Published in final edited form as:
Nat Rev Drug Discov. 2011 March ; 10(3): 197–208. doi:10.1038/nrd3367.

Probing the links between *in vitro* potency, ADMET and physicochemical parameters

M. Paul Gleeson[*], **Anne Hersey**[‡], **Dino Montanari**[§,‖], and **John Overington**[‡]

[*]Department of Chemistry, Faculty of Science, Kasetsart University, 50 Phaholyothin Rd, Chatuchak, Bangkok 10900, Thailand [‡]EMBL-EBI European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridgeshire, CB10 1SD, United Kingdom [§]Neurosciences CEDD, GlaxoSmithKline Medicines Research Centre, Via A. Fleming, 2, 37135, Verona, Italy

Abstract

A common underlying assumption in current drug discovery strategies is that compounds with higher *in vitro* potency at their target(s) have greater potential to translate into successful, low-dose therapeutics. This has led to the development of screening cascades with *in vitro* potency embedded as an early filter. However, this approach is beginning to be questioned, given the bias in physicochemical properties that it can introduce early in lead generation and optimization, which is due to the often diametrically opposed relationship between physicochemical parameters associated with high *in vitro* potency and those associated with desirable absorption, distribution, metabolism, excretion and toxicity (ADMET) characteristics. Here, we describe analyses that probe these issues further using the ChEMBL database, which includes more than 500,000 drug discovery and marketed oral drug compounds. Key findings include: first, that oral drugs seldom possess nanomolar potency (50 nM on average); second, that many oral drugs have considerable off-target activity; and third, that *in vitro* potency does not correlate strongly with the therapeutic dose. These findings suggest that the perceived benefit of high *in vitro* potency may be negated by poorer ADMET properties.

The past two decades have seen the evolution of chemical lead discovery strategies in the pharmaceutical industry, increasingly focusing on speed and efficiency of chemical synthesis and biological screening, in order to satisfy the greater desire for potent, selective small molecules for novel drug targets1. The shift from the historical strategy based on low-throughput *in vivo* pharmacology to a focus on specific molecular targets selected on the basis of disease biology knowledge derived from ‘omics’ approaches — typically explored using high-throughput *in vitro*-dominated screening cascades2 — has not arrested the increase in costs or improved the rate of output of marketed new chemical entities. Data indicate that, of the compounds that are selected as clinical candidates, only ~10% will make it to the market, showing that the overall process is far from optimal3,4.

[‖]Present address: Aptuit Srl, Medicines Research Centre, Via A. Fleming, 4, 37135, Verona, Italy.

Competing interests
The authors declare no competing financial interests.

Europe PMC Funders Author Manuscripts

The current drug discovery process favours candidate molecules with high *in vitro* potency at a single target and good absorption, distribution, metabolism, excretion and toxicity (ADMET) properties. For a molecule to reach the candidate selection stage, it should fit a profile that is established towards the beginning of a drug discovery programme; selection typically requires appropriate target validation data, a compound with high *in vitro* potency (usually in the nanomolar range) at the target, off-target selectivity, good ADMET parameters and no intellectual property issues. One perceived benefit of molecules with higher *in vitro* potency is their assumed greater potential to translate into treatments that achieve their therapeutic effects at low (milligram) doses. This could have the advantage of reducing the risk of more generic off-target interactions with proteins linked to toxic effects (such as the HERG potassium channel and drug-induced QT interval prolongation), as well as longer-term toxicity and risks of adverse drug reactions due to reactive metabolites5. Another advantage of high potency at the intended target is a potentially reduced risk of interactions with related targets.

The pharmaceutical industry has invested considerably in high-throughput screening (HTS) technologies to test large numbers of compounds in a range of biochemical and cell-based assays. However, as the assays measuring *in vitro* potencies (XC_{50} values) often have considerably greater throughput and lower costs per compound than assays used to assess ADMET characteristics, there is a tendency for the former measurements to have a more substantial role in the selection and optimization of a given lead series6. In particular, although some simple *in vitro* ADMET assays have comparable costs and throughput to those for measuring compound affinity or potency and can be employed essentially in parallel, a more thorough ADMET profile, obtained from more complex and lower-throughput *in vivo* assays measuring parameters such as bioavailability and toxicity — which are key determinants of the decision to select a particular lead series — often only become available at later stages in lead optimization. Overall, as not all compounds synthesized during the course of the programme can be tested in each assay owing to prohibitive costs, a screening cascade is typically implemented in which an initial filter is *in vitro* target potency2,7.

Given the considerable difficulty of optimizing potency and ADMET properties side by side owing to their often diametrically opposed relationship with physicochemical properties, a greater emphasis on high *in vitro* potency may make the search for new drugs more challenging. Indeed, several groups of researchers have previously questioned whether suboptimal physicochemical properties that arise during lead optimization are at least partly responsible for the continuing high candidate attrition rates8–18. For example, it has been highlighted that the mean molecular mass of drugs has increased gradually over time11,14,19 and, although it seems we can still find drug molecules in less favourable areas of chemical property space, the key question is whether the industry as a whole might be more productive searching within drug-like space more thoroughly18–21.

In an attempt to understand the underlying basis of these issues further, we have performed a range of analyses on the comprehensive ChEMBL database (see Further information), a publicly available database of drugs, drug-like small molecules and their targets compiled by the European Bioinformatics Institute (EBI) (see Box1 for further information on the full

dataset for these analyses), which is part of the European Molecular Biology Laboratory (EMBL). The database contains manually extracted and curated data from the primary medicinal chemistry and pharmacology literature on >500,000 compounds, with >2.4 million records of their effects on biological systems. As such, it represents an ideal open-access database to perform cheminformatics analyses. We have used this data to investigate the relationship between: first, ADMET parameters and physicochemical parameters; second, *in vitro* potency and physicochemical parameters; third, selectivity and physicochemical parameters; and fourth, therapeutic dose and XC_{50} values for targets implicated in the disease. On the basis of this analysis, we suggest altered guidelines on the levels of potency, selectivity and physicochemical properties for potential oral drugs.

Analysis

In this study, we assess the relationship between biological parameters that are frequently measured in drug discovery programmes and two fundamental molecular properties: molecular mass and lipophilicity (logP). We focus on these two generic molecular properties as they have been shown to be universally relevant to potency and ADMET parameters measured in the pharmaceutical industry8,22,23, and because there are numerous reports of their importance in defining the chemical space of small-molecule therapeutics8–14,16,18–20,24–29. The analyses have been performed on the ChEMBL database, which contains two-dimensional chemical structures, bioactivity data (such as binding constants and pharmacology) and calculated molecular properties, with bioactivity data tagged to show links between molecular targets and published assays (Box 1).

Relationship between ADMET parameters and physicochemical properties

The seminal paper by Lipinski and colleagues8 published in 1997, which introduced the ‘rule of five’ (RO5), highlighted the relationship between poor absorption, solubility and permeability, and increased values of molecular mass, logP and the number of hydrogen-bond donors and acceptors for a dataset of compounds in clinical development. Many reports9–29 since have also indicated that the physicochemical properties associated with desirable ADMET properties correspond to lower molecular mass and logP, more akin to the properties that are characteristic of historical oral drugs27. For example, it was recently shown using proprietary data from GlaxoSmithKline that a range of 15 ADMET parameters that are routinely used in the pharmaceutical industry deteriorate with either increasing molecular mass, clogP or both23. It is also possible to demonstrate this effect on data in the ChEMBL database (see Supplementary Information S1 (figure)).

Further Information
Cerep Bioprint database: http://www.cerep.fr/Cerep/Users/index.asp
ChEMBL database: https://www.ebi.ac.uk/chembl/
Drugbank: http://www.drugbank.ca
JMP 6.0: www.jmp.com
PipelinePilot 8.0: http://www.accelrys.com
Pubmed: http://www.ncbi.nlm.nih.gov/sites/entrez
RXList: http://www.rxlist.com/script/main/hp.asp
Statistica 7: www.statsoft.com
The Journal of Pharmacology and Experimental Therapeutics: http://jpet.aspetjournals.org/

Another way of illustrating the link between oral-drug space and these two key physicochemical properties is by generating a simple score that is derived from the deviation of a compound's properties from the mean values associated with oral drugs (see equation 1). The mean and standard deviation of approved oral drugs have a particular relevance as these compounds have successfully passed through all of the development hurdles. The score for a compound is computed as the modulus (that is, negative numbers arising within the modulus symbols are treated as positive) of the mean minus the molecular mass or AlogP of the compound in question, divided by the standard deviation. It should be noted that only compounds with a molecular mass above the mean are penalized as this property does not show a strong parabolic relationship with ADMET parameters, unlike AlogP. The use of the mean and standard deviation has particular relevance given that these are representative of the boundary of oral drug space with respect to known key properties.

$$Score = \frac{|2.5 - c\log P|}{2.0} + \frac{|330 - MWT\ ^*|}{120} (^* \text{if MWT} < 330,\ \text{MWT} = 0) \quad \text{Equation 1}$$

This score helps to highlight the link between ADMET parameters, molecular mass and logP, and also the difference in properties between oral drugs and compounds from the ChEMBL database (Fig 1a). The compounds from the ChEMBL database can be considered broadly representative of medicinal chemistry space that has been under investigation in industry, or contained in typical screening collections, as it is compiled from reports of drug discovery projects published in the literature (that is, it consists of both active and inactive molecules from structure–activity relationship studies across multiple target classes). The percentage of compounds with scores <1 for ChEMBL-derived literature compounds is almost half that of the oral drugs set19 (31% versus 56%, respectively), as might be expected given the markedly higher molecular masses and AlogP values of the former set. For a score >2, the difference is almost triple (39% versus 14%, respectively). This clearly illustrates the discrepancy in physicochemical properties between the types of compounds contained in the ChEMBL database and oral drugs, which also has implications for ADMET parameters. It also raises the question of why this discrepancy exists and whether it is due to a bias introduced by the focus on *in vitro* potency in the screening cascade.

The relevance of this simple ADMET score is illustrated in Supplementary information S2 (figure), in which compounds that deviate furthest from oral drug space have increased liabilities on average when assessed against a representative set of ADMET parameters (solubility, permeability, bioavailability and cytochrome P450 3A4 inhibition). The combination of both molecular mass and logP into a single score may offer a better way of prioritizing promising compounds or screening out compounds with poor ADMET properties than simple cut-off rules based on each parameter individually. For example, in this scheme, a compound that might be excluded using traditional hard cut-offs based on two parameters used independently — for example, a compound with a molecular mass of 520 and an AlogP of 2, which has a score of 1.8 — could be ranked as preferable to a compound that might pass traditional hard cut-offs for each individual parameter, but close to their extreme limits, making it less likely that the compound has good ADMET properties; for

example, a compound with a molecular mass of 499 and an AlogP of 4.9 has a higher (worse) score of 2.6.

Relationship between potency and physicochemical properties

In vitro potency is typically the key parameter used to select hits from an HTS campaign in lead generation, and it is one of the most important parameters used in lead optimization to select candidates for full preclinical evaluation. The link between general molecular properties and target affinity (which is typically directly related to potency) has been well studied from a functional group30 and property31 perspective, and more recently, the concept of ligand efficiency22,26,32 has been proposed.

It is particularly important that the relationship between *in vitro* potency and physicochemical properties is understood so that any focus on *in vitro* potency early in the screening cascade does not adversely affect the ADMET properties of molecules generated during lead optimization, given that most ADMET assays cannot be done in parallel with potency measurements owing to resource limitations. We therefore analysed the ChEMBL dataset of 201,355 unique compounds, with 402,496 activity measurements across multiple targets, to assess the link between molecular mass, logP and potency.

The analysis was performed in two ways: by unique compound and by unique measurement (that is, the same compounds may be present more than once). However, as both analyses show the same trends, we report the results of the unique-measurements analysis of potency only (FIG. 2). For this dataset, it is apparent that the mean pXC_{50} value shows a strong dependence on the overall molecular mass and logP. Interestingly, we found that a 100 Da increase in molecular mass seems to have a stronger effect on the potency than a 1 log unit increase in logP, although this is target-dependent. Furthermore, from FIG. 2c it can be seen that the effect of both parameters shows a degree of independence, as for a given AlogP value, a concomitant increase in molecular mass increases the mean pXC_{50} value and vice versa. This general effect can also be shown when compounds are divided according to individual chemotype; five examples are highlighted in Supplementary information S3 (figure).

Lead generation and optimization requires a balance of multiple parameters: potency, selectivity, ADMET and physicochemical properties, which will allow a molecule to be absorbed, transported and interact at the required receptor(s), for a sufficiently long period to elicit the required pharmacological response. As noted above, the early focus on *in vitro* target potency in the lead generation step, and its subsequent use as a key driver of programme decisions in lead optimization, may make it more difficult to achieve the optimal balance in potency and ADMET properties given their often diametrically opposed relationship to simple molecular properties12,18. Indeed, an analysis by Wenlock and colleagues27 suggested that the balance seems to be suboptimal: clinical candidates in earlier development phases displayed considerably higher molecular masses and logP values than oral drugs, whereas moving through the phases, the values begin to converge with those molecules that made it to market. Furthermore, it has been shown that, although the molecular mass of marketed drugs has risen gradually over time, logP has essentially remained constant14,18,19,33. In addition, Oprea *et al.* found that compounds being

synthesized in the medicinal chemistry community, exemplified by ~35,000 compounds described in the literature as part of structure–activity relationship investigations, tended to have less desirable properties than oral drugs, and the most potent compounds were even less desirable from a physicochemical perspective29.

We explored the possible implications of drug discovery programmes imposing high *in vitro* potency constraints using a ChEMBL-derived dataset of 201,355 unique compounds with 402,496 measured activities at more than 2,000 unique targets. Table 1 shows the distribution of molecular mass and AlogP for the full ChEMBL dataset, a subset of the ChEMBL dataset with nanomolar potency or lower, and oral drugs19. We also include the proportion of compounds failing the RO5, which is probably the most frequently used filter in the medicinal chemistry community.

Considering a molecular mass cut-off of 400, the majority of oral drugs (79%) have values below this cut-off, compared to less than half of the total ChEMBL set (44%), and close to a quarter for ChEMBL compounds with nanomolar potency (26%) (Table 1). Considering an ALogP cut-off of 4, 77% of oral drugs lie below this value, compared to 58% of the ChEMBL set and a comparable value for the nanomolar potency ChEMBL compounds (50%). Strikingly, only 8% of oral drugs have both a molecular mass >400 and AlogP >4, compared to 30% and 41% of ChEMBL and nanomolar potency ChEMBL compounds, representing an approximately threefold and fivefold increase, respectively, over oral drugs. In addition, only 12% of oral drugs fail two or more of the four RO5 criteria, compared with 14% of the ChEMBL set and 21% of the nanomolar potency ChEMBL set.

It should be noted that we have deliberately focused on molecular mass and logP cut-offs of 4 and 400, respectively, as opposed to the more commonly used values of molecular mass <500 and AlogP <5. This is because the commonly used values are a minimum requirement and are not necessarily representative of ideal oral-drug space25. Of course, the RO5 and its associated molecular mass and logP cut-offs were not intended to define ideal values for oral drugs, but rather to act as an initial filter to remove compounds that were unlikely to be orally absorbed.

Nonetheless, the pharmaceutical industry is increasingly focusing on new targets, the druggability of which is often unknown34, and this fact is commonly used by researchers as the principal reason for the exploration of novel areas of physicochemical-property space. For example, it has been reported that the molecular mass and AlogP of newer oral drugs targeting peptide G protein-coupled receptors (GPCRs) and kinases are considerably higher than the mean of oral drugs35, suggesting there are merits in searching outside of typical drug-like space. However, although there is a distinct difference in the properties of drugs that modulate these different target classes, they are much smaller in magnitude than the differences in properties between oral drugs and ChEMBL compounds of nanomolar potency reported here.

Selectivity versus promiscuity

In many drug discovery programmes over the past two decades, selective molecules for a given target have been sought to reduce the likelihood of unwanted off-target binding and

thus undesired side effects. However, it has become apparent that some off-target activity might prove beneficial when there is substantial redundancy within a given disease-gene network36,37. For example, lapatinib (Tykerb; GlaxoSmithKline), which is used in the treatment of breast cancer, inhibits the kinase activity of two receptor tyrosine kinases in the epidermal growth factor receptor (EGFR) family: EGFR (also known as ERBB1) and human epidermal growth factor receptor 2 (HER2; also known as ERBB2)7.

Efforts to assess the relationship between promiscuity and physicochemical properties have yielded variable results, being dependent on the source and size of the dataset studied and the method of analysis. The promiscuity of a set of 75,000 Pfizer compounds measured in 220 in-house assays was found to decrease with increasing molecular mass38. This was proposed to be consistent with the 'reduced complexity' argument of Hann *et al.*10, although molecular mass is not necessarily a good surrogate for complexity. Similar findings were reported based on Organon's proprietary Scope database39, which consists of a limited dataset of preclinical compounds. However, another group found that 3,138 Novartis compounds tested in 50–79 in-house assays showed the opposite trend — promiscuity increased with increasing molecular mass — with acids being particularly selective40. An analysis by Leeson *et al.* of 200 assays performed on 2,333 compounds from the Cerep Bioprint database (see Further information) suggested a similar trend for molecular mass, and that increasing lipophilicity and the presence of a basic moiety also lead to increased promiscuity18. Subsequent analyses on 213 Roche compounds screened in Cerep found the same trends with respect to lipophilicity and basicity as Leeson *et al.,* but none with respect to molecular mass41. These variable results clearly indicate that further work is needed to better understand the relationship between physicochemical properties and promiscuity.

To try to shed further light on this issue, we extracted all the compounds with at least three measured pXC_{50} values from the ChEMBL database (N = 40,408, number of unique target activities = 191,417, >500 unique targets). The advantages of the ChEMBL dataset used here is that the analysis is performed on confirmed pXC_{50} values derived from the medicinal chemistry literature, rather than relying on percentage inhibition values at single concentrations that have not necessarily been confirmed by more rigorous follow-up analyses. However, although the number of compounds used in the analysis is considerable, and the use of pXC_{50} values preferred, a clear limitation of the ChEMBL dataset is the low number of measured pXC_{50} values per compound. Nevertheless, this does not mean that meaningful trends cannot be ascertained from an analysis of this dataset.

An assessment of the dependency of promiscuity on molecular mass is shown in Fig 3a. The number of unique hits (defined here as a compound with $pXC_{50} \geq 6$ (≤ 1 μM potency) for a particular target) shows only a weak relationship with molecular mass; however, the total number of measurements per molecular mass bin decreases slightly with increasing molecular mass, which could mask the true effect of this molecular property. The number of hits seems to show a dependence on the number of targets against which molecules in a given property bin have been measured; that is, oral drugs (which are expected to have lower molecular mass) may have been tested more often, leading to greater apparent promiscuity of compounds with low molecular mass. It therefore seems preferable to look at both the

number of hits (≤1 μM potency) for a given property bin and the ratio of hits with micromolar potency to total measurements per compound.

If we focus on the ratio of hits with micromolar potency to actual measurements, we see that as the molecular mass increases, the probability of hitting another target increases18,40. When normalized for the total number of measurements, compounds with molecular mass <200 have a micromolar activity ~35% of the time on average (Fig 3a). This compares to ~75% for molecules with molecular mass >500.

The relationship with logP and promiscuity is displayed in Fig 3b. Looking first at the mean number of hits with ≤1 μM potency, we see that they also tend to decrease as AlogP increases; however, this is again an artefact of the total number of measurements associated with each logP bin. If we instead look at the ratio of the hits with micromolar potency to the total number of measurements, we see a general trend of increasing promiscuity with AlogP, consistent with other reports18,41.

With this dataset, we observe that acidic molecules tend to be less promiscuous than neutral or basic molecules (fig 3c) (acids < zwitterions < neutrals = bases). However, it is apparent from Fig 3c that there is no statistically significant difference between bases and neutral molecules. This is an artefact of the underlying distribution of molecular mass within the given ionization state categories. Breaking the data down by molecular mass reveals that there is a clear difference between basic and neutral molecules18,41, at least for molecular mass <400 (Fig 4). Interestingly, the difference in promiscuity between bases and neutral molecules seems to tail off at higher molecular mass.

The use of datasets of different sizes and chemical coverage, different promiscuity cut-offs based on percentage activity or pXC_{50} values, the range and number of targets used, as well as presumably different assay technologies (binding, functional, cellular and so on) are all factors that will complicate such analyses. Although the ChEMBL database is lacking in terms of the biological space covered per compound compared to the more biologically ideal Cerep dataset, the trends observed are in good overall agreement with the most recent analysis of this data18. It is clear nevertheless that analyses on more comprehensive datasets are required to shed further light on this issue

Characteristics of oral drugs

For many indications, the goal for a new drug therapy is administration via the oral route, once daily, to improve convenience for patients and consequently adherence. It is also preferred that the dose itself is <10 mg per day to reduce cost of goods and to maximize the therapeutic window. To investigate how many oral drugs match these criteria and to understand the potential value of high *in vitro* potency in the successful development of oral drugs, we first extracted information on the set of 792 oral drugs with dosing information from the CHEMBL database for which the dose was administered as a capsule, tablet or suspension. Where possible (see below), we also matched the doses and therapeutically relevant pXC_{50} values. We hypothesize that the mean values of these parameters will have relevance in defining more suitable benchmark values for drug discovery in general and

might help to illuminate the nature of the relationship between *in vitro* potency and the dose, which is possibly a more relevant measure of *in vivo* efficacy. As the distribution of both the therapeutically relevant pXC_{50} values and oral doses show non-normal distributions, we report the median values for all discussion regarding Fig 5.

Analysis of the dataset of 792 oral drugs shows that the median value associated with the mean formulated doses is 92 μmol, or 34 mg, with the median of the lowest (minimum) dose reported being 55 μmol (20 mg), and the highest 148 μmol (55 mg) (Fig 5a). A more detailed analysis of the distribution shows that 62% of the 792 oral drugs have a minimum formulated dose >10 mg, and 36% have a value >50 mg.

An analysis of the *in vitro* potencies of oral drugs is much more challenging because the mode of action of many is unknown, or multiple targets are involved, making the assignment of individual potency values difficult. To estimate the potency distribution of oral drugs, we focused on the 792 oral drugs already extracted, because we were also interested in assessing the correlation between dose and potency. For each drug, we searched the ChEMBL database for any reported activities at any targets. From this search, we obtained activity values for 392 of the oral drugs at many diverse targets; however, not all of these target activities were related to the therapeutic targets of the drug in question. We subsequently determined the therapeutic target for each oral drug from a search of literature sources, which then allowed us to manually identify whether any therapeutically relevant activities had been reported for each drug. This resulted in a dataset of 261 drugs in total for the analysis. In cases in which more than one target is implicated (~30%), we calculate the mean, maximum and minimum pXC_{50} values. In a small number of cases in which only a subset of the relevant activities was present, the compounds were still considered in the analysis.

The distribution of the largest therapeutically relevant pXC_{50} values reported for each molecule has a mean value of 7.3 and a median of 7.7 (Fig 5b). The corresponding median pXC_{50} values broken down by target class are: ion channels (6.4, $N = 25$, mean molecular mass = 359 Da), enzymes (7.0, $N = 93$, mean molecular mass = 371 Da), GPCRs (7.9, $N = 106$, mean molecular mass =372 Da) and nuclear receptors (8.1, $N = 20$, mean molecular mass = 347 Da). The lower median potency of drugs that target ion channels could be due in part to their lower mean molecular mass, although drugs that target nuclear receptors have a similar mean molecular mass but display higher median potencies, suggesting other factors such as target type or mode of action may also be involved.

The data obtained on these 261 oral drugs with therapeutically relevant pXC_{50} values suggest that a more realistic pXC_{50} value to be sought during early research and development might be ~7.5. This has important implications, given that a pXC_{50} of 6 (corresponding to 1 μM potency) is a common cut-off used to select hits from an HTS screen, whereas a pXC_{50} of 9 (corresponding to 1 nM potency) is the desired output from a lead optimization programme. If a benchmark pXC_{50} of ~7.5, for example, were sought, it would be easier to justify selecting hits with low molecular mass and low micromolar potency from an HTS campaign, which might have a better chance of being optimized into drug candidates that possess both potency and ADMET characteristics required of an orally

administered drug molecule. Further analysis shows that ~20% of this dataset do not have any therapeutically relevant pXC_{50} values of ≥6, meaning they would be unlikely to pass typical lead-like potency filters42. Additionally, for 34% of oral drugs, the highest therapeutically relevant pXC_{50} value is <7, meaning their selection as candidate molecules might be challenging based on current approaches.

The absolute level of activity for a given compound can vary considerably between assays7, and so the pXC_{50} values reported per compound might also be affected by the underlying technologies used to generate them (for example, competitive-binding assays, enzyme inhibition assays, or cell-based functional assays for GPCRs and nuclear receptors). Nevertheless, the median pXC_{50} value observed here is almost identical to the median value of 7.8 reported for a set of oral and non-oral drugs1. However, the value reported here is considerably lower than that observed in another study of 60 recent drugs (median pXC_{50} = 8.5)33. This might be explained in part by the considerably higher molecular masses of this set of drugs (median of 436 Da versus 346 Da for the specific oral drugs set used here). This difference over time may have arisen owing to the targeting of different receptor types with different ligand requirements43. Alternatively, the difference in potency might be a reflection of the increased focus on *in vitro* potency as a parameter in lead optimization.

Relationship between *in vitro* potency and therapeutic dose

As highlighted above, the level of *in vitro* potency of a lead for its intended target(s) is a key parameter at the beginning of a drug discovery programme. This idea is extended further in lead generation–optimization based on the assumption that the most potent molecules *in vitro* are generally more likely to translate into more effective therapies, possessing a greater overall safety profile and requiring lower therapeutic doses. However, with regard to the first assumption, from this study and many others in the literature, it seems that the safety profile (ADMET and off-target-related effects) is more likely to be inversely correlated with the level of target potency given their opposing relationship with regard to key physicochemical properties. In an attempt to assess the validity of the second assumption — that high *in vitro* potency favourably affects the therapeutic dose — we have assessed the correlation between the two parameters for the set of 261 oral drugs above.

Assessing the relationship between *in vitro* potency and dosage amount is not trivial. This is because a drug molecule may be dosed at a range of possible concentrations and a range of time intervals that can depend on the nature of the illness, progression and so on. However, most molecules are generally formulated in a dose that directly reflects their efficacy, as there is a desire for a once-daily dosing interval. We therefore use the absolute dose size for the analysis. As we assessed the relationship between dose and potency on the log scale, doubling the dose, for example, will have a minor effect on the overall correlation.

An important limitation here is that we do not consider information regarding the complex pharmacokinetic profile for each drug; however, neither do researchers during early lead optimization, so it is still of general interest to understand how well both parameters correlate. It is also worth noting that the lack of published information (target identification and potencies) on oral drugs substantially hampers such analyses. Indeed, this information might not be known for drugs identified primarily through screening in animal models of

disease, or through serendipity. Nevertheless, while accepting these issues, the analysis of this compiled dataset represents an important first step in the assessment of the actual strength of the relationship between *in vitro* potency and therapeutic dose.

The correlation between the therapeutically relevant pXC_{50} values and the formulated dose for 261 oral drugs was assessed using linear regression (Fig 5c). In many cases, more than one target activity or more than one formulated dose are available, and so we assessed the correlation between the corresponding mean, maximum and minimum values of each for completeness. The best correlation was obtained between the mean formulated dose and the mean therapeutically relevant pXC_{50} reported, with a squared correlation coefficient (r^2) of 0.26. The r^2 values obtained can be roughly interpreted as meaning that the *in vitro* pXC_{50} data can explain ~30% of the variation in the dosage for the 261 compounds. This corresponds to a fivefold prediction error in the dose using the *in vitro* pXC_{50} alone.

The correlation between therapeutic dose and pXC_{50} when broken down by target class is variable: ion channels ($r^2 = 0.31$, $N = 25$), enzymes ($r^2 = 0.17$, $N = 93$), GPCRs ($r^2 = 0.24$, $N = 106$) and nuclear receptors ($r^2 = 0.45$, $N = 20$). This might suggest that higher potencies may be of more utility in achieving low-dose treatments for ion channels and nuclear receptors, although the number of observations are small and the correlation is still below 50% overall, meaning other important factors are involved.

It is not surprising that the correlation between *in vitro* potency and dose is relatively weak overall given the extensive role of ADME parameters in determining the concentration of drug that is available at the site of action to elicit a biological response. Indeed, considerable efforts are ongoing to improve our understanding of the link between pharmacokinetics–pharmacodynamics and receptor occupancy, particularly in the area of the central nervous system44–47.

Promiscuity of oral drugs

It is known that many drugs exhibit a therapeutic effect by acting on multiple pathways, thereby overcoming redundancy in the disease-gene network36–39,48–51. Indeed, Leeson *et al.*18 reported on the promiscuity of 2,133 drugs and reference compounds from the Cerep BioPrint database, noting that drugs in general hit large number of targets at >30% inhibition at 10 μM. These results seem to contradict to some extent the most common strategy in drug discovery of searching for highly potent, selective molecules for a single target.

We assessed the promiscuity of oral drugs further using our dataset of 392 oral drugs, which have 2,864 measured pXC_{50} values in total. The biological coverage of the dataset is less good than that covered by Cerep, as discussed before. However, the use of quantitative pXC_{50} values rather than percentage inhibition values, as well as a focus on oral drugs alone, makes the analysis relevant.

Fig 6 shows that 29% of the 392 oral drugs studied did not have any reported hits with potency ≤1 μM in the database. This may be because they are either only weakly active molecules or their activities have not been reported in the literature. Eighteen percent of the oral drugs display only 1 hit at or below the 1-μM threshold, 22% showed between 2–5 hits,

14% showed between 5–10 hits, and 16% showed >10 hits. Although we cannot directly relate the number of hits to their efficacy, these findings and the reports of others seem to suggest that the therapeutic effect of drug molecules is more complex than the popular 'one disease, one target' paradigm.

It should be noted that this analysis will underestimate the promiscuity of oral drugs, as the 392 oral drugs have only ~7.3 unique target activity measurements on average. This also helps to explain why this set seems less promiscuous than those in the network pharmacology studies of others48–51. We would therefore expect that, as additional targets are assayed, the probability of additional off-target hits will increase.

Conclusions

The pharmaceutical industry is increasingly focusing on new targets, the druggability of which is often unknown34, and this is one of the principal reasons for the exploration of novel areas of physicochemical property space. However, it is increasingly clear from many publications in the literature that the focus on less well explored chemical space is neither likely to increase the probability of bringing new, safe drugs to market, nor reduce overall attrition rates. A key question is whether it is possible to combine the characteristics needed for oral administration with those needed for a drug to show efficacy at novel targets. In this respect, the fragment-based approach52 is an attractive option, provided the leads identified are not subsequently incorporated into an optimization cascade that seeks the highest target potencies if there are no data to support such a strategy. The quest for high target potency should not be pursued blindly without an understanding of its relevance to efficacy. Obtaining high potency at a single target when a network of targets are important, or at the cost of favourable ADMET and physicochemical properties, is futile.

Although molecules that bind effectively to the target(s) implicated in a particular medical condition are required, we need to carefully consider what level of *in vitro* potency and selectivity is needed to achieve efficacy, especially given that the historical oral drugs analysed here show median pXC_{50} values of ~7.7 (~50 nM), and what might be considered a less than 'clean' selectivity profile, even in this limited dataset. On the basis of this analysis, we conclude that the balance of parameters that are being sought in pharmaceutical candidate molecules are often suboptimal, and we postulate that this discrepancy is in part an artefact of the screening cascade6, which puts great reliance on simplistic *in vitro* assays of drug targets53 early in the selection–optimization process. These assays are certainly not infallible given that only ~10% of what are deemed sufficiently potent candidate molecules make it to the market as effective drugs3,4, with 30% failing for reasons of efficacy alone54.

In summary, we think a drug discovery process with improved focus on quantifying the therapeutic potential of the molecules through a more complete consideration of the necessary biological parameters will be more successful. The definition of a more sophisticated hypothesis-driven approach at the beginning of a drug discovery programme (for example, considering receptor occupancy45–47, receptor off-rates53 and pharmacokinetics–pharmacodynamics44), based on a more thorough knowledge of the

biology of the disease and the molecular properties required for oral drugs, is crucial to minimize compound attrition on the path to the market.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

M.P.G., A.H. and D.M. are grateful to many former colleagues at GlaxoSmithKline for insightful discussion on this topic. M.P.G. gratefully acknowledges support from the Faculty of Science at Kasetsart University as well as additional assistance from S. Hannongbua and S. Ruchirawat. The authors thank J. Proudfoot (Boehringer Ingelheim Pharmaceuticals) for kindly providing a copy of his oral-drug dataset. ChEMBL data collection, database maintenance and support are funded through a Wellcome Trust award.

References

1. Overington JP, Al-Lazikani B, Hopkins AL. How many drug targets are there? Nature Rev Drug Discov. 2006; 5:993–996. [PubMed: 17139284]
2. Li D, Kerns EH. Application of pharmaceutical profiling assays for optimization of druglike properties. Curr Opin Drug Discov Devel. 2005; 8:495–504.
3. Peck RW. Driving earlier clinical attrition: if you want to find the needle, burn down the haystack. Considerations for biomarker development. Drug Discov Today. 2006; 12:289–294.
4. Paul SM, et al. How to improve R&D productivity: the pharmaceutical industry's grand challenge. Nature Rev Drug Discov. 2010; 9:203–214. [PubMed: 20168317]
5. Kalgutkar AS, et al. A comprehensive listing of bioactivation pathways of organic functional groups. Curr Drug Metabol. 2005; 6:161–225.
6. Keseru GM, Makara GM. The influence of lead discovery strategies on the properties of drug candidates. Nature Rev Drug Discov. 2009; 8:203–212. [PubMed: 19247303]
7. Lackey K. Lessons from the drug discovery of lapatinib, a dual ErbB1/2 tyrosine kinase inhibitor. Curr Topics Med Chem. 2006; 6:435–460.
8. Lipinski CA, Lombardo F, Dominy BW, Feeney PJ. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. Adv Drug Deliv Rev. 1997; 23:3–25. [**This paper highlighted for the first time the link between drug-likeness and key physicochemical properties (that is, the rule of 5).**]
9. Teague SJ, Davis AM, Leeson PD, Oprea T. The design of leadlike combinatorial libraries. Angew Chem Int Ed. 1999; 38:3743–3748.
10. Hann MM, Leach AR, Harper G. Molecular complexity and its impact on the probability of finding leads for drug discovery. J Chem Inf Comput Sci. 1999; 41:856–864.
11. Leeson PD, Davis AM, Steele J. Drug-like properties: guiding principles for design — or chemical prejudice? Drug Discov Today. 2004; 1:189–195.
12. Lajiness MS, Vieth M, Erickson J. Molecular properties that influence oral drug-like behaviour. Curr Opin Drug Disc Devel. 2004; 7:470–477.
13. Hann MM, Oprea TI. Pursuing the leadlikeness concept in pharmaceutical research. Curr Opin Chem Biol. 2004; 8:255–263. [PubMed: 15183323]
14. Leeson PD, Davis AM. Time-related differences in the physical property profiles of oral drugs. J Med Chem. 2004; 47:6338–6348. [PubMed: 15566303]
15. Li D, Kerns EH. Biological assay challenges from compound solubility: strategies for bioassay optimization. Drug Discov Today. 2006; 11:446–451. [PubMed: 16635808]
16. Wunberg T, et al. Improving the hit-to-lead process: data-driven assessment of drug-like and lead-like screening hits. Drug Discov Today. 2006; 11:175–180. [PubMed: 16533716]
17. De Witte RS. Avoiding physicochemical artefacts in early ADME–Tox experiments. Drug Discov Today. 2006; 11:855–859. [PubMed: 16935755]

18. Leeson P, Springthorpe B. The influence of drug-like concepts on decision-making in medicinal chemistry. Nature Rev Drug Discov. 2007; 6:881–890. [PubMed: 17971784] [**An excellent paper that describes, with well-chosen examples, the importance of physicochemical properties in medicinal chemistry research.**]
19. Proudfoot J. The evolution of synthetic oral drug properties. Bioorg Med Chem Lett. 2005; 15:1087–1090. [PubMed: 15686918]
20. Johnson TJ, Dress KR, Edwards M. Using the Golden Triangle to optimize clearance and oral absorption. Bioorg Med Chem Lett. 2009; 19:5560–5564. [PubMed: 19720530]
21. Waring MJ. Defining optimum lipophilicity and molecular weight ranges for drug candidates — molecular weight dependent lower logD limits based on permeability. Bioorg Med Chem Lett. 2009; 19:2844–2851. [PubMed: 19361989]
22. Hopkins AL, Groom CR, Alex A. Ligand efficiency: a useful metric for lead selection. Drug Discov Today. 2004; 9:430–431. [PubMed: 15109945]
23. Gleeson MP. Generation of a set of simple, interpretable ADMET rules of thumb. J Med Chem. 2008; 51:817–834. [PubMed: 18232648] [**An interesting paper that assesses the link between molecular mass, logP and ionization state for a range of ADMET parameters that are routinely measured in industry.**]
24. Sneader, W. Drug Prototypes and their Exploitation. Wiley; Chichester: 1996.
25. Oprea TI, Davis AM, Teague SJ, Leeson PD. Is there a difference between leads and drugs? A historical perspective. J Chem Inf Comput Sci. 2001; 41:1308–1315. [PubMed: 11604031]
26. Hadjuk PJ. Fragment-based drug design: how big is too big? J Med Chem. 2006; 49:6972–6976. [PubMed: 17125250] [**This paper highlighted the benefits of selecting the most ligand-efficient molecular templates in lead generation.**]
27. Wenlock MC, Austin RP, Barton P, Davis AM, Leeson PD. A comparison of physiochemical property profiles of development and marketed oral drugs. J Med Chem. 2003; 46:1250–1256. [PubMed: 12646035] [**This study showed that, as compounds in different phases of development get closer to the market, their mean molecular mass and logP tend to converge towards those of marketed drugs.**]
28. Tyrchana C, Blomberga N, Engkvista O, Kogeja T, Muresan S. Physicochemical property profiles of marketed drugs, clinical candidates and bioactive compounds. Bioorg Med Chem Lett. 2009; 19:6943–6947. [PubMed: 19879759]
29. Oprea TI, et al. Lead-like, drug-like or "pub-like": how different are they? J Comput Aided Mol Des. 2007; 21:113–119. [PubMed: 17333482]
30. Andrews PR, Craik DJ, Martin JL. Functional group contributions to drug-receptor interactions. J Med Chem. 1984; 27:1648–1657. [PubMed: 6094812]
31. Kuntz ID, Chen K, Sharp KA, Kollman PA. The maximal affinity of ligands. Proc Natl Acad Sci USA. 1999; 96:9997–10002. [PubMed: 10468550]
32. Abad-Zapatero C, Metz JT. Ligand efficiency indices as guideposts for drug discovery. Drug Discov Today. 2005; 10:464–469. [PubMed: 15809192]
33. Perola E. An analysis of the binding efficiencies of drugs and their leads in successful drug discovery programs. J Med Chem. 2010; 53:2986–2997. [PubMed: 20235539]
34. Hadjuk PJ, Huth JR, Tse C. Predicting protein druggability. Drug Discov Today. 2005; 10:1675–1682. [PubMed: 16376828]
35. Vieth M, Sutherland JJ. Dependence of molecular properties on proteomic family for marketed oral drugs. J Med Chem. 2009; 49:3451–3453.
36. Goh K, et al. The human disease network. Proc Natl Acad Sci USA. 2007; 104:8685–6690. [PubMed: 17502601]
37. Zimmermann GR, Lehár J, Keith CT. Multi-target therapeutics: when the whole is greater than the sum of the parts. Drug Discov Today. 2007; 12:34–42. [PubMed: 17198971]
38. Hopkins AL, Mason JS, Overington JP. Can we rationally design promiscuous drugs? Curr Opin Struct Biol. 2006; 16:127–136. [PubMed: 16442279]
39. Morphy R, Rankovic Z. Fragments, network biology and designing multiple ligands. Drug Discovery Today. 2007; 12:156–160. [PubMed: 17275736]

40. Azzaoui K, Hamon J, Faller B, Whitebread S, Jacoby E, Bender A, Jenkins JL, Urban L. Modeling Promiscuity Based on in vitro Safety Pharmacology Profiling Data. Chem Med Chem. 2007; 2:874–880. [PubMed: 17492703]
41. Peters J-U, Schnider P, Mattei P, Kansy M. Pharmacological Promiscuity: Dependence on Compound Properties and Target Specificity in a Set of Recent Roche Compounds. Chem Med Chem. 2009; 4:680–186. [PubMed: 19266525]
42. Davis AM, Keeling DJ, Steele J, Tomkinson NP, Tinker AC. Components of Successful Lead Generation. Curr Topics Med Chem. 2005; 5:421–439.
43. Morphy R. The Influence of Target Family and Functional Activity on the Physicochemical Properties of Pre-Clinical Compounds. J Med Chem. 2006; 49:2969–2978. [PubMed: 16686538]
44. McGinnity DF, Collington J, Austin RP, Riley RJ. Evaluation of Human Pharmacokinetics, Therapeutic Dose and Exposure Predictions Using Marketed Oral Drugs. Curr Drug Metabolism. 2007; 8:463–479.
45. Jeffrey P, Summerfield S. Assessment of the blood–brain barrier in CNS drug discovery. Neurobiology Disease. 2010; 37:33–37.
46. Summerfield SG, Stevens AJ, Cutler L, Carmen-Osuna MD, Hammond B, Tang S, Hersey A, Spalding DJ, Jeffrey P. Improving the in Vitro Prediction of in Vivo Central Nervous System Penetration: Integrating Permeability, P-glycoprotein Efflux, and Free Fractions in Blood and Brain. JPET. 2006; 316:1282–1290.
47. Watson J, Coggon S, Lucas A, Clarke KL, Viggers J, Cheetham S, Jeffrey P, Porter R, Read KD. Receptor Occupancy and Brain Free Fraction. Drug Metab Dispos. 2009; 37:753–760. [PubMed: 19158315]
48. Hopkins AL. Network pharmacology. Nat Biotechnol. 2007; 11:1110–1111.
49. Yildirim MA, Goh KI, Cusick ME, Barabasi AL, Vidal M. Drug-target network. Nat Biotechnol. 2007; 25:1119–1126. [PubMed: 17921997]
50. Hopkins AL. Network pharmacology: the next paradigm in drug discovery. Nature Rev Chem Biol. 2008; 4:682–690.
51. Janga SC, Tzakos A. Structure and organization of drug-target networks: insights from genomic approaches for drug discovery. Mol Biosyst. 2009; 5:1536–1548. [PubMed: 19763339]
52. Congreve M. A 'Rule of Three' for fragment-based lead discovery? Drug Discovery Today. 2003; 8:876–877.
53. Copeland RA, Pompliano DL, Meek TD. Drug–target residence time and its implications for lead optimization. Nature Rev Drug Disc. 2006; 5:730–739. [**This paper discusses issues associated with current biochemical screening technologies, and advocates the assessment of receptor off-rates to facilitate the optimization of compound efficacy.**]
54. Kola I, Landis J. Can the pharmaceutical industry reduce attrition rates? Nature Rev Drug Discovery. 2004; 3:711–715. [PubMed: 15286737]
55. Ekins S, Williams AJ. Reaching Out to Collaborators: Crowdsourcing for Pharmaceutical Research. Pharm Res. 2010; 27:393–395. [PubMed: 20107873]
56. Young D, Martin T, Venkatapathy R, Harten P. Are the Chemical Structures in Your QSAR Correct? QSAR Comb Sci. 2008; 27:1337–1345.
57. Fourches D, Muratov E, Tropsha A. Trust, But Verify: On the Importance of Chemical Structure Cheminformatics and QSAR Modeling Research. J Chem Inf Model. 2010; 50:1189–1204. [PubMed: 20572635]
58. Daugan A, Grondin P, Ruault C, Le Monnier de Gouville A, Coste H, Kirilovsky J, Hyafil F, Labaudiniere R. The Discovery of Tadalafil: A Novel and Highly Selective PDE5 Inhibitor. 1:5,6,11,11a-Tetrahydro-1H-imidazo[1'5':1,6]pyrido[3,4-b]indole-1,3(2H)-dione Analogues. J Med Chem. 2003; 46:4525–4532. [PubMed: 14521414]
59. Moriguchi I, Hirono S, Liu Q, Nakagome I, Matsushita Y. Simple method of calculating octanol/water partition coefficient. Chem Pharm Bull. 1992; 40:127–130.

Box 1

Data and Methods

Literature-derived data

The data used in this analysis were extracted from the publicly accessible ChEMBL database (see Further information for a web link and Supplementary information S4 (table) and Supplementary information S5 (table) for the specific datasets analysed) in October 2009. Publicly available databases such as ChEMBL are becoming an increasingly important source of information in pharmaceutical research55. The key advantage of such data is that analyses can be done in an open manner, and thus can be critically assessed by others. This contrasts to many informatics analyses reported on proprietary datasets of large pharmaceutical companies. Nevertheless, limitations of publicly available datasets have been highlighted recently, with data translation errors ranging from 0.1 to 3.4% depending on the database in question56. In addition, such errors can complicate the generation of quantitative structure–activity relationship models on specific end points57. Even so, valuable lessons can be learned from the analysis of such databases8,18, and to mitigate against such errors in this analysis we focus on large populations of samples and assess mean differences for rather broad physicochemical groups (that is, qualitative structure–activity relationships).

The dataset of 201,355 compounds was extracted from ChEMBL07. In order to simplify the analysis, the data were filtered to include only XC_{50} values from assays in which the target was a protein and in which the data were recorded in nM or μM units. XC_{50} values reported as 'greater than' or 'less than' were also excluded, as were very high (>12) or low (<0) values of the negative logarithm of XC_{50} (pXC_{50}). The XC_{50} values for compounds for which only inhibition constant (Ki) values were reported were obtained by conversion assuming competitive inhibition ($Ki = IC_{50}/2$). In some cases, the target information in the original paper did not record the specific isoform of the receptor, and so the data are associated with multiple targets and flagged as such. For example, IC_{50} values for sildenafil against the targets phosphodiesterase 1 (PDE1)–PDE5 are recorded, but the specific PDE isoform (A1, A2 and so on) is not known and so has been recorded as against all isoforms58. For our promiscuity analysis, this would suggest the compound was non-selective and so data for which the specific isoform is not known were excluded from the analysis. Having extracted the data, XC_{50} values recorded against the same target and compound were averaged.

Oral drugs with dose information were identified from the ChEMBL database ($N = 792$). Drugs were extracted only where the dosing vehicle corresponded to a capsule, tablet or suspension. All doses were converted to the logarithmic (molar) scale and the mean, maximum and minimum values were computed for further analysis. Therapeutic target information was extracted for a subset of the 792 oral drugs for which we could find target pXC_{50} values in ChEMBL ($N = 392$). The target information was then used to match the therapeutic doses to relevant therapeutic target pXC_{50} values. This resulted in a dataset of 261 drugs with both dose and therapeutically relevant pXC_{50} values. Target information was taken from: ChEMBL, Drugdex, Martindales, Physicians' Desk

Reference, Index Nominum, Pubmed, RXList, Drugbank and *The Journal of Pharmacology and Experimental Therapeutics* (see Further information for web links). For our assessment of the promiscuity of oral drugs, we determined the total number of unique target hits ≥1 μM for all 392 oral drugs with reported potency data in ChEMBL.

Statistical analyses

We used the analysis of variance (ANOVA) technique to assess the majority of the relationships in the current analysis, owing to their simplicity and ease of interpretation. ANOVA is particularly suitable for assessing weak relationships (whereas regression is preferable when a relationship is considered moderate to strong). For each analysis, molecular mass and logP (the logarithm of the octanol–water partition coefficient) were binned into 4–6 separate categories for the analysis, with bins chosen to ensure an even distribution so that the mean values could be assessed reliably.

Results obtained from a given analysis have been displayed only if the relationships observed are statistically significant above the commonly used 95% confidence level based on either regression or Student's *t* tests or ANOVA testing. ANOVA (one-way and two-way) calculations were performed in Statistica 7 and regression analyses in JMP 6.0 using the default settings. Molecular mass, AlogP59 and the negative logarithm of the acid dissociation constant (p*K*a) were calculated for each molecule in PipelinePilot 8.0. Compounds that reported an error in PipelinePilot 8.0 were excluded from the analysis. In most cases, we report the mean values for the distributions. Median values are shown for the dose–potency correlations of oral drugs because the distributions are non-normal18.

Europe PMC Funders Author Manuscripts

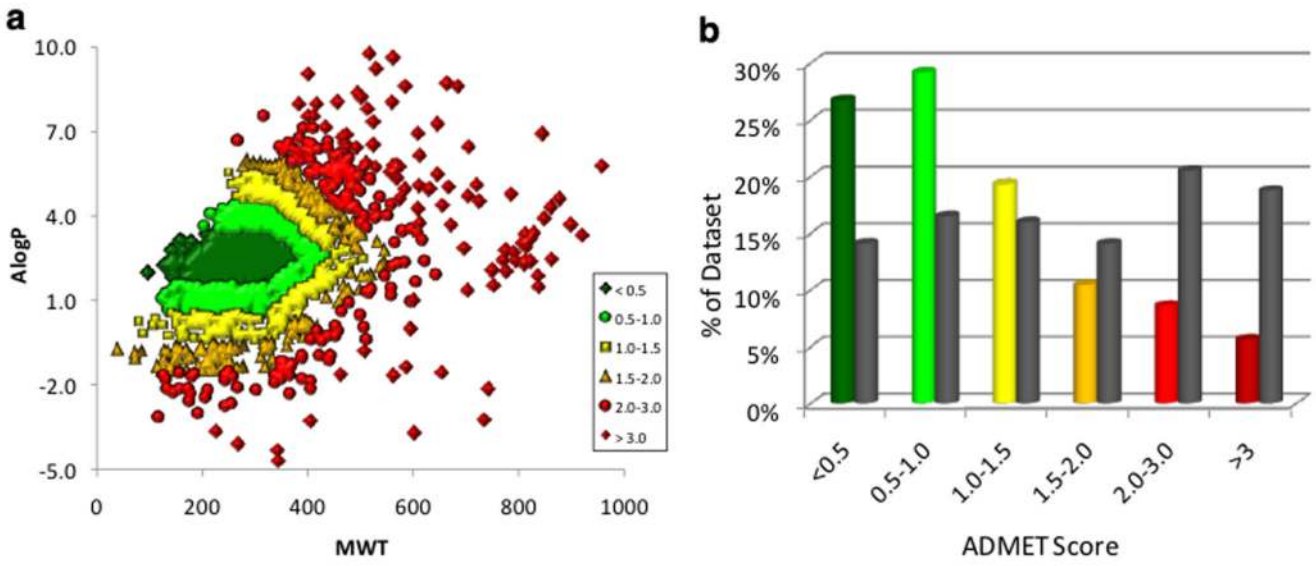

Figure 1.

Relationship between ADMET parameters and physicochemical properties. (a) Plot of molecular mass versus AlogP for 1,791 oral drugs. The data are coloured according to the absorption, distribution, metabolism, excretion and toxicity (ADMET) score, which is a measure of the deviation from oral drug space as given by molecular mass and AlogP. (b) A comparison of the ADMET score distribution of oral drugs (coloured according to the score) and 201,355 unique compounds with measured target potency data from the ChEMBL database (coloured in black). Approximately 14% of oral drugs have a score >2 compared to ~39% for drug discovery compounds reported in ChEMBL.

Europe PMC Funders Author Manuscripts

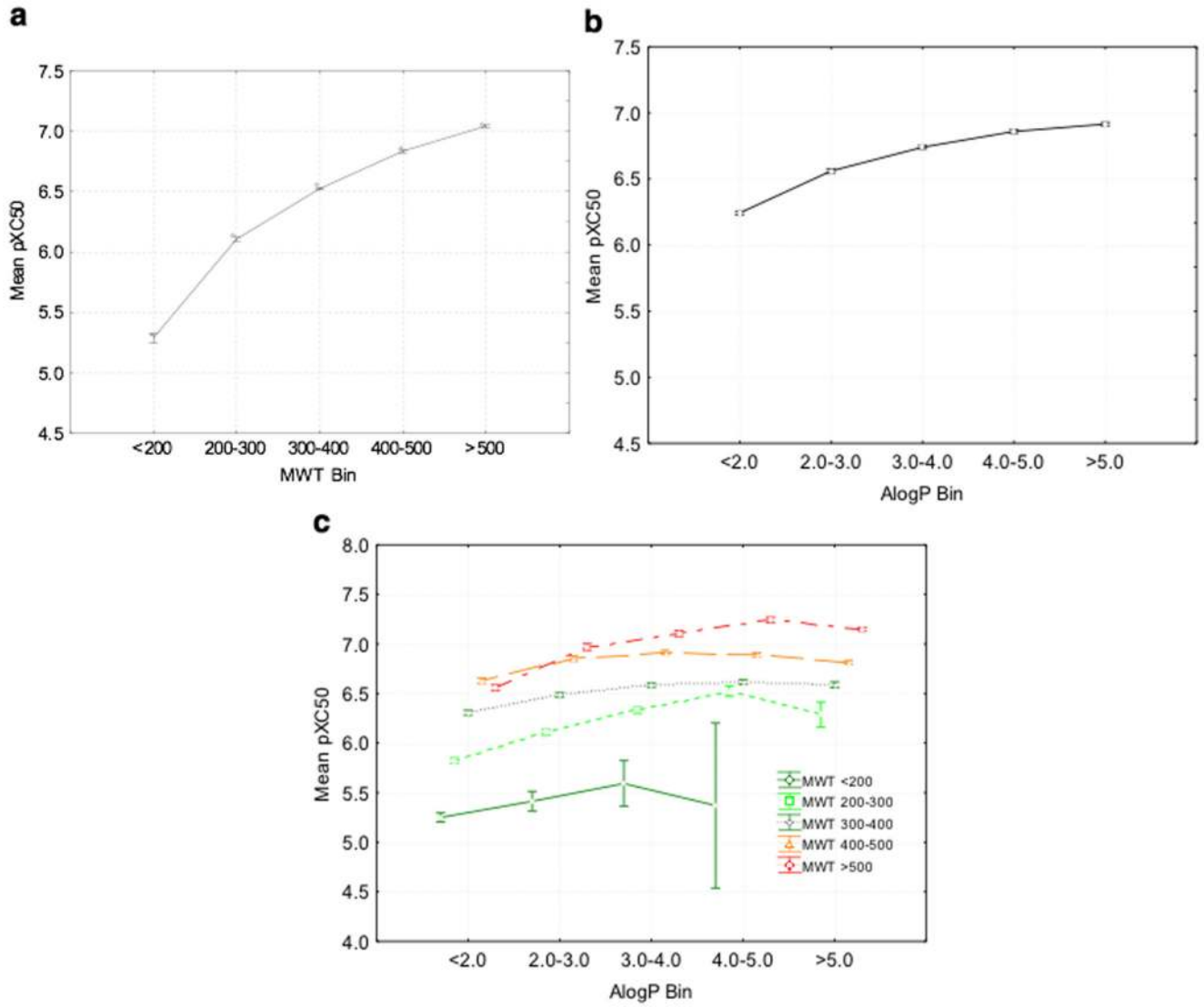

Figure 2.

Relationship between in vitro potency and physicochemical properties. The graphs illustrate the relationship between the mean reported *in vitro* potency (shown here as the negative logarithm of the XC_{50} value; pXC_{50}) and molecular mass (**a**) or AlogP (**b**) for 201,355 compounds with reported activity measurements in the ChEMBL database. **c** | The effect of both parameters shows some independence as, for a given AlogP value, a concomitant increase in molecular mass increases the mean pXC_{50} value and vice versa. Only 9 observations are present in the AlogP 4.0–5.0, molecular mass <200 category, hence the large error bars. Plotted values are offset within x-axis bins to aid visualization. Error bars denote the 95% confidence interval in the means

Europe PMC Funders Author Manuscripts

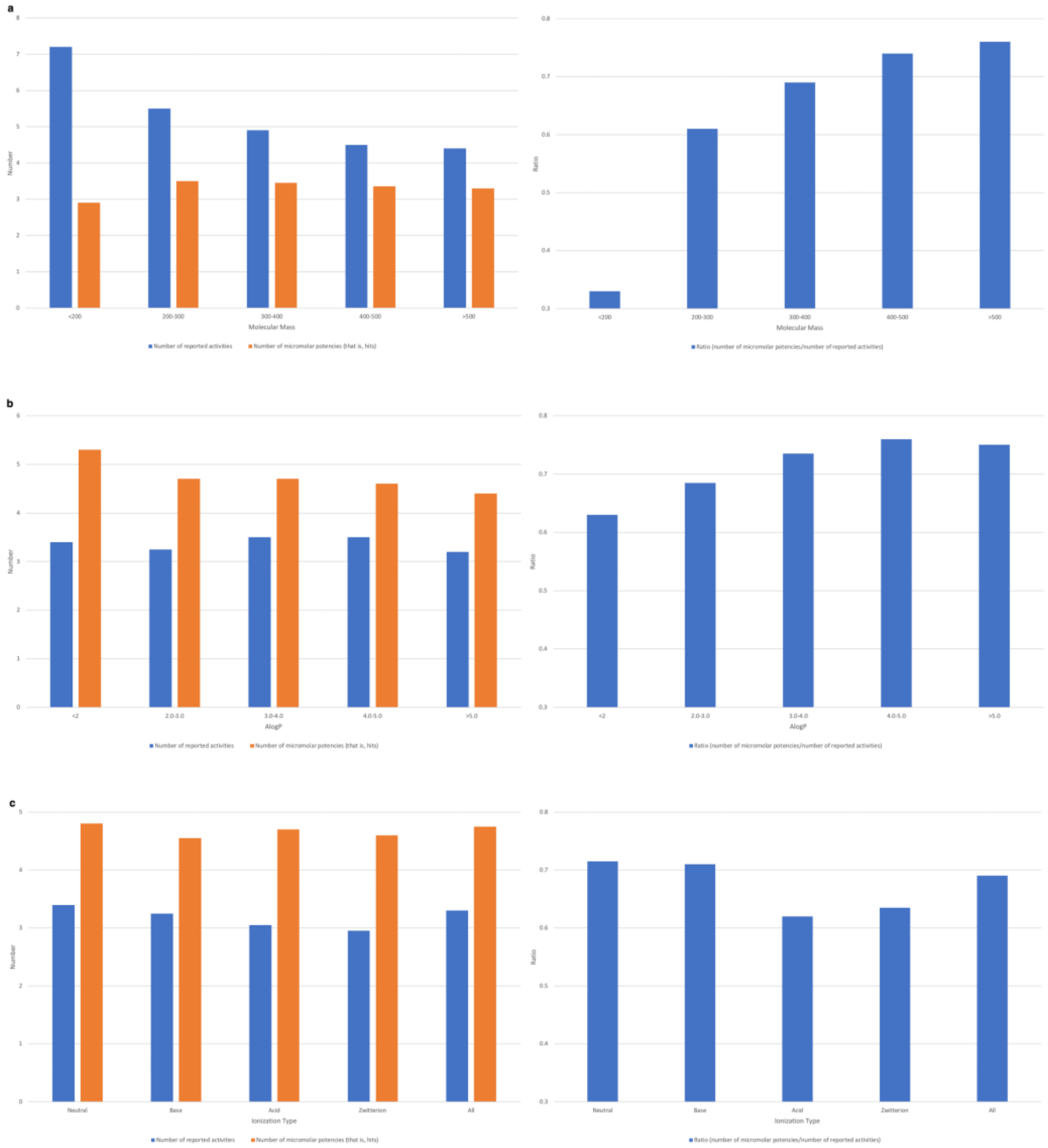

Figure 3.

Relationship between promiscuity and physicochemical properties.

The graphs illustrate the relationship between the mean promiscuity and molecular mass (**a**), AlogP (**b**) or ionization state (**c**) for a diverse set of 40,408 molecules in the ChEMBL database. Each compound has ≥3 activity measurements reported in ChEMBL. The total

number of observations in each molecular mass bin, in order of increasing mass are: 924, 5,121, 12,287, 12,514 and 9,562. The corresponding values for AlogP are: 8,741, 6,839, 9,033, 7,632 and 8,163, respectively. The corresponding values for ionization state are: 22,060, 10,893, 5,862 and 1,593, respectively.

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

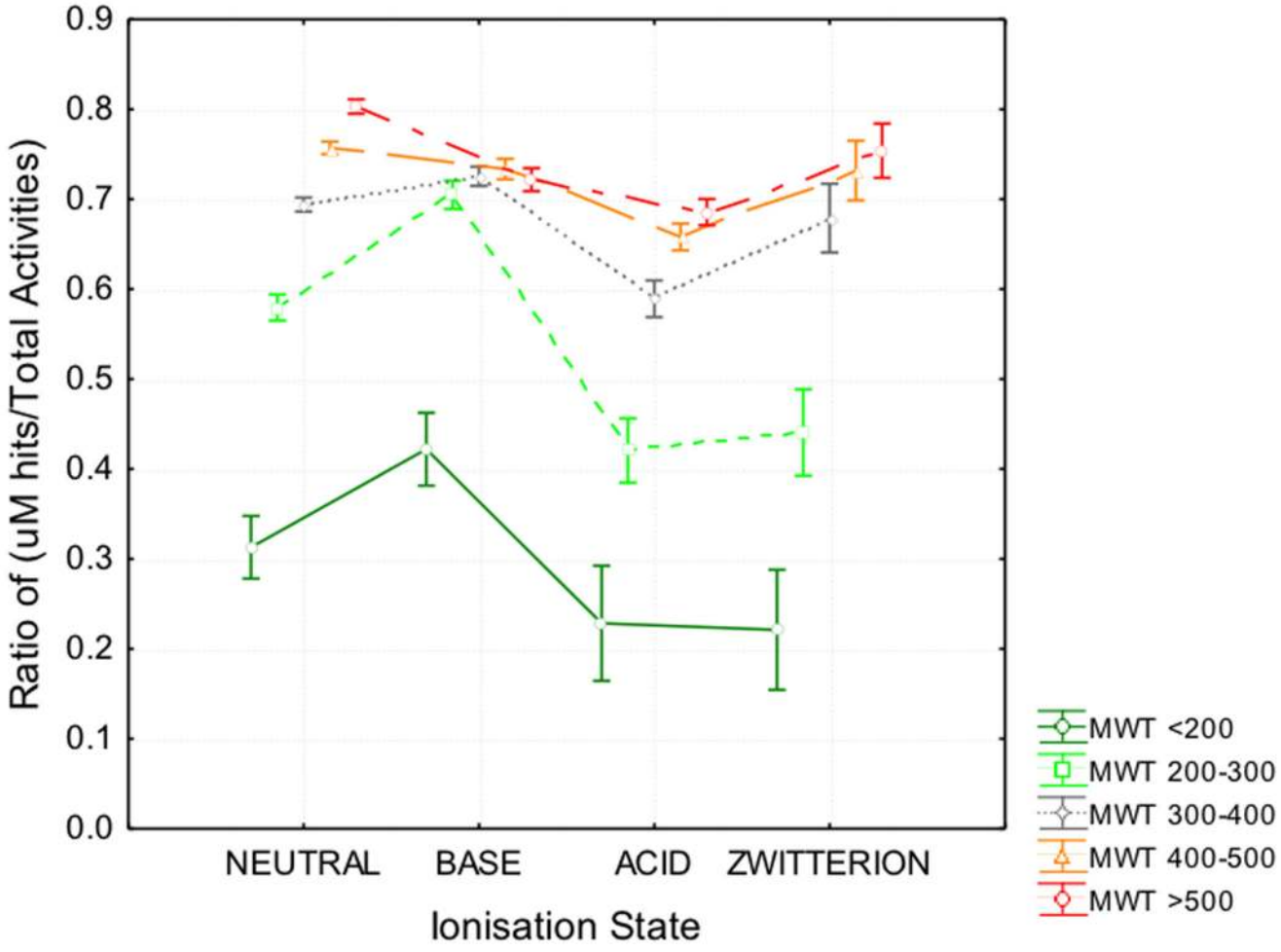

Figure 4.

Relationship between promiscuity and ionization state.

The data shown are for a diverse set of 40,408 molecules and are broken down by molecular mass to aid analysis. Error bars denote the 95% confidence interval in the means. Plotted values are offset within x-axis categories to aid visualization.

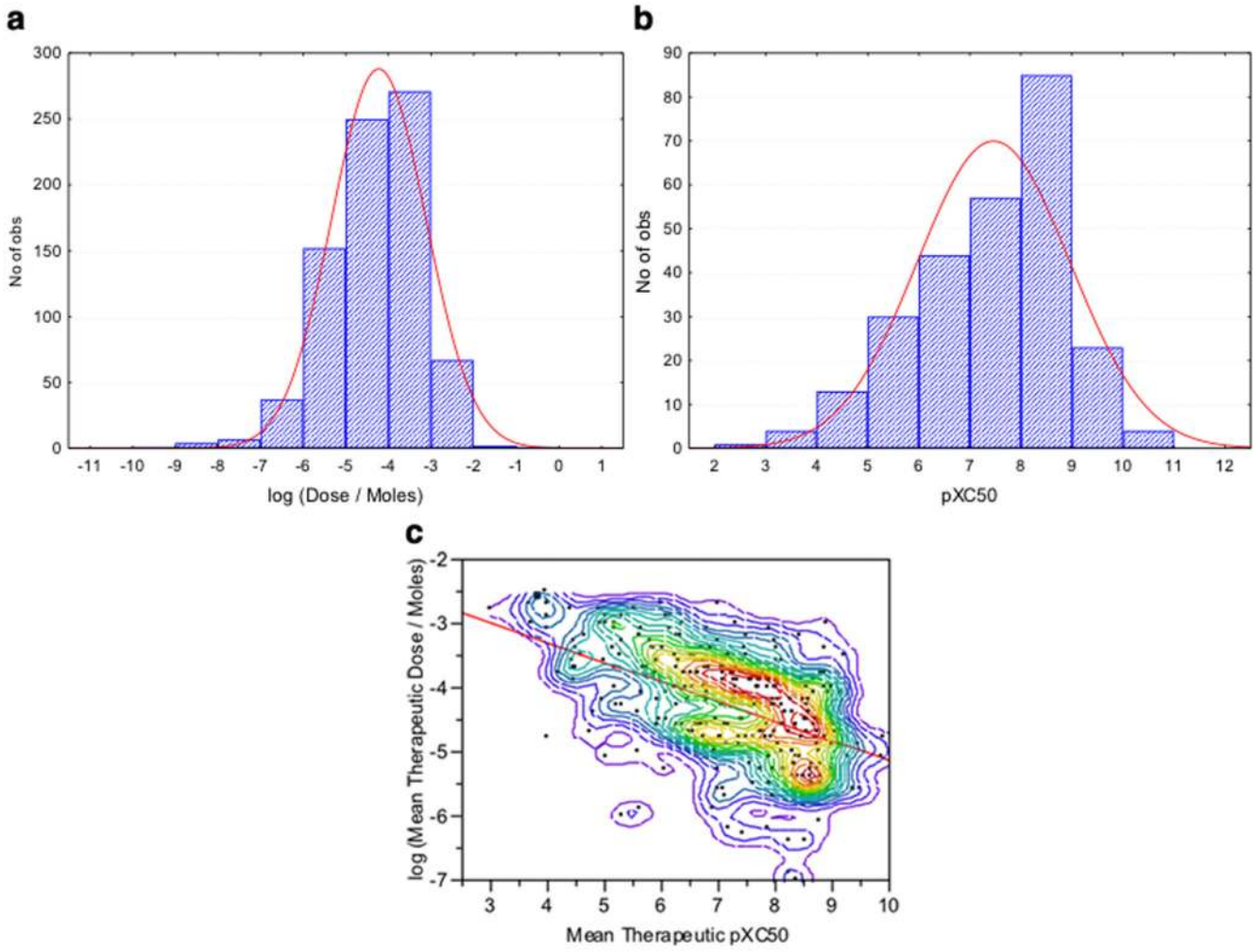

Figure 5.

Relationship between in vitro potency and dose for oral drugs.

Distribution of the lowest reported oral dose for 792 oral drugs (**a**) and the maximum reported *in vitro* pXC_{50} for a subset of 261 drugs for which the target(s) are known and relevant potency data are available. The distribution of the minimum, mean and maximum doses per compound have median values corresponding to 63 μmol, 125 μmol and 223 μmol, or 24 mg, 47 mg and 83 mg, respectively. The corresponding median values for the minimum, mean and maximum therapeutically relevant pXC_{50} values are 7.0, 7.3 and 7.7, respectively. **c** | Relationship between the reported mean therapeutic dose and the mean therapeutically relevant pXC_{50} for the set of 261 oral drugs. The relationship between therapeutic dose and pXC_{50} is displayed using a regression plot rather than analysis of variance (ANOVA), as it is a comparatively strong relationship. The correlation between therapeutic dose and molecular mass has a correlation coefficient of $r^2 = 0.10$, whereas the corresponding value between molecular mass and pXC_{50} is $r^2 = 0.16$.

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

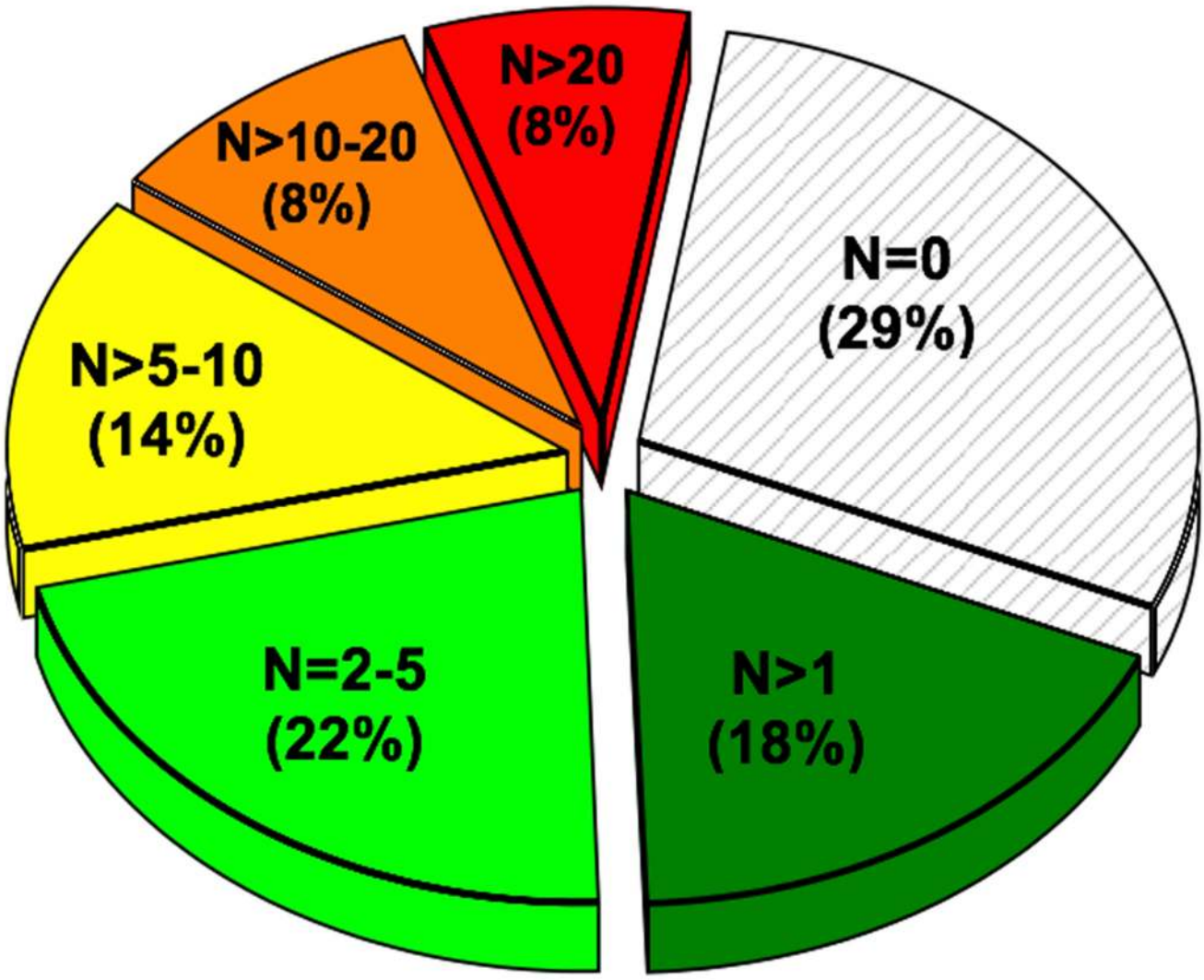

Figure 6.

Promiscuity of oral drugs.

The number of hits ≤ μM reported for a subset of 392 oral drugs extracted from ChEMBL. The percentage of compounds with the number of reported hits indicated are shown next to each portion of the pie chart. This is likely to underestimate the promiscuity of oral drugs given the scarcity of biological data per compound in ChEMBL.

Europe PMC Funders Author Manuscripts

Table 1

Distribution of physicochemical properties for the ChEMBL dataset and oral drugs ADMET: absorption, distribution, metabolism, excretion and toxicity, SD: standard deviation

Property bin	Oral drugs (N = 1,791)	ChEMBL (all) (N = 201,355)	ChEMBL (≤1 nM) (N = 15,934)
Molecular mass >400 and AlogP >4	8%	30%	41%
Molecular mass >500 and AlogP >5	2%	9%	16%
>2 rule of five failures	12%	14%	21%
Molecular mass mean/median (SD)	333/316 (121)	430/418 (131)	490/475 (132)
Molecular mass ≤300	42%	14%	6%
Molecular mass ≤400	79%	44%	26%
Molecular mass ≤500	93%	75%	58%
AlogP mean/median (SD)	2.5/2.6 (2.0)	3.5/3.6 (2.1)	4.0/3.9 (2.0)
AlogP <3	58%	36%	27%
AlogP <4	77%	58%	50%
AlogP <5	91%	77%	73%
ADMET score mean/median (SD)	1.2/0.9 (1.0)	1.9/1.6 (1.4)	2.4/2.1 (1.5)