higher calculated mean molecular weight and logP³⁰, it is clear that the Novartis and St Jude datasets are much closer to the mean values of the Mtb actives. This would suggest to us that these libraries may also be quickly repurposed or, at the very least, prioritized for screening against Mtb (after filtering of reactive compounds) as they cover similar molecular property space. We have previously described how computational models can be used to enrich screening libraries with Mtb actives and enable more efficient screening and identification of hits^{5, 11}. The addition of physicochemical property and reactive compound alerts filtering will also be useful selection criteria for compounds to follow up.

Large compound libraries screened against Mtb and *P. Falciparum* show that active compounds have higher mean molecular weights and logP values^{5, 9, 11} and, in the majority of cases, the overlap in these values is near identical. Compounds screened against *P. Falciparum* have a high proportion of compounds that fail the Abbott filters for reactivity when compared to drugs and antimalarials which is in agreement with our observations for compounds active against Mtb¹¹ and these compounds should be carefully studied before further optimization. Understanding the chemical properties and characteristics of compounds used against Mtb and malaria may assist in the selection of better compounds for lead optimization.

Abbreviations: CDD, Collaborative Drug Discovery; GSK, GlaxoSmithKline; HBA, hydrogen bond acceptor; HBD, hydrogen bond donor; RBN, rotatable bond number;

Acknowledgements

The authors thank Dr. Jeremy Yang and colleagues (University of New Mexico) for kindly providing access to the Smartsfilter web application and Dr David J. Sullivan (Johns Hopkins University) for providing the dataset of drugs tested against Malaria. We gratefully acknowledge the many groups that have provided antimalarial datasets. S.E. acknowledges colleagues at CDD for developing the software and assistance with large datasets and our collaborators.

Competing interests: Sean Ekins is a consultant for Collaborative Drug Discovery Inc. on a Bill and Melinda Gates Foundation Grant#49852 “Collaborative drug discovery for TB through a novel database of SAR data optimized to promote data archiving and sharing” He is also on the advisory board for ChemSpider. Antony Williams is employed by the Royal Society of Chemistry which owns ChemSpider and associated technologies.

Table 1. Mean $\pm$ SD of molecular descriptors from the CDD database for the malaria and drug datasets. MW = molecular weight, HBD = Number of Hydrogen bond donors, HBA = Number of Hydrogen bond acceptors, Lipinski = Rule of 5 score, PSA = polar surface area, RBN = Number of rotatable bonds. Molecular properties were calculated using the Marvin plug-in (ChemAxon, Budapest, Hungary) within the CDD database. *The analysis for the GSK dataset is in press ³⁰ and has been compared to Mtb active datasets in a separate study ¹¹.

Dataset	MW	logP	HBD	HBA	Lipinski rule of 5 alerts	PSA ($\text{\AA}^2$)	RBN
GSK data (N = 13,471)*	478.2 $\pm$ 114.3	4.5 $\pm$ 1.6	1.8 $\pm$ 1.0	5.6 $\pm$ 2.0	0.8 $\pm$ 0.8	76.8 $\pm$ 30.0	7.2 $\pm$ 3.4
St Jude (N = 1524)	385.3 $\pm$ 71.2	3.8 $\pm$ 1.6	1.1 $\pm$ 0.8	4.9 $\pm$ 1.8	0.2 $\pm$ 0.4	72.2 $\pm$ 29.3	5.2 $\pm$ 2.3
Novartis (N = 5695)	398.2 $\pm$ 105.3	3.7 $\pm$ 2.0	1.2 $\pm$ 1.1	4.7 $\pm$ 2.1	0.4 $\pm$ 0.7	74.7 $\pm$ 37.9	5.6 $\pm$ 3.0
Johns Hopkins All FDA drugs (N = 2615)	349.1 $\pm$ 355.8	1.2 $\pm$ 3.4	2.4 $\pm$ 4.6	5.1 $\pm$ 5.5	0.3 $\pm$ 0.8	96.0 $\pm$ 139.8	5.4 $\pm$ 9.6
Johns Hopkins Subset >	458.0 $\pm$ 298.6	2.2 $\pm$ 2.7	2.1 $\pm$ 3.4	5.4 $\pm$ 4.7	0.6 $\pm$ 0.9	90.6 $\pm$ 104.4	7.1 $\pm$ 7.7

50% malaria inhibition at							
96h (N = 165)							
<hr/>							
Antimalarial drugs (N =	341.6 ± 67.0	3.8 ± 1.6	1.8 ± 1.0	5.3 ± 1.5	0.2 ± 0.6	53.4 ± 21.2	5.8 ± 3.0
14)							

Table 2. Summary of SMARTS filter failures for various datasets. The Abbott ALARM²⁴, Glaxo²² and Blake³¹ SMARTS filter calculation were performed through the Smartsfilter web application, Division of Biocomputing, Dept. of Biochem & Mol Biology, University of New Mexico, Albuquerque, NM, (<http://pangolin.health.unm.edu/tomcat/biocomp/smartsfilter>). The GSK malaria screening data was obtained⁹ from the CDD database. The St Jude malaria data was obtained from¹⁰. The Novartis dataset was obtained from ChEMBL²⁸. We also used the Microsource US Drugs dataset as a reference set of “drug-like” molecules. Large datasets > 1000 molecules were fragmented into smaller sdf files before running through this website. *The analysis for the GSK and Microsource datasets is in press³⁰ and has been compared to Mtb active datasets in a separate study¹¹.

Dataset (N)	Number failing Abbott ALARM filters²⁴ (%)	Number failing Pfizer LINT filters * (%)	Number failing Glaxo filters²² (%)
GSK Malaria hits. (13,355)*	10124 (75.8)	7683 (57.5)	129 (0.01)
St Jude (N = 1524)	1291 (84.7)	621 (40.7)	83 (5.4)
Novartis (N = 5695)	4542 (79.7)	2371 (41.6)	169 (7.5)
Johns Hopkins –All	1442 (53.5)	1264 (46.9)	401 (14.9)

FDA drugs tested

against malaria (N =

2615)

Johns Hopkins Subset	104 (63.0)	91 (55.2)	41 (24.8)
-----------------------------	------------	-----------	-----------

> 50% malaria

inhibition at 96h (N =

165)

Microsource US FDA	688 (66.1)	516 (49.6)	143 (13.7)
---------------------------	------------	------------	------------

drugs (N = 1041)

Antimalarial drugs (N	8 (57.1)	8 (57.1)	2 (14.3)
------------------------------	----------	----------	----------

= 14)

*Originally provided as a Sybyl script to Tripos by Dr. James Blake (Array Biopharma)

while at Pfizer and also known as the Blake filter

<http://pasilla.health.unm.edu/tomcat/biocomp/smartsfilter?help=TRUE>.

References

1. D. A. Fidock, *Nature*, 2010, **465**, 297-298.
2. T. S. Balganesh, P. M. Alzari and S. T. Cole, *Trends Pharmacol Sci*, 2008, **29**, 576-581.
3. S. Nwaka and R. G. Ridley, *Nat Rev Drug Discov*, 2003, **2**, 919-928.
4. M. Hohman, K. Gregory, K. Chibale, P. J. Smith, S. Ekins and B. Bunin, *Drug Disc Today*, 2009, **14**, 261-270.
5. S. Ekins, J. Bradford, K. Dole, A. Spektor, K. Gregory, D. Blondeau, M. Hohman and B. Bunin, *Mol BioSystems*, 2010, **6**, 840-851.
6. Editorial, *The Lancet*, 2010, **375**, 956.
7. J. A. Maddry, S. Ananthan, R. C. Goldman, J. V. Hobrath, C. D. Kwong, C. Maddox, L. Rasmussen, R. C. Reynolds, J. A. Secrist, 3rd, M. I. Sosa, E. L. White and W. Zhang, *Tuberculosis (Edinburgh, Scotland)*, 2009, **89**, 354-363.
8. S. Ananthan, E. R. Faaleolea, R. C. Goldman, J. V. Hobrath, C. D. Kwong, B. E. Laughon, J. A. Maddry, A. Mehta, L. Rasmussen, R. C. Reynolds, J. A. Secrist, 3rd, N. Shindo, D. N. Showe, M. I. Sosa, W. J. Suling and E. L. White, *Tuberculosis (Edinburgh, Scotland)*, 2009, **89**, 334-353.
9. F.-J. Gamo, L. M. Sanz, J. Vidal, C. de Cozar, E. Alvarez, J.-L. Lavandera, D. E. Vanderwall, D. V. S. Green, V. Kumar, S. Hasan, J. R. Brown, C. E. Peishoff, L. R. Cardon and J. F. Garcia-Bustos, *Nature*, 2010, **465**, 305-310.

10. W. A. Guiguemde, A. A. Shelat, D. Bouck, S. Duffy, G. J. Crowther, P. H. Davis, D. C. Smithson, M. Connelly, J. Clark, F. Zhu, M. B. Jimenez-Diaz, M. S. Martinez, E. B. Wilson, A. K. Tripathi, J. Gut, E. R. Sharlow, I. Bathurst, F. El Mazouni, J. W. Fowble, I. Forquer, P. L. McGinley, S. Castro, I. Angulo-Barturen, S. Ferrer, P. J. Rosenthal, J. L. Derisi, D. J. Sullivan, J. S. Lazo, D. S. Roos, M. K. Riscoe, M. A. Phillips, P. K. Rathod, W. C. Van Voorhis, V. M. Avery and R. K. Guy, *Nature*, 2010, **465**, 311-315.
11. S. Ekins, T. Kaneko, C. A. Lipinski, J. Bradford, K. Dole, A. Spektor, K. Gregory, D. Blondeau, S. Ernst, J. Yang, N. Goncharoff, M. Hohman and B. Bunin, *Molecular bioSystems*, 2010, **In press**.
12. G. Chen, S. Zheng, X. Luo, J. Shen, W. Zhu, H. Liu, C. Gui, J. Zhang, M. Zheng, C. M. Puah, K. Chen and H. Jiang, *Journal of combinatorial chemistry*, 2005, **7**, 398-406.
13. V. V. Zernov, K. V. Balakin, A. A. Ivashchenko, N. P. Savchuk and I. V. Pletnev, *J Chem Inf Compu Sci*, 2003, **43**, 2048-2056.
14. Y. Takaoka, Y. Endo, S. Yamanobe, H. Kakinuma, T. Okubo, Y. Shimazaki, T. Ota, S. Sumiya and K. Yoshikawa, *J Chem Inf Comput Sci*, 2003, **43**, 1269-1275.
15. W. P. Walters and M. A. Murcko, *Adv Drug Del Rev*, 2002, **54**, 255-271.
16. I. Muegge, S. L. Heald and D. Brittelli, *J Med Chem*, 2001, **44**, 1-6.
17. C. A. Lipinski, *J Pharm Toxicol Methods*, 2000, **44**, 235-249.

18. Ajay., W. P. Walters and M. A. Murcko, *J Med Chem*, 1998, **41**, 3314-3324.
19. M. P. Gleeson, *J Med Chem*, 2008, **51**, 817-834.
20. C. A. Lipinski, F. Lombardo, B. W. Dominy and P. J. Feeney, *Adv Drug Del Rev*, 1997, **23**, 3-25.
21. A. J. Williams, V. Tkachenko, C. Lipinski, A. Tropsha and S. Ekins, *Drug Discovery World*, 2009, **10**, **Winter**, 33-38.
22. M. Hann, B. Hudson, X. Lewell, R. Lifely, L. Miller and N. Ramsden, *J Chem Inf Comput Sci*, 1999, **39**, 897-902.
23. B. C. Pearce, M. J. Sofia, A. C. Good, D. M. Drexler and D. A. Stock, *J Chem Inf Model*, 2006, **46**, 1060-1068.
24. J. R. Huth, R. Mendoza, E. T. Olejniczak, R. W. Johnson, D. A. Cothron, Y. Liu, C. G. Lerner, J. Chen and P. J. Hajduk, *J Am Chem Soc*, 2005, **127**, 217-224.
25. J. R. Huth, D. Song, R. R. Mendoza, C. L. Black-Schaefer, J. C. Mack, S. A. Dorwin, U. S. Lador, J. M. Severin, K. A. Walter, D. M. Bartley and P. J. Hajduk, *Chem Res Toxicol*, 2007, **20**, 1752-1759.
26. J. T. Metz, J. R. Huth and P. J. Hajduk, *J Comput Aided Mol Des*, 2007, **21**, 139-144.
27. J. B. Baell and G. A. Holloway, *J Med Chem*, 2010, **53**, 2719-2740.

28. K. Gagaring, R. Borboa, C. Francek, Z. Chen, J. Buenviaje, D. Plouffe, E. Winzeler, A. Brinker, T. Diagen, J. Taylor, R. Glynne, A. Chatterjee and K. Kuhen, ChEMBL-NTD (www.ebi.ac.uk/chemblntd)
29. D. L. Wheeler, T. Barrett, D. A. Benson, S. H. Bryant, K. Canese, V. Chetvernin, D. M. Church, M. DiCuccio, R. Edgar, S. Federhen, L. Y. Geer, W. Helmberg, Y. Kapustin, D. L. Kenton, O. Khovayko, D. J. Lipman, T. L. Madden, D. R. Maglott, J. Ostell, K. D. Pruitt, G. D. Schuler, L. M. Schriml, E. Sequeira, S. T. Sherry, K. Sirotkin, A. Souvorov, G. Starchenko, T. O. Suzek, R. Tatusov, T. A. Tatusova, L. Wagner and E. Yaschenko, *Nucleic Acids Res*, 2006, **34**, D173-180.
30. S. Ekins and A. J. Williams, *Drug Disc Today*, 2010, **In Press**.
31. J. F. Blake, *Medicinal chemistry (Shariqah (United Arab Emirates))*, 2005, **1**, 649-655.
32. C. R. Chong and D. J. Sullivan, Jr., *Nature*, 2007, **448**, 645-646.
33. C. R. Chong, J. Xu, J. Lu, S. Bhat, D. J. Sullivan, Jr. and J. O. Liu, *ACS chemical biology*, 2007, **2**, 263-270.
34. C. R. Chong, X. Chen, L. Shi, J. O. Liu and D. J. Sullivan, Jr., *Nat Chem Biol*, 2006, **2**, 415-416.
35. C. R. Chong, D. Z. Qian, F. Pan, Y. Wei, R. Pili, D. J. Sullivan, Jr. and J. O. Liu, *J Med Chem*, 2006, **49**, 2677-2680.
36. T. I. Oprea, *J Comput Aided Mol Des*, 2002, **16**, 325-334.

37. T. I. Oprea, A. M. Davis, S. J. Teague and P. D. Leeson, *J Chem Inf Comput Sci*, 2001, **41**, 1308-1315.
38. P. Axerio-Cilies, I. P. Castaneda, A. Mirza and J. Reynisson, *Eur J Med Chem*, 2009, **44**, 1128-1134.
39. J. Rosen, J. Gottfries, S. Muresan, A. Backlund and T. I. Oprea, *J Med Chem*, 2009, **52**, 1953-1962.
40. J. Seidler, S. L. McGovern, T. N. Doman and B. K. Shoichet, *J Med Chem*, 2003, **46**, 4477-4486.
41. G. M. Rishton, *Curr Opin Chem Biol*, 2008, **12**, 340-351.
42. G. M. Rishton, *Medicinal chemistry (Sharjah (United Arab Emirates))*, 2005, **1**, 519-527.
43. T. I. Oprea, C. G. Bologa, S. Boyer, R. F. Curpan, R. C. Glen, A. L. Hopkins, C. A. Lipinski, G. R. Marshall, Y. C. Martin, L. Ostopovici-Halip, G. Rishton, O. Ursu, R. J. Vaz, C. Waller, H. Waldmann and L. A. Sklar, *Nat Chem Biol*, 2009, **5**, 441-447.
44. K. E. Coan and B. K. Shoichet, *J Am Chem Soc*, 2008, **130**, 9606-9612.
45. B. Y. Feng, A. Simeonov, A. Jadhav, K. Babaoglu, J. Inglese, B. K. Shoichet and C. P. Austin, *J Med Chem*, 2007, **50**, 2385-2390.

46. A. Jadhav, R. S. Ferreira, C. Klumpp, B. T. Mott, C. P. Austin, J. Inglese, C. J. Thomas, D. J. Maloney, B. K. Shoichet and A. Simeonov, *J Med Chem*, **53**, 37-51.
47. A. K. Doak, H. Wille, S. B. Prusiner and B. K. Shoichet, *J Med Chem*.
48. C. Schmidt, *Nat Biotechnol*, 2010, **28**, 185-186.