Published in final edited form as:
Nat Chem. ; 4(2): 90–98. doi:10.1038/nchem.1243.

Quantifying the chemical beauty of drugs

G. Richard Bickerton[1], **Gaia V. Paolini**[2], **Jérémy Besnard**[1], **Sorel Muresan**[3], and **Andrew L. Hopkins**[1]

[1]Division of Biological Chemistry and Drug Discovery, College of Life Sciences, University of Dundee, Dundee, DD1 5EH, UK

[2]Gaia Paolini Ltd, 29 High Street, Bridge, Canterbury, CT4 5JZ, UK

[3]DECS Computational Compound Sciences, Computational Chemistry, AstraZeneca R&D Mölndal, S-431 83 Mölndal, Sweden

Abstract

Druglikeness is a key consideration when selecting compounds during the early stages of drug discovery. However, evaluation of druglikeness in absolute terms does not adequately reflect the whole spectrum of compound quality. More worryingly, widely used rules may inadvertently foster undesirable molecular property inflation as they permit the encroachment of rule-compliant compounds toward their boundaries. We propose a measure of druglikeness based on the concept of desirability called Quantitative Estimate of Druglikeness (QED). The empirical rationale of QED reflects the underlying distribution of molecular properties. QED is intuitive, transparent, straightforward to implement in many practical settings and allows compounds to be ranked by their relative merit. We extend the utility of QED by applying it to the problem of molecular target druggability assessment by prioritizing a large set of published bioactive compounds. The measure may also capture the abstract notion of aesthetics in medicinal chemistry.

The concept of druglikeness provides useful guidelines for early stage drug discovery [1, 2]. Analysis of the observed distribution of some key physicochemical properties of approved drugs, including molecular weight, hydrophobicity and polarity, reveals they preferentially occupy a relatively narrow range of possible values[3]. Compounds that fall within this range are described as “druglike.” Note that this definition holds in the absence of any obvious structural similarity to an approved drug. It has been shown that preferential selection of druglike compounds increases the likelihood of surviving the well-documented high rates of attrition in drug discovery[4].

Druglikeness can be rationalized by consideration of how simple physicochemical properties impact molecular behavior *in vivo*, with particular respect to solubility, permeability, metabolic stability and transporter effects. Indeed druglikeness is often used as a proxy for

Correspondence should be addressed to A.L.H. (a.hopkins@dundee.ac.uk)..

Additional Information

Supplementary information is available online at XXXX. We have implemented QED as simple functions in Python, SQL (Structure Query Language), Accelrys Pipeline Pilot and Microsoft Excel, the codes for which are available in the Supplementary Information. The Microsoft Excel example also includes data on the 771 oral drugs used to derive the desirability functions. Pre-calculated QED values and desirability functions for 657,736 compounds from ChEMBL (release ChEMBL09) are also available.

Author contributions

A.L.H. conceived the approach and designed the algorithm. G.R.B. implemented the algorithm, performed the calculations, identified the functions and performed the analysis. G.V.P. developed the use of Shannon entropy as the weighting scheme. J.B. wrote the Pipeline Pilot implementation. S.M. coordinated the survey of AstraZeneca chemists. A.L.H and G.R.B. wrote the manuscript. G.R.B. produced the figures. All authors commented on the manuscript.

oral bioavailability. However, druglikeness provides a broad composite descriptor that implicitly captures several criteria, with bioavailability amongst the most prominent.

In practical terms, assessment of druglikeness is most commonly manifested as rules, the original and most well known of which is Lipinski's Rule of Five (Ro5)[5]. The rule states that a compound is more likely to exhibit poor absorption or permeation when two or more of the following physicochemical criteria are fulfilled: the molecular weight (MW) is greater than 500Da; the calculated logP (ClogP) is greater than 5; there are more than 5 hydrogen-bond donors or the number of hydrogen-bond acceptors (nitrogen and oxygen atoms) is greater than 10. The rule does not apply to substrates of biological transporters or natural products. Aside from its predictive power, the widespread adoption of the Ro5 as a guideline for compound evaluation can also be attributed to the fact that it is conceptually simple and straightforward to implement.

Lipinski's insight - that the great majority of orally absorbed drugs occupy a privileged area of molecular property space[5, 6] - has resulted in greater awareness of the importance of molecular properties in determining oral bioavailability. The rule has inspired numerous refinements and investigations into the concept of druglikeness: a comprehensive review of the area is provided by Ursu *et al.* [2]. The rule of five is not without its critics[7], yet in detail the issues tend to be with its qualitative nature, or the focus on oral drug space, as opposed to druglike thinking *per se*.

Paradoxically, since the publication of Lipinski's seminal paper[5] there appears to be a growing epidemic, of what Hann has termed "molecular obesity" [8] amongst new pharmacological compounds (Supplementary Figure 1). Compounds with higher molecular weight and lipophilicity have a higher probability of attrition at each stage of clinical development [4, 9-11]. Thus, the inflation of physico-chemical properties that increases the risks associated with clinical development may partly explain the decline in productivity of small molecule drug discovery over the past two decades[4]. However, the mean molecular properties of new pharmacological compounds are still considered Lipinski compliant, despite the fact their property distributions are far from historical norms.

Whilst the Ro5 is predictive of oral bioavailability, 16% of oral drugs violate at least one of the criteria and 6% fail two or more (although this does include natural products and substrates of transporters) (Supplementary Figure 2a and Supplementary Table 1). Notably, high profile drugs such as atorvastatin (Lipitor) and montelukast (Singulair), fail more than one of the Lipinski rules (Supplementary Figure 2b). Despite Lipinski's recommendation that the rule be considered as a guideline in reality it is routinely used to filter libraries of compounds. The implementation of rules as filters means that no discrimination is achieved beyond a qualitative pass or fail – all compounds that comply with the rules are considered equal, as are all that breach.

The response to such issues is not to define more refined rules. Instead, methods to quantify druglikeness are required [12-14]. However, scoring schemes proposed to date, often derived by machine learning methods, have lacked the intuitiveness, transparency and ease of implementation of the Ro5. To quantify compound quality we apply the concept of desirability[15] to provide a quantitative metric for assessing druglikeness that we call QED (Quantitative Estimate of Druglikeness). QED values can range between zero (all properties unfavourable) and one (all properties favourable). The desirability approach can be used to generate functions to describe any set of compounds depending on requirements. Here we will demonstrate the utility of the approach by describing desirability functions derived from a set of orally absorbed approved drugs.

Desirability provides a simple yet powerful approach to multi-criteria optimization. It is finding increasing utility in a number of applications in drug discovery including compound selection [16], library design [17, 18], molecular target prioritisation, central nervous system penetration [19] and estimating the reliability of screening data [20].The concept was originally introduced by Harrington[15] in the area of process engineering and further refined by Derringer[21]. Desirability takes multiple numeric or categoric parameters measured on different scales and describes each by an individual desirability function. These are then integrated into a single dimensionless score. In the case of compounds, a series of desirability functions (*d*) are derived, each corresponding to a different molecular descriptor. Combining the individual desirability functions into the QED is achieved by taking the geometric mean of the individual functions, as shown in Equation 1.

$$QED=\exp\left(\frac{1}{n}\Sigma_{i=1}^{n}\ln d_i\right) \quad \text{Equation (1)}$$

Conventionally, desirability functions are defined arbitrarily, usually as monotonic decreasing or increasing functions, or “hump” functions at defined parameter ranges and inflection points. Importantly, whereas previous approaches have used functions defined by user experiences and expectations[16, 19], our approach differs fundamentally in that the functions are derived empirically by describing the underlying property distributions of a set of approved drugs, much as the boundaries defined by Lipinski were. The data used comprises a carefully curated collection of 771 orally dosed approved drugs. Eight widely-used molecular properties were selected on the basis of published precedence for their relevance in determining druglikeness[3, 5, 22, 23]: molecular weight (MW), octanol-water partition coefficient (ALOGP)[24], number of hydrogen bond donors (HBD), number of hydrogen bond acceptors (HBA), molecular polar surface area (PSA), number of rotatable bonds (ROTB), the number of aromatic rings (AROM)[25, 26] and number of structural alerts (ALERTS)[27]. The molecular properties were chosen on the basis that they have all been shown to influence the likelihood of attrition and can all be calculated robustly at high-throughput. Histograms showing the distribution of the eight molecular properties across the set of oral drugs are shown in Figure 1. We found that the property distribution data are consistently best modelled as asymmetric double sigmoidal (ADS) functions, which are also shown in Figure 1 over the same range. The general ADS function is shown in Equation 2 where *d(x)* is the desirability function for molecular descriptor *x*.

$$d\,(x)=a+\frac{b}{\left[1+\exp\left(-\frac{x-c+\frac{d}{2}}{e}\right)\right]}\bullet\left[1-\frac{1}{\left[1+\exp\left(-\frac{x-c-\frac{d}{2}}{f}\right)\right]}\right] \quad \text{Equation (2)}$$

The parameters (*a, b, c, d, e* and *f*) for each of the ADS functions d_{MW}, d_{ALOGP}, d_{HBD}, d_{HBA}, d_{PSA}, d_{ROTB}, d_{AROM} and d_{ALERTS} are shown in Supplementary Table 2, as are the R^2 values and the rank amongst a library of non-linear functions.

Weighted desirability functions

The chosen molecular descriptors may vary in the importance of their contribution to druglikeness, so each can be weighted by their relative significance. The unweighted QED shown in Equation 1 would then be replaced with a weighted QED (QED_w), shown in Equation 3,

$$QED_w = \exp\left(\frac{\sum_{i=1}^{n} w_i \ln d_i}{\sum_{i=1}^{n} w_i}\right) \quad \text{Equation (3)}$$

where d is the individual desirability function, w is the weight applied to each function and n is the number of descriptors. Rather than assigning weights subjectively, we rationalized that the optimal set of weights is that which maximises information content, which can be measured by calculating the Shannon entropy[28, 29] (Supplementary Figures 3a and 3b). An exhaustive search of possible weight combinations was performed for the set of approved drugs (see Methods). Three resulting sets of weights were considered (Table 1): *i)* the set of weights that gave the maximal information content ($\mathrm{QED_w}^{\mathrm{max}}$), *ii)* the mean weights of the 1,000 weight combinations giving the highest information content ($\mathrm{QED_w}^{\mathrm{mo}}$) and *iii)* all weights as unity *i.e.* unweighted ($\mathrm{QED_w}^{\mathrm{u}}$). Interestingly, the $\mathrm{QED_w}^{\mathrm{max}}$ series gave zero weight to the PSA and HBA parameters suggesting the information in these parameters is redundant. To help explain the relative weights we performed a Principal Component Analysis of the unweighted desirability functions (Supplementary Figure 3c). The results were consistent with the entropy analysis in that the least correlated descriptors were weighted most highly. The pair-wise cross correlations between each of the properties is shown in Supplementary Figure 3d and listed in Supplementary Table 4. The complete weighted QED is given in Equation 4.

$$QED_w = \exp\left[\frac{W_{MW} \ln d_{MW} + W_{ALOGP} \ln d_{ALOG} + W_{HBA} \ln d_{HBA} + W_{HBD} \ln d_{HBD} + W_{PSA} \ln d_{PSA} + W_{ROTB} \ln d_{ROTB} + W_{AROM} \ln d_{AROM} + W_{ALERTS} \ln d_{ALERTS}}{W_{MW} + W_{ALOGP} + W_{HBA} + W_{HBD} + W_{PSA} + W_{ROTB} + W_{AROM} + W_{ALERTS}}\right] \quad \text{Eq}$$

The property descriptors that are considered, the weights given to those descriptors and the set of data that the functions are derived from, can all be varied according to requirements. In this study we have considered approved drugs that are dosed orally, but given appropriate data sets, desirability functions could be derived with relative ease to describe the relevant chemical space for parenteral administration, blood-brain barrier penetration[30] or taxonomic species with different permeability barriers.

Benchmarking

A benchmark study was designed to determine the relative performance of QED, the druglike classifiers defined by Lipinski[5], Veber[23] and Ghose[22] and the quantitative score of Gleeson *et al.*[31] in distinguishing a set of drugs from a background set of compounds. The issue of assessing whether a compound is objectively druglike or otherwise is non-trivial and, as we have already argued, such a binary classification is somewhat misleading. With this consideration in mind, for the purposes assessing the *relative* performance of rule-based classifiers and QED we attempted to benchmark their performance qualitatively. The Drugbank database [32] was used as the positive set whilst the small molecule ligands of the Protein Data Bank (PDB) were used as the negative set.

The results of the benchmark study are shown in Figure 2a and Supplementary Figure 4 in the form of a Receiver-Operator Characteristic (ROC) plot. QED outperforms the Ro5 and Ghose rules and performs marginally better than the Veber rule at a QED of 0.35 (the threshold at which Veber closely approaches that of QED). However, unlike rule-based approaches this threshold could be modulated to give different levels of sensitivity and specificity according to requirements. The Ghose rule is less sensitive but more specific than the Veber and Lipinski rules. Interestingly this benchmark suggests that the Veber rule outperforms the Ro5. $\mathrm{QED_w}^{\mathrm{mo}}$ and $\mathrm{QED_w}^{\mathrm{u}}$ outperform the quantitative measure of Gleeson regardless of threshold. $\mathrm{QED_w}^{\mathrm{max}}$ outperforms Gleeson above 0.37, below which it performs comparably. Performance of $\mathrm{QED_w}^{\mathrm{max}}$, $\mathrm{QED_w}^{\mathrm{mo}}$ and unweighted $\mathrm{QED_w}^{\mathrm{u}}$ is generally comparable, suggesting that the optimally weighted index (QED_w^{max}) can provide similar discrimination despite using fewer molecular descriptors. The best performing measure alternates between QED_w^{mo} and the unweighted QED_w^{u} depending on the range being considered, with QED_w^{max} performing marginally worse. Given the somewhat artificial nature of the benchmark, in practical terms QED_w^{mo} and QED_w^{u} could be used interchangeably; only 18 (2.3%) of the DrugStore oral drugs have ΔQED (between QED_w^{mo} and QED_w^{u}) of >0.15 and 100 drugs (13.0%) have ΔQED >0.10 (Supplementary Figures 5a and 5b). Therefore, QED_w^{mo} is used in all further analyses described here.

Direct comparison of the Ro5 and QED is illustrated in Figures 2b and 2c for the set of 771 oral drugs. An advantage of QED is its ability to rank compounds whether they fail the Ro5 or not. Interestingly, oral drugs that fail the Ro5 show QED values over a very wide range from nearly 0 to 0.8 (Figure 2c). Figure 2d shows the differences in the distribution of QED scores for compounds in the ChEMBL database of small molecule bioactivities[33], small molecule ligands from the PDB and the set of oral drugs from DrugStore used to derive the functions. Such comparative analyses provide the means of establishing the relative druglikeness of any library of compounds.

Chemical aesthetics

As beauty is in the eye of the beholder, so chemical attractiveness is in the eye of the chemist[34, 3536]. A study that compared the ability of chemists to assess druglikeness, revealed that while chemists would agree on the ‘attractive’ or drug-like structures, subjective human analysis is inconsistent in rejecting undesirable or ‘ugly’ compounds[35]. In an attempt to use chemists collective experiences as a means to evaluate druglikeness Takaoka *et al.* found the correlation coefficient between druglike scores assigned by individual chemist to be 0.5 – 0.6 [34]. Lipinski has argued that pattern recognition is the forte of the chemist[37, 38]. Wipke and Rogers have described the chemist’s knowledge of chemical structures as *Gestalt* pattern recognition process[39]. Thus we suggest that QED is an objective score that may correlate with the tacit knowledge of chemists’ subjective assessment of druglikeness or chemical attractiveness. The advantage of a codified metric on which chemical attractiveness can be judged is its application to ranking very large numbers of compounds. To aid the interpretation, it may be useful to consider QED values in the context of the observed distribution of a large reference set. To illustrate this, QED values corresponding to key percentiles from the ChEMBL database are shown in Supplementary Table 3 and a complete list is provided in the Supplementary Information.

Compared to the binary classification of the Ro5, QED exhibits a continuous scale from the most druglike drugs (Figure 3a) to the least druglike (Figure 3b). Comparison of the most druglike drugs that fail Ro5 (Figure 3c) and the least druglike drugs that pass the Ro5 (Figure 3d) illustrate the potential of QED to objectively rank compounds by the elusive quality of chemical attractiveness.

To assess whether QED reflects chemists’ opinions of chemical attractiveness we compared QED with the manually assigned annotations for 17,117 diverse compounds scored by a survey of 79 chemists from across AstraZeneca’s chemistry community (see Supplementary Information). Each chemist was asked to provide a yes or no answer to the question “would you undertake chemistry on this compound if it were a hit?” for approximately 200 compounds each. Less than one third (31.8%) of the compounds were considered as attractive chemical starting point for hit optimisation (5,457). Of 11,660 compounds that were considered unattractive, 4,497 (38.6%) were considered to be “too complex” and 5,243 (45.0%) considered “too simple”, the remainder having no reason assigned. The mean QED is 0.67 (S.D. = 0.16) for the attractive compounds, 0.49 (S.D. = 0.23) for the unattractive compounds and 0.34 (S.D. = 0.24) for the unattractive compounds considered "too complex" (Figure 3e and 3f). The difference in QEDs between the attractive and unattractive compounds is statistically significant. The estimated difference in the medians of the attractive and unattractive compounds is 0.164 (Wilcoxon rank-sum test, 95% confidence interval 0.157-0.171). The equivalent value for the difference between the attractive and the "too complex" set is 0.349 (95% confidence interval 0.340-0.358).

The Chemical Beauty of Drug Targets

A logical extension of the concept of compound druglikeness is to apply it to the problem of target druggability assessment. Hopkins and Groom [40] postulated that if there are physico-chemical limitations to the properties of compounds that are likely to be oral drugs (as Lipinski proposed [5, 6]), then drug binding sites should have complementary properties. An implication of this idea is that not all ligand binding sites have the appropriate physico-chemical and topological properties to non-covalently bind small molecule drugs with sufficient affinity. Binding sites that do have these characteristics are described as druggable. Note that this definition is independent of any wider biological considerations. A number of algorithms have been developed to determine the druggability of proteins based on analysis of the structural and physico-chemical properties of an identified binding site [41-44]. A common feature of structure-based druggability analysis methods is the classification of a binding site into the categories of druggable or undruggable based on predefined training data.

Much as we have argued for the benefits of considering druglikeness in quantitative terms, the druggability of a protein can also be considered as a continuum of chemical tractability [45] rather than as a simple binary categorical assignment, thereby enabling the prioritization of druggable binding sites. QED provides an efficient means to quantify and rank the druggability of targets according to the chemical attractiveness of their associated ligands. QEDs were calculated for each compound in the ChEMBL database[33] of published bioactivity (release ChEMBL09) having an affinity <10uM for a defined human protein target. The resulting 167,045 compounds are associated with 1,729 human proteins.

Top ranking targets by three different schemes are shown in Table 2. The first scheme involves ranking targets by the mean QED of their associated ligands (Table 2 and Supplementary Table 5). The mean QED for all targets in the list is 0.478. For the targets of approved drugs the mean QED is 0.492 and for the targets of approved oral drugs the mean QED is 0.539 (with an average standard deviation for a target of 0.231). Drug targets are indeed enriched towards the more highly desirable targets with 70% of the drug targets being found in the top 50% of the prioritized target list.

Within a set of ligands for a target, it is useful to consider the QED of distinct chemical series, as even targets that are perceived as being relatively intractable may have a small proportion of associated chemical matter that is druglike and of potential interest. To approximate distinct chemical series, all by all Tanimoto similarity matrices[46] were calculated for each of the 1,729 human protein targets in ChEMBL. Compound similarity is represented as a network with chemical series being identified as distinct subgraphs within the network using a Tanimoto similarity threshold of 0.7. We define a chemical series as a cluster comprising at least 5 compounds and an active chemical series as one where the proportion of actives is at least 0.7 (with an activity threshold of 10μM or Ligand Efficiency of 0.3). The number of compounds, series and active series for all ChEMBL targets is listed in the Supplementary Information. Chemical similarity networks for four targets are shown in Figure 4. The chemical network representation in Figure 4 illustrates the presence of highly desirable chemotypes even for some targets with low mean QED. The mean QED of the most druglike active series for each target provides the second ranking scheme, listed in Table 2 and Supplementary Table 6. The mean QED of the best compound cluster of all ChEMBL targets is 0.569 (where the cluster comprises at least 5 compounds). The mean QED of the best cluster of human drug targets is 0.693. The mean QED when considering only the best cluster of the targets of oral drugs is 0.766.

A third approach to ranking targets is to consider the degree of enrichment of druglike series. Here, targets are ranked by the proportion of active series that have a mean QED above that of the top 10% of the ChEMBL database (0.796) (Table 2 and Supplementary Table 7).

Conclusion

QED provides the means to rank chemical structures by their merit relative to a target function, which in this case are the properties of oral drugs. Furthermore, by extension of the concept to the set of ligands associated with a drug target, QED provides an efficient means to quantify and rank the druggability of targets. Lipinski's Rule of Five has gained considerable traction in early stage drug discovery largely because it is predictive, intuitive and simple to implement. We believe QED compares favourably in each of these regards. Compared to the rule-based approaches QED offers a richer, more nuanced view of druglikeness. The QED functions are based on the underlying distribution data of drug properties and unlike rule-based metrics can identify cases when a generally unfavourable property can be tolerated where the other parameters are close to ideal. In so doing the phenomenon of druglikeness is evolved from a binary 'black and white' assessment to a more realistic and gradated description of the continuum of compound quality.

METHODS

Data set of known drugs

A non-redundant data set comprising 771 approved drugs was derived from the ChEMBL DrugStore database[47]. The selected compounds were all (i) marketed drugs, (ii) classified as small molecular weight therapeutics (i.e. no nutritional supplements, diagnostic agents or biologics), (iii) of specified molecular structure, (iv) composed of at least six atoms, (v) dependent on a biological macromolecule for their mode of action (i.e. exclude chelators and buffers), (vi) orally administered, (vii) systemically absorbed (i.e. exclude compounds whose site of action is in the gastro-intestinal (GI) tract e.g. orlistat targets gastric lipase, acarbose targets enteric alpha glucosidase).

Molecular properties

Physico-chemical properties were calculated using the Pipeline Pilot Chemistry Collection (version 8.0.1.500) from Accelrys (San Diego, CA, USA). The properties calculated were Molecular Weight (MW), octanol-water partition coefficient (ALOGP) (using the atom-based method by Ghose and Crippen[24]), number of hydrogen bond donors (HBD), number of hydrogen bond acceptors (HBA), molecular polar surface area (PSA), number of rotatable bonds (ROTB) and the number of aromatic rings (AROM)[25, 26]. Finally, a substructure search was performed against each drug using a curated reference set of 94 functional moieties that are potentially mutagenic, reactive or have unfavourable pharmacokinetic properties[27]. The number of matches for each compound was captured (ALERTS). We chose to omit the acid dissociation constant (pKa) as the available high-throughput computational approaches do not provide sufficient accuracy[48].

Fitting of Desirability functions

Histograms were plotted reflecting the distribution of each of the 8 molecular properties for the oral drugs. For discrete variables (HBD, HBA, ROTB, AROM, ALERTS) a bin size of 1 was used. For continuous variables (MW, ALOGP and PSA) the optimal bin size D was estimated by optimising the cost function C(Δ)[49] (Equation 5):

$$C\,(\Delta) = \frac{2k - v}{\Delta^2} \quad \text{Equation (5)}$$

where k and v are the mean and variance of the occupancy of bins of size Δ respectively. For PSA a local minimum C(Δ) was used. A library of functions was fitted to the distributions using TableCurve 2D version 5.01 (Systat Software, CA, USA). Asymmetric Double Sigmoidal (ADS) functions (Equation 2) were found to be the most consistently high-ranking non-linear functions (Supplementary Table 1) and also reflected the important underlying asymmetry. Each function was then normalized by dividing by the maximum function value $d(x)_{max}$ to give a value between 0 and 1.

QED

The individual desirability functions were combined into the QED by taking the geometric mean, which, by logarithmic identities, can be expressed as the exponent of the arithmetic mean of the logarithm transformed identities (Equations 1 and 3).

Assignment of Weights

We rationalized that the optimal set of weights is that which maximises information content, as measured by Shannon entropy[28] (Equation 6):

$$ShannonEntropy_w = -\sum_{i=1}^{n} QED_w \log_2 QED_w \quad \text{Equation (6)}$$

where QED_w is the weighted QED calculated with a set of weights w. Each possible combination of weights between 0 and 1 at increments of 0.25 were exhaustively enumerated for all 8 molecular descriptors, giving 5^8 (390,625) weight combinations of for each of the 771 drugs. The combination of weights giving the highest entropy gives QED_w^{max} (Table 1). Inspection of the ranked weight combinations revealed a “spike” of higher entropy values over the highest-scoring 1,000 combinations (Supplementary Figure 3a). The mean of each individual molecular property weight over these 1,000 highest ranked entropy scores gives the mean optimal weighted QED_w^{mo} (Table 1). QED_w^{mo} may more accurately sample the high entropy combinations whilst attenuating the quantized nature of the weight increments. The robustness of this procedure was established by assessing the relationship of individual descriptor weights to the ranked entropy scores compared to a randomized series (Supplementary Figure 3b).

Principle Component Analysis (PCA)

PCA was performed on the 8 unweighted desirability functions calculated on the ChEMBL database (release ChEMBL09) using Pipeline Pilot’s R Statistics Component Collection (Supplementary Figure 3c).

Benchmark study

The benchmarking assessment involves assignment of positive and negative compound sets. The DrugBank database[32] was used to derive the positive set. 771 compounds having the word “oral” in their “Route of Administration” field were selected. Whilst we endeavoured to obtain a truly independent positive set for the benchmark inevitably significant overlap was found between the DrugBank set and the drugs used to derive QED. 554 of the 771 compounds were structurally identical and a further 30 had significant structural similarity (Tanimoto score > 0.8). Small molecule ligands from the Protein Data Bank's (PDB's) Ligand Dictionary[50] was selected as the negative set as it provides a large and diverse source of chemical tools, metabolites, natural products, crystallographic buffers as well as drugs. To prevent ambiguity, 475 compounds were removed that had significant structural similarity to the positive set (Tanimoto score > 0.8), leaving a negative set of 10,250.

Performance measures

The following performance measures were used:

$$Sensitivity = \frac{TP}{TP+FN} \quad \text{Equation (7)}$$

$$Specificity = \frac{TN}{FP+TN} \quad \text{Equation (8)}$$

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP+FP) * (TP+FN) * (TN+FP) * (TN+FN)}} \quad \text{Equation (9)}$$

Where MCC = Mathews Correlation Coefficient, TP = True Positives, TN = True Negatives, FP = False Positives and FN = False Negatives.

Target Druggability Methods

The ChEMBL database includes a highly heterogeneous assortment of published bioactivity data. Bioactivity endpoints were only considered when (i) there was a defined protein molecular target, (ii) the activity type was either IC_{50}, K_i or K_d, (iii) the relation was '=', '<' or '<=', (iv) standard units were defined as 'nM' and (v) the activity was greater than 10^{-6} nM (largely to remove misannotations). A broad range of bioactivity values are typically reported for a given combination of target and ligand due to a combination of biological, technical and annotation errors. Selection of the "correct" value is non-trivial, particularly when using large-scale automated procedures. Simple calculation of a mean is sensitive to outliers. As such, for each combination of target and ligand we identified the modal log unit of bioactivity and calculated the mean value of activities within that range. Consideration of only human targets results in 167,045 unique compounds being associated with 1,729 proteins, giving 310,551 compound target pairs.

For each protein target the Tanimoto structural similarity of each associated compound to every other associated compound was calculated using Pipeline Pilot (FCFP_4 fingerprints) to give an all-against-all similarity matrix. Compound networks were derived from these matrices using the Python package NetworkX.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This research received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement N° 223461 and the Scottish Universities Life Sciences Alliance. We thank John Overington for providing the DrugStore data, Ruth Brenk for the provision of SMARTS for structural alerts and Ian Carruthers for assistance with DrugStore database queries. We also express our thanks to the chemistry community of AstraZeneca for participating in the chemistry survey.

References

1. Keller TH, Pichota A, Yin Z. A practical view of 'druggability'. Curr. Opin. Chem. Biol. 2006; 10:357–361. [PubMed: 16814592]
2. Ursu O, Rayan A, Goldblum A, Oprea TI. Understanding drug-likeness. Wiley Interdis. Rev.: Comp. Mol. Sci. 2011; 1 doi: 10.1002/wcms.1052.
3. Oprea TI. Property distribution of drug-related chemical databases. J. Comput. Aided Mol. Des. 2000; 14:251–264. [PubMed: 10756480]
4. Leeson PD, Springthorpe B. The influence of drug-like concepts on decision-making in medicinal chemistry. Nature Rev. Drug Discov. 2007; 6:881–890. [PubMed: 17971784]
5. 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 Del. Revs. 1997; 23:3–25.
6. Lipinski CA. Drug-like properties and the causes of poor solubility and poor permeability. J. Pharmacol. Toxicol. Methods. 2000; 44:3–25.
7. Abad-Zapatero C. A sorcerer's apprentice and The Rule of Five: from rule-of-thumb to commandment and beyond. Drug Discov. Today. 2007; 12:995–997. [PubMed: 18061876]
8. Hann MM. Molecular obesity, potency and other addictions in drug discovery. MedChemComm. 2011 DOI: 10.1039/c1md00017a.
9. Hughes JD, et al. Physicochemical drug properties associated with in vivo toxicological outcomes. Bioorg. Med. Chem. Lett. 2008; 18:4872–4875. [PubMed: 18691886]
10. Wenlock M, 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]
11. Proudfoot JR. The evolution of synthetic oral drug properties. Bioorg. Med. Chem. Lett. 2005; 15:1087–1090. [PubMed: 15686918]
12. Xu J, Stevenson J. Drug-like index: a new approach to measure drug-like compounds and their diversity. J. Chem. Inf. Comput. Sci. 2000; 40:1177–1187. [PubMed: 11045811]
13. Rayan A, Marcus D, Goldblum A. Predicting oral druglikeness by iterative stochastic elimination. J. Chem. Info. Model. 2010; 50:437–445.
14. Ohno K, Nagahara Y, Tsunoyama K, Orita M. Are there differences between launched drugs, clinical candidates, and commercially available compounds? J. Chem. Inf. Model. 2010; 50:815–821. [PubMed: 20394396]
15. Harrington EC Jr. The Desirability Function. Ind. Qual. Control. 1965; 21:494–498.
16. Cruz-Monteagudo M, et al. Desirability-based methods of multiobjective optimization and ranking for global QSAR Studies. Filtering safe and potent drug candidates from combinatorial libraries. J. Comb. Chem. 2008; 10:897–913. [PubMed: 18855460]
17. Le Bailly de Tilleghem C, Beck B, Boulanger B, Govaerts B. A fast exchange algorithm for designing focused libraries in lead optimization. J. Chem. Inf. Model. 2005; 45:758–767. [PubMed: 15921465]
18. Mandal A, Johnson K, Wu CFJ, Bornemeier D. Identifying Promising Compounds in Drug Discovery: Genetic Algorithms and Some New Statistical Techniques. J. Chem. Inf. Model. 2007; 47:981–988. [PubMed: 17425300]
19. Wager TT, Hou X, Verhoest PR, Villalobos A. Moving beyond Rules: The Development of a Central Nervous System Multiparameter Optimization (CNS MPO) Approach To Enable Alignment of Druglike Properties. ACS Chemical Neuroscience. 2010; 1
20. Paolini GV, Lyons R, Laflin P. How desirable are your IC50s? A method to enhance screening-based decision making. J. Biomol. Screen. 2010; 15:1183–1193. [PubMed: 20980692]
21. Derringer G, Suich R. Simultaneous optimization of several response variables. J. Quality. Technol. 1980; 12:214–219.

22. Ghose AK, Viswanadhan VN, Wendoloski JJ. A knowledge-based approach in designing combinatorial or medicinal chemistry libraries for drug discovery. 1. A qualitative and quantitative characterization of known drug databases. J Comb Chem. 1999; 1:55–68. [PubMed: 10746014]

23. Veber DF, et al. Molecular properties that influence the oral bioavailability of drug candidates. J. Med. Chem. 2002; 45:2615–2623. [PubMed: 12036371]

24. Ghose AK, Crippen GMJ. Atomic Physicochemical Parameters for Three-Dimensional Structure-Directed Quantitative Structure-Activity Relationships I. Partition Coefficients as a Measure of Hydrophobicity. J. Comput. Chem. 1986; 7:565–577.

25. Lovering F, Bikker J, Humblet C. Escape from Flatland: Increasing Saturation as an Approach to Improving Clinical Success. J. Med. Chem. 2009; 52:6752–6756. [PubMed: 19827778]

26. Ritchie TJ, Macdonald SJ. The impact of aromatic ring count on compound developability - are too many aromatic rings a liability in drug design? Drug Discov. Today. 2009; 14:1011–1120. [PubMed: 19729075]

27. Brenk R, et al. Lessons learnt from assembling screening libraries for drug discovery for neglected diseases. Chem Med Chem. 2008; 3:435–444. [PubMed: 18064617]

28. Shannon CE. A Mathematical Theory of Communication. Bell System Technical Journal. 1948; 27:379–423.

29. Hosseinzadeh Lotfi F, Fallahnejad R. Imprecise Shannon's Entropy and Multi Attribute Decision Making. Entropy. 2010; 12:53–62.

30. Wager TT, et al. Defining Desirable Central Nervous System Drug Space through the Alignment of Molecular Properties, in Vitro ADME, and Safety Attributes. ACS Chemical Neuroscience. 2010; 1:420–434. [PubMed: 22778836]

31. Gleeson MP, Hersey A, Montanari D, Overington J. Probing the links between in vitro potency, ADMET and physicochemical parameters. Nature Rev. Drug Discov. 2011; 10:197–208. [PubMed: 21358739]

32. Knox C, et al. DrugBank 3.0: a comprehensive resource for 'omics' research on drugs. Nucleic Acids Res. 2011; 39:D1035–1041. [PubMed: 21059682]

33. ChEMBL. https://www.ebi.ac.uk/chembldb/

34. Takaoka Y, et al. Development of a method for evaluating drug-likeness and ease of synthesis using a data set in which compounds are assigned scores based on chemists' intuition. J. Chem. Inf. Comput. Sci. 2003; 43:1269–1275. [PubMed: 12870920]

35. Lajiness MS, Maggiora GM, Shanmugasundaram V. Assessment of the consistency of medicinal chemists in reviewing sets of compounds. J. Med. Chem. 2004; 47:4891–4896. [PubMed: 15369393]

36. Muresan, S.; Sadowski, J. Molecular Drug Properties - Measurement and Prediction. Mannhold, R., editor. Wiley-VCH Verlag; 2008.

37. Lipinski, CA. Molecular Informatics: Confronting Complexity. Beilstein-Institut; Bozen, Italy: 2002.

38. Lipinski CA. Overview of Hit to Lead: The Medicinal Chemist's Role from HTS Retest to Lead Optimisation hand Off Top. Med. Chem. 2009; 5:1–24. [PubMed: 19149644]

39. Wipke WT, Rogers D. Artificial intelligence in organic synthesis. SST: starting material selection strategies. An application of superstructure search. J. Chem. Inf. Comput. Sci. 1984; 24:71–81. [PubMed: 6547445]

40. Hopkins AL, Groom CR. The Druggable Genome. Nat. Rev. Drug Discov. 2002; 1:727–730. [PubMed: 12209152]

41. An J, Totrov M, Abagyan R. Comprehensive identification of "druggable" protein ligand binding sites. Genome Inform. 2004; 15:31–41. [PubMed: 15706489]

42. Cheng AC, et al. Structure-based maximal affinity model predicts small-molecule druggability. Nat. Biotechnology. 2007; 25:71–75.

43. Halgren TA. Identifying and characterizing binding sites and assessing druggability. J. Chem. Inf. Model. 2009; 49:377–389. [PubMed: 19434839]

44. Schmidtke P, Barril X. Understanding and predicting druggability. A high-throughput method for detection of drug binding sites. J. Med. Chem. 2010; 53:5858–5867. [PubMed: 20684613]

45. Southan C, Boppana K, Jagarlapudi SA, Muresan S. Analysis of in vitro bioactivity data extracted from drug discovery literature and patents: Ranking 1654 human protein targets by assayed compounds and molecular scaffolds. J. Cheminform. 2011; 3:14. [PubMed: 21569515]

46. Tanimoto TT. IBM Internal Report. 1957

47. Overington JP, Al-Lazikani B, Hopkins AL. How many drug targets are there? Nat. Rev. Drug Discov. 2006; 5:993–996. [PubMed: 17139284]

48. Manchester J, Walkup G, Rivin O, You Z. Evaluation of pKa estimation methods on 211 druglike compounds. J. Chem. Inf. Model. 2010; 50:565–571. [PubMed: 20225863]

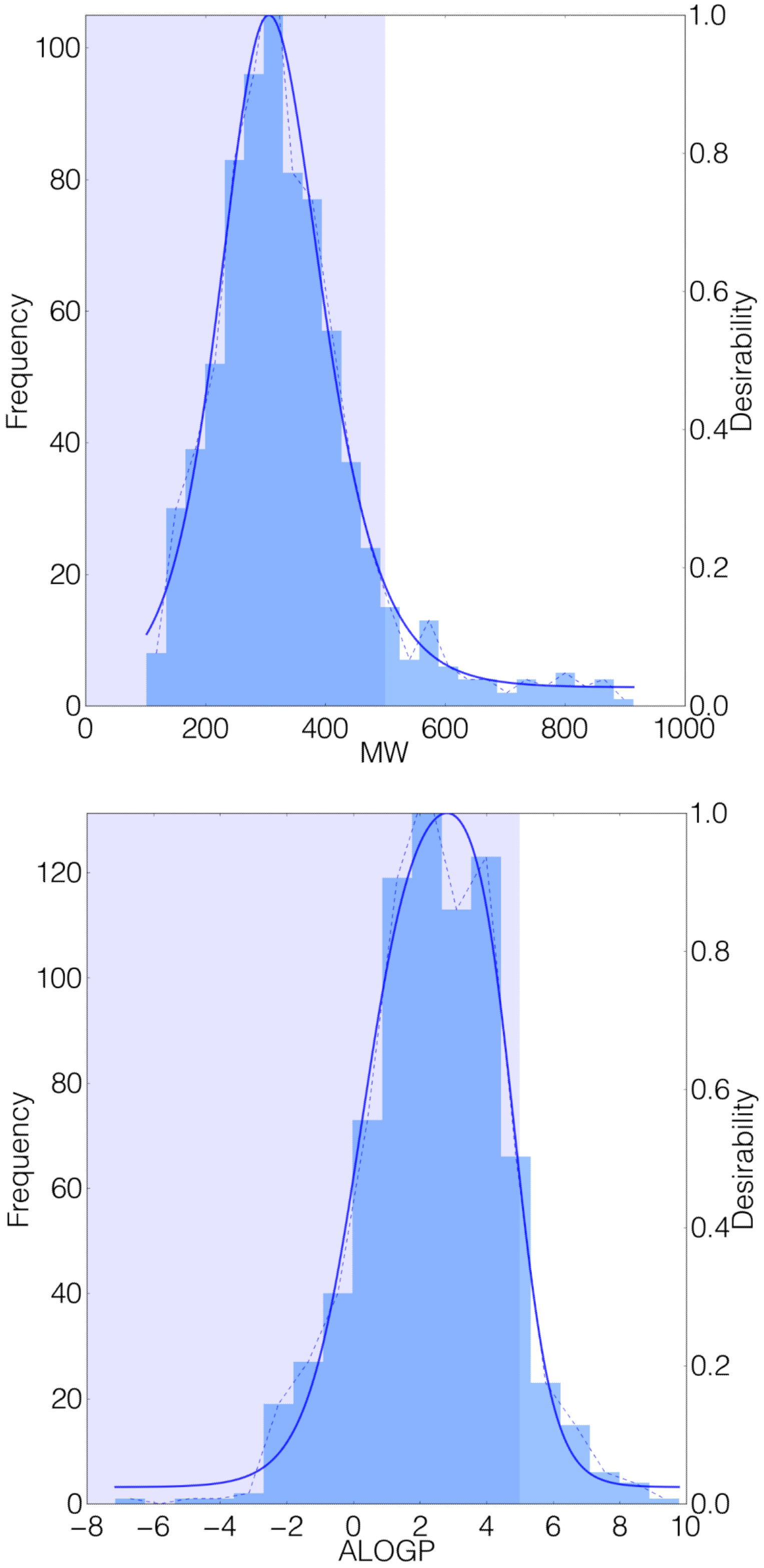
49. Shimazaki, H.; Shinomoto, S. Advances in Neural Information Processing Systems. Schölkopf, B.; Platt, J.; Hoffman, T., editors. Vol. 19. 2007.

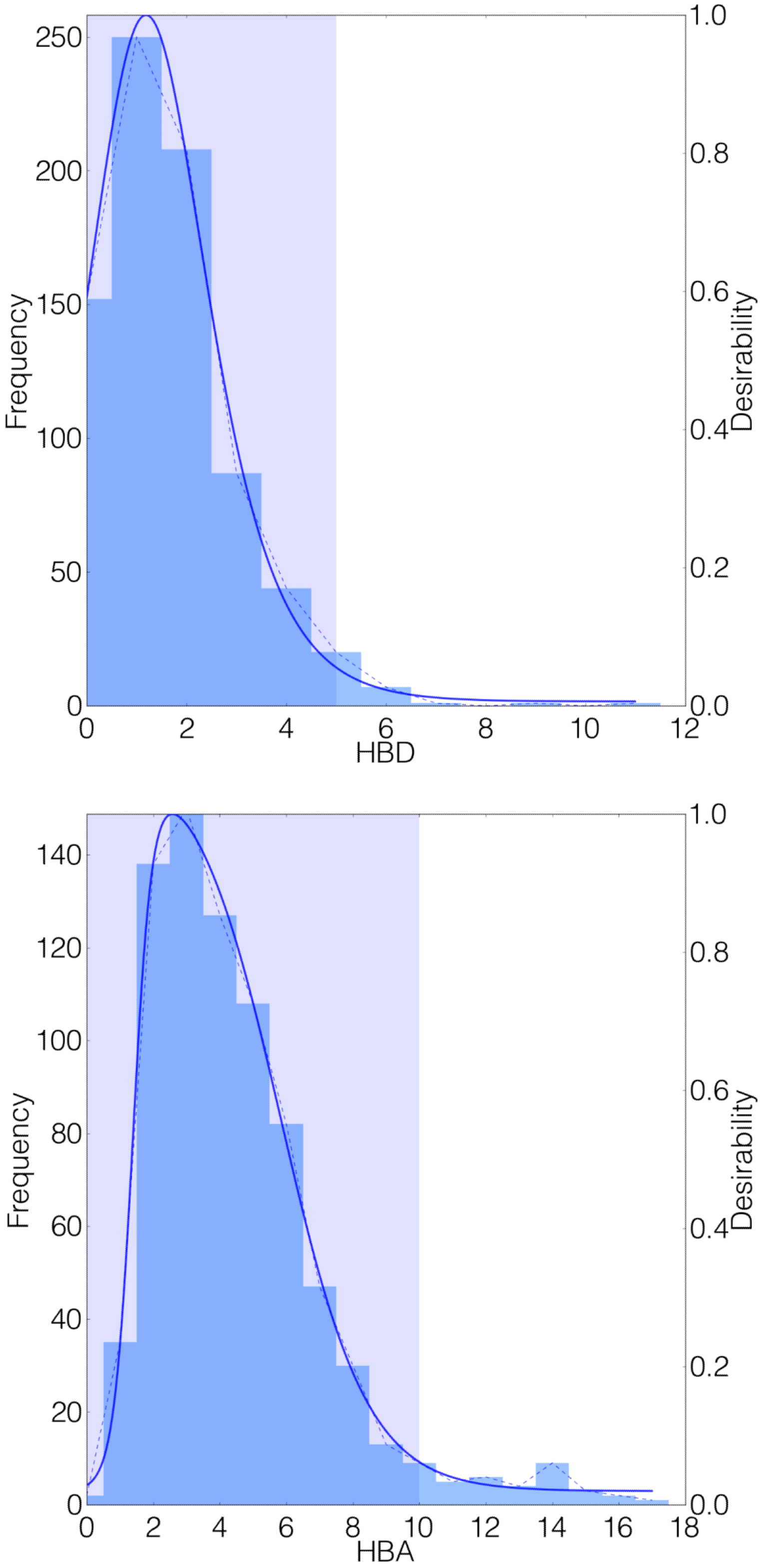
50. Dimitropoulos, D.; Ionides, J.; Henrick, K. Current Protocols in Bioinformatics. Baxevanis, AD.; Page, RDM.; Petsko, GA.; Stein, LD.; Stormo, GD., editors. John Wiley & Sons; Hoboken, NJ, USA: 2006. p. 14.13.11-14.13.13.

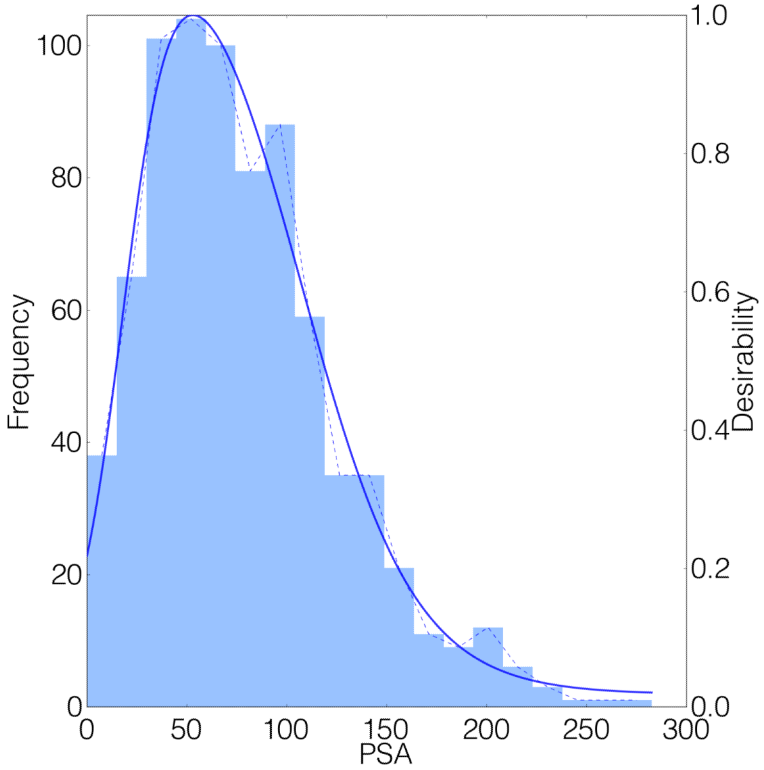
51. Gleeson MP. Generation of a set of simple, interpretable ADMET rules of thumb. J. Med. Chem. 2008; 51:817–834. [PubMed: 18232648]

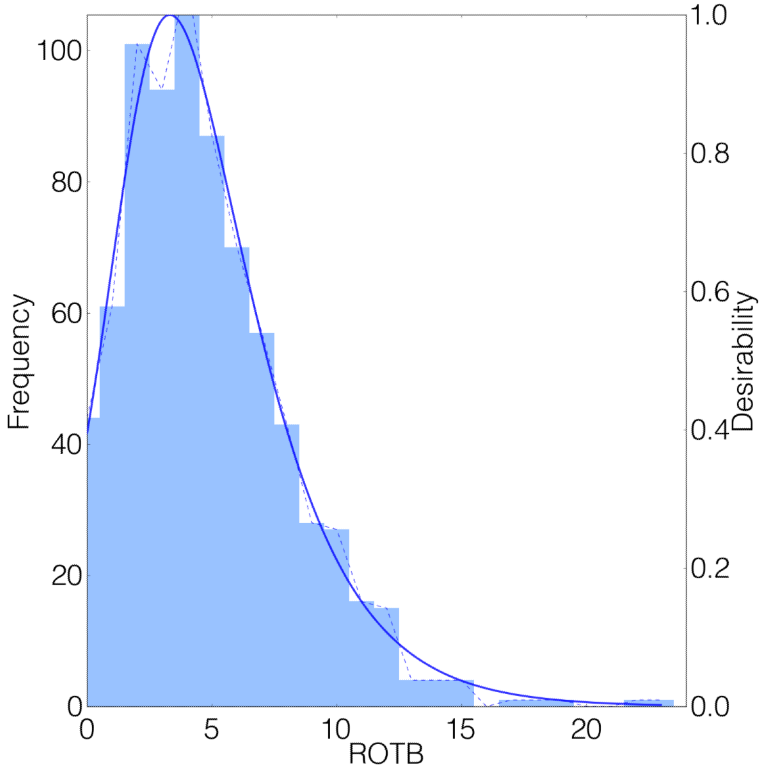
52. Congreve M, Carr R, Murray C, Jhoti H. A ‘rule of three’ for fragment-based lead discovery? Drug Discov. Today. 2003; 8:876–877. [PubMed: 14554012]

53. Luker T, et al. Strategies to improve in vivo toxicology outcomes for basic candidate drug molecules. Bioorganic & Medicinal Chemistry Letters. 2011 In Press.


Frequency
0
20
40
60
80
100
Desirability
0.0
0.2
0.4
0.6
0.8
1.0
0
200
400
600
800
1000
MW
Frequency
0
20
40
60
80
100
120
Desirability
0.0
0.2
0.4
0.6
0.8
1.0
−8
−6
−4
−2
0
2
4
6
8
10
ALOGP


Frequency
250
200
150
100
50
0
0
2
4
6
8
10
12
HBD
Desirability
1.0
0.8
0.6
0.4
0.2
0.0
Frequency
140
120
100
80
60
40
20
0
0
2
4
6
8
10
12
14
16
18
HBA
Desirability
1.0
0.8
0.6
0.4
0.2
0.0


Frequency
Desirability
PSA
0
20
40
60
80
100
0
50
100
150
200
250
300
0.0
0.2
0.4
0.6
0.8
1.0


Frequency
Desirability
ROTB
0
20
40
60
80
100
0
5
10
15
20
0.0
0.2
0.4
0.6
0.8
1.0

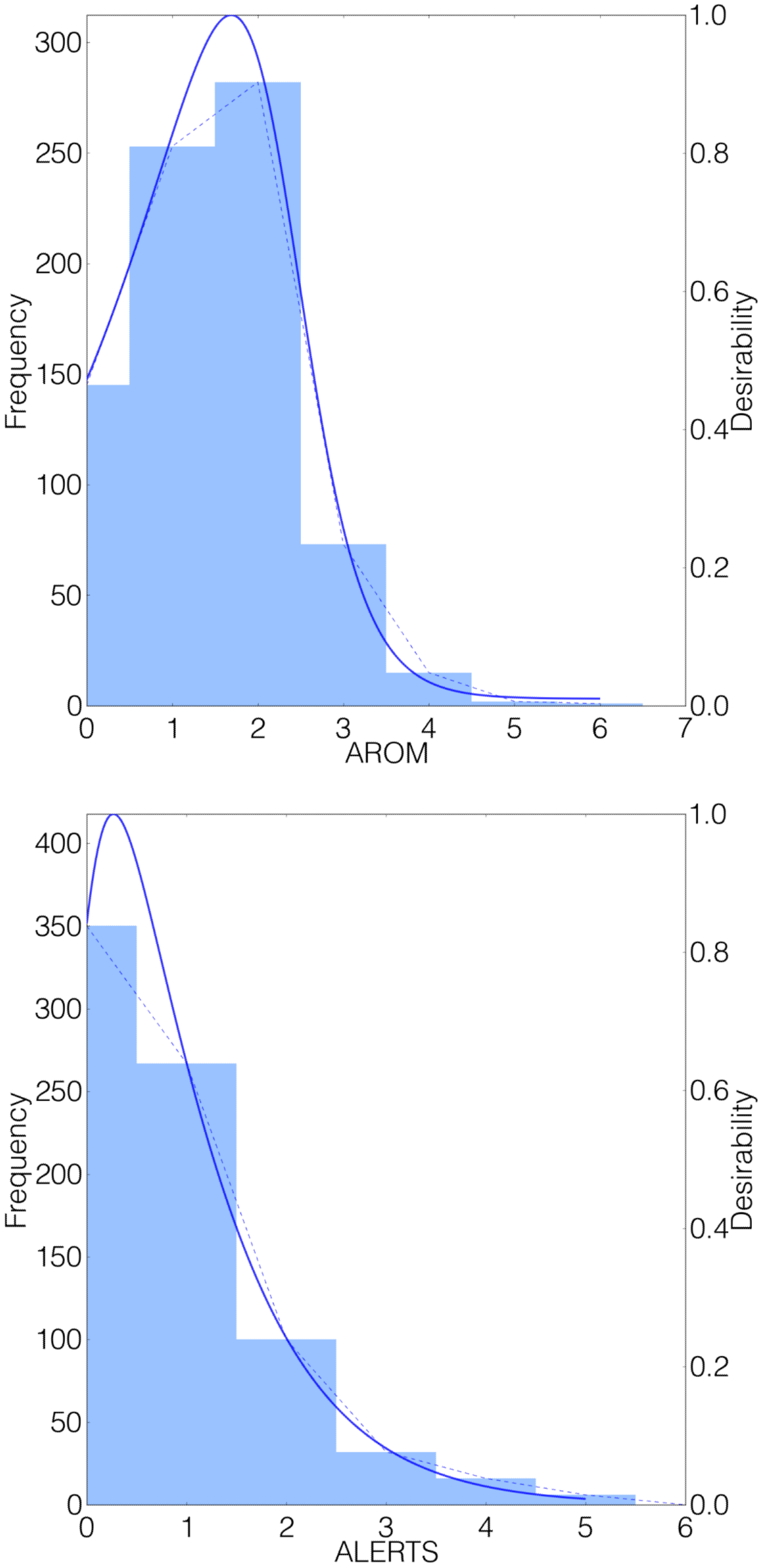
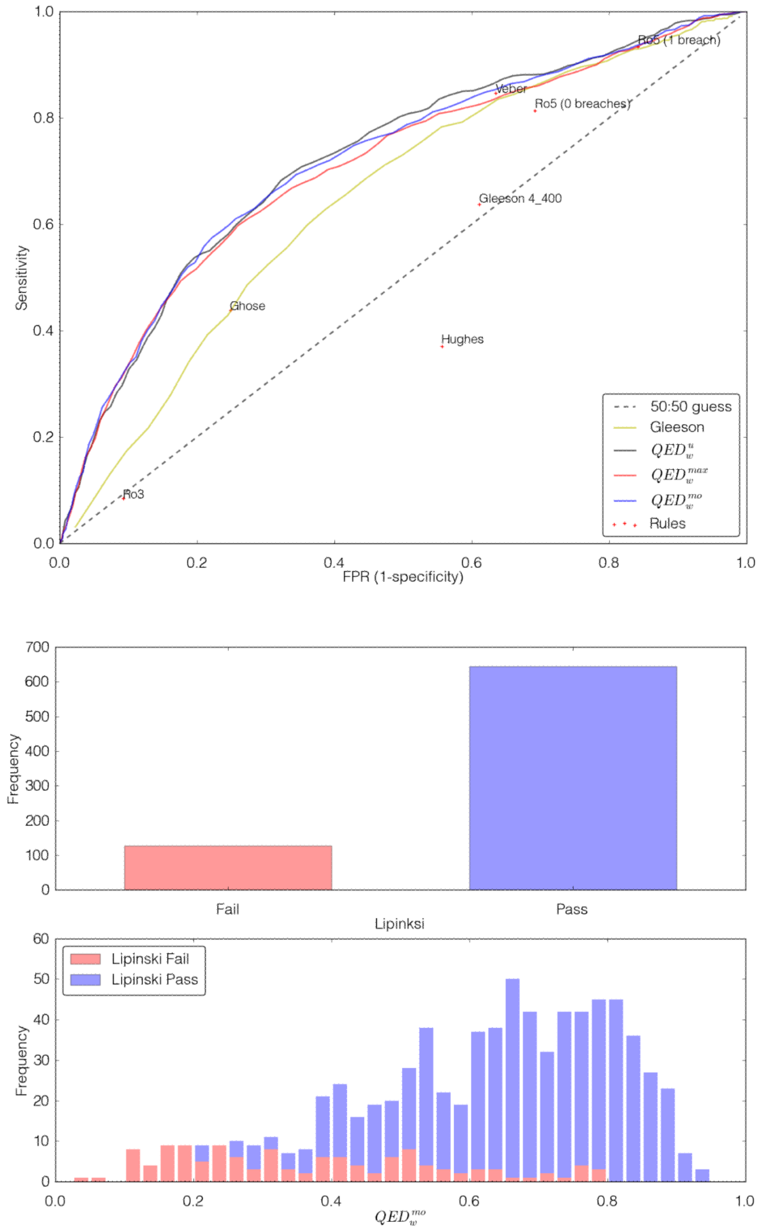


Figure 1. Histograms of 8 selected molecular properties for a set of 771 orally absorbed small molecule drugs

The solid blue line describes the Asymmetric Double Sigmoidal (ADS) functions (Equation 2) used to model the histogram. The parameters for each function are shown in Supplementary Table 1. The Lipinski compliant areas are shown in pale blue in **Figures 1 (a), (b), (c)** and **(d)**. The molecular properties are: **(a)** Molecular Weight (MW), **(b)** Lipophilicity estimated by atomic based prediction of octanol-water partition coefficient (ALOGP), **(c)** number of hydrogen bond donors (HBD), **(d)** number of hydrogen bond acceptors (HBA), **(e)** polar surface area (PSA), **(f)** number of rotatable bonds (ROTB), **(g)** number of aromatic rings (AROM) and **(h)** number of structural alerts (ALERTS).


Sensitivity
FPR (1-specificity)
0.0
0.2
0.4
0.6
0.8
1.0
Ro5 (1 breach)
Veber
Ro5 (0 breaches)
Gleeson 4_400
Ghose
Hughes
Ro3
50:50 guess
Gleeson
QED_w^u
QED_w^{max}
QED_w^{mo}
Rules
Frequency
700
600
500
400
300
200
100
0
Fail
Pass
Lipinksi
60
50
40
30
20
10
Lipinski Fail
Lipinski Pass
QED_w^{mo}

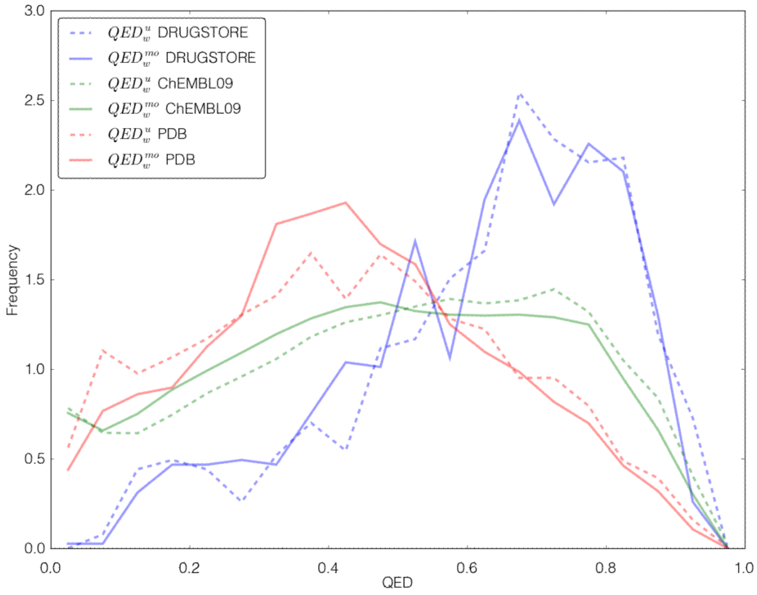
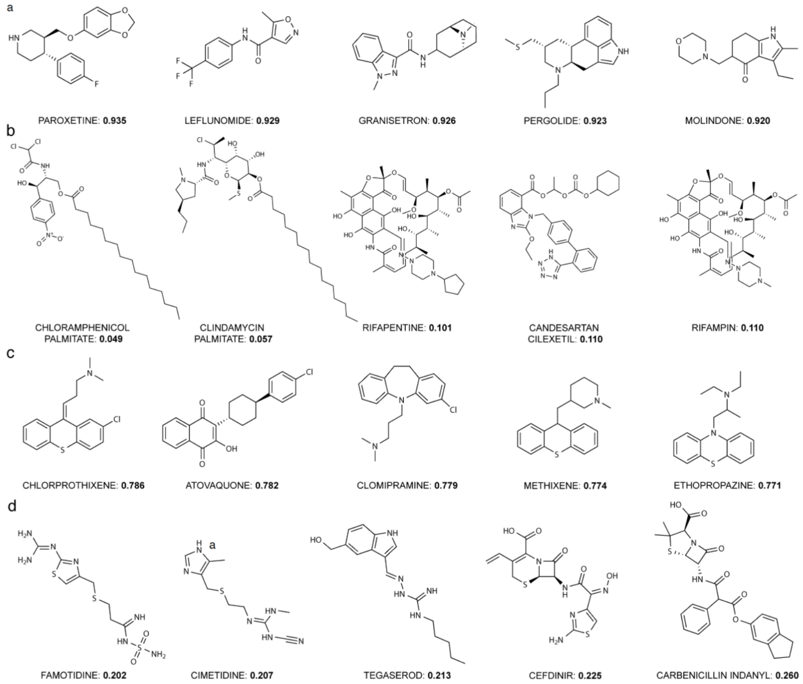
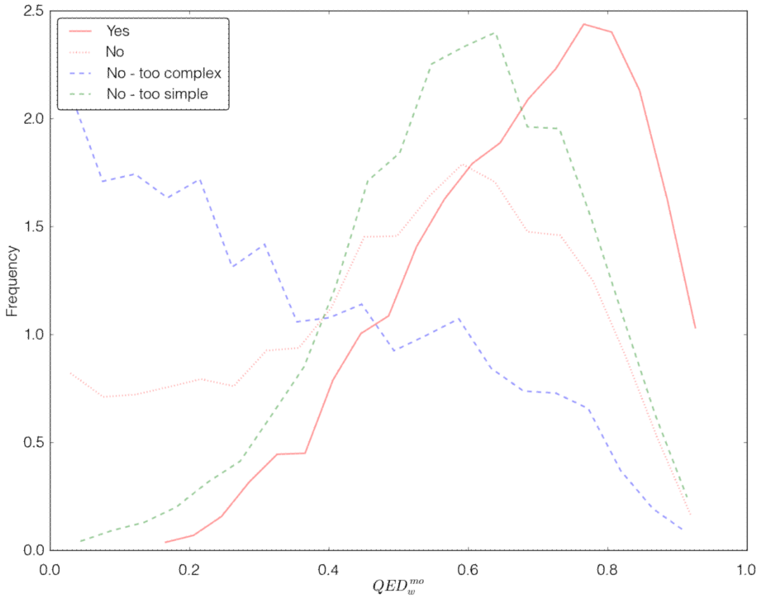


Figure 2. Benchmarking of QED against other measures of druglikeness

(a) ROC curve (Receiver operating characteristic) of true positive rate (sensitivity) against false positive rate (1-specificity) describing the difference in performance of different approaches in classifying compounds as druglike or otherwise. The performance of the rules of Lipinski[5], Veber[23] and Ghose[22], Gleeson (4/400)[51], Congreve (Ro3)[52], Hughes[9] and the quantitative method of Gleeson[31] is compared to three different QED weighting schemes (maximal entropy (QED_w^{max}), mean optimal entropy (QED_w^{mo}) and unweighted (QED_w^{u})). Veber *et al.* observed that compounds with fewer than 10 rotatable bonds and Polar Surface Area (PSA) less than or equal to 140Å^2 (or fewer than or equal to 12 hydrogen donors and acceptors) had an increased oral bioavailability in rats[23]. Ghose *et al.* suggested a qualifying range that could be used in the development of druglike chemical libraries and recommended the following constraints: molecular weight between 160 and 480; calculated logP between -0.4 and 5.6; molar refractivity between 40 and 130 and total number of atoms between 20 and 70[22]. Gleeson *et al.* has proposed the most desirable region for ADME properties lies between MW<400 and AlogP<4[51] and recently suggested a quantitative ADMET score based on molecular weight and AlogP [31]. For comparison the [52] 'Rule of Three' for fragment selection (Ro3) is also plotted (where MW <300, AlogP ≤3, PSA ≤60, the number of hydrogen bond donors ≤3, the number of hydrogen bond acceptors ≤3). At a threshold that provides an equivalent level of sensitivity as the Ro5, a QED_w^{mo} of 0.40 offers 48% greater specificity than the Ro5. Equally, for the same degree of specificity as the Ro5 a QED_w^{mo} of 0.26 offers 12% greater sensitivity. The dashed line represents the line of no discrimination – the level of performance that would be achieved by a random guess. **(b)** Direct comparison of Ro5 and QED. Drugs failing (red) and passing (blue) Lipinski's Ro5. **(c)** Equivalent plot of the QED results of the same set of compounds. The overlapping distributions indicate the greater resolution provided by the quantitative measure – some rather druglike Lipinski failures are observed as are some undruglike passes. **(d)** QED distribution for three small molecule databases: the ChEMBL database of small molecule bioactivities (green), small molecule ligands from the PDB (red) and the set of oral drugs used to derive the functions (blue). Both weighted (QED_w^{mo}) (solid lines) and unweighted (QED_w^{u}) (dashed line) indices are shown.


a
PAROXETINE: 0.935
LEFLUNOMIDE: 0.929
GRANISETRON: 0.926
PERGOLIDE: 0.923
MOLINDONE: 0.920
b
CHLORAMPHENICOL PALMITATE: 0.049
CLINDAMYCIN PALMITATE: 0.057
RIFAPENTINE: 0.101
CANDESARTAN CILEXETIL: 0.110
RIFAMPIN: 0.110
c
CHLORPROTHIXENE: 0.786
ATOVAQUONE: 0.782
CLOMIPRAMINE: 0.779
METHIXENE: 0.774
ETHOPROPAZINE: 0.771
d
FAMOTIDINE: 0.202
CIMETIDINE: 0.207
TEGASEROD: 0.213
CEFDINIR: 0.225
CARBENICILLIN INDANYL: 0.260


Yes
No
No - too complex
No - too simple
2.5
2.0
1.5
1.0
0.5
0.0
Frequency
0.0
0.2
0.4
0.6
0.8
1.0
QED_w^{mo}

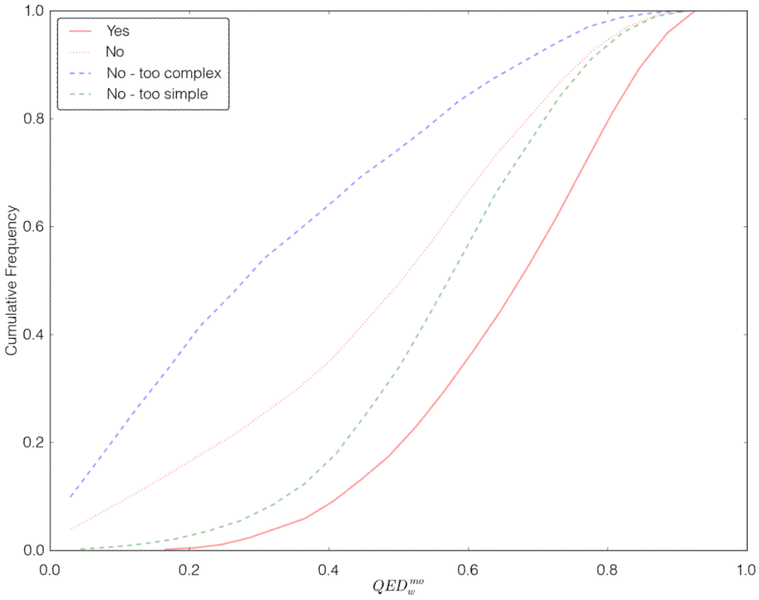
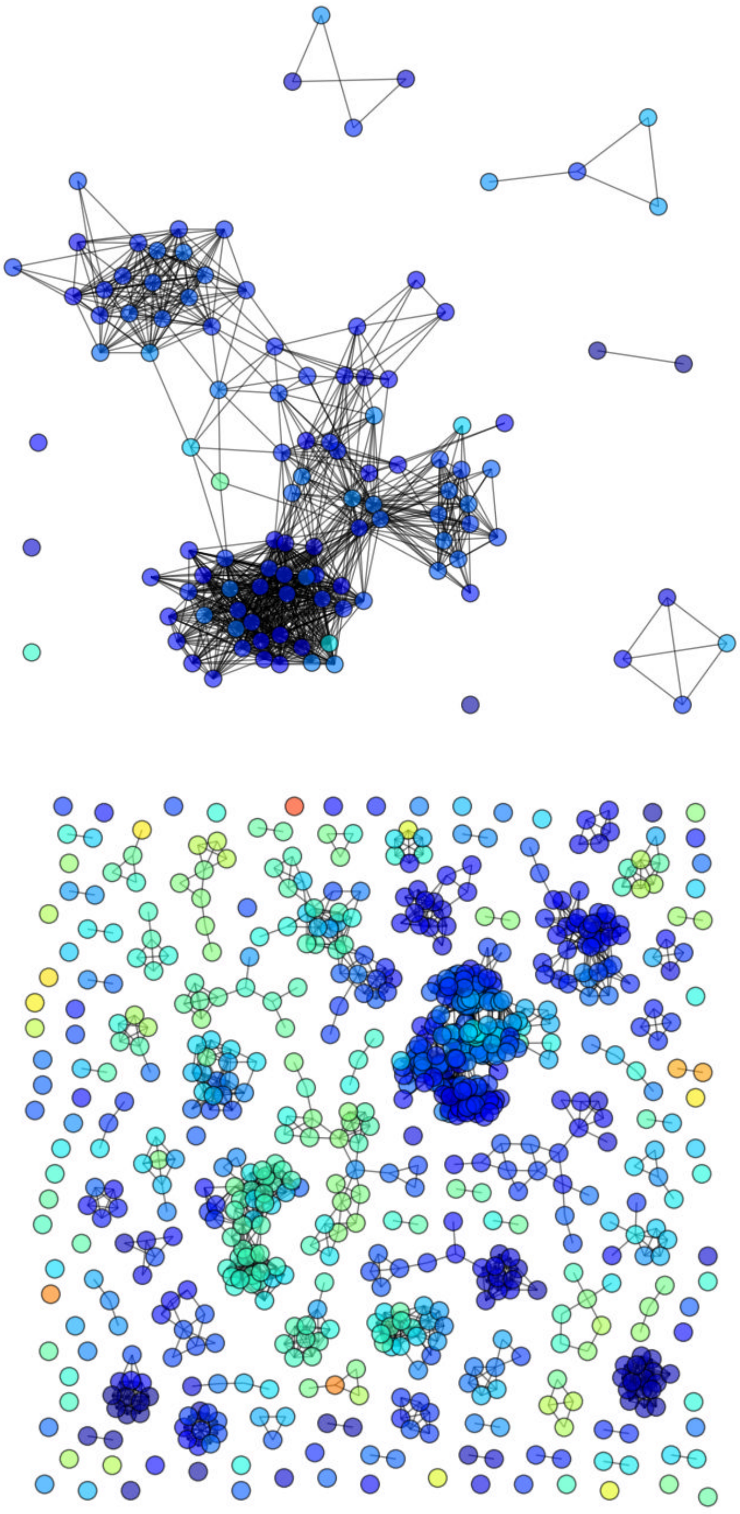


Figure 3. Chemical aesthetics

Illustrative subsets of the oral drugs from DrugStore. (a) The 5 most druglike drugs. (b) The 5 least druglike drugs. (c) The 5 most druglike Ro5 failures. (d) The 5 least druglike Ro5 passes (also see Supplementary Figure 8). (e) Results of chemical survey: QED distributions between compounds annotated chemically attractive and unattractive. (f) Cumulative QED distribution of chemical survey results.



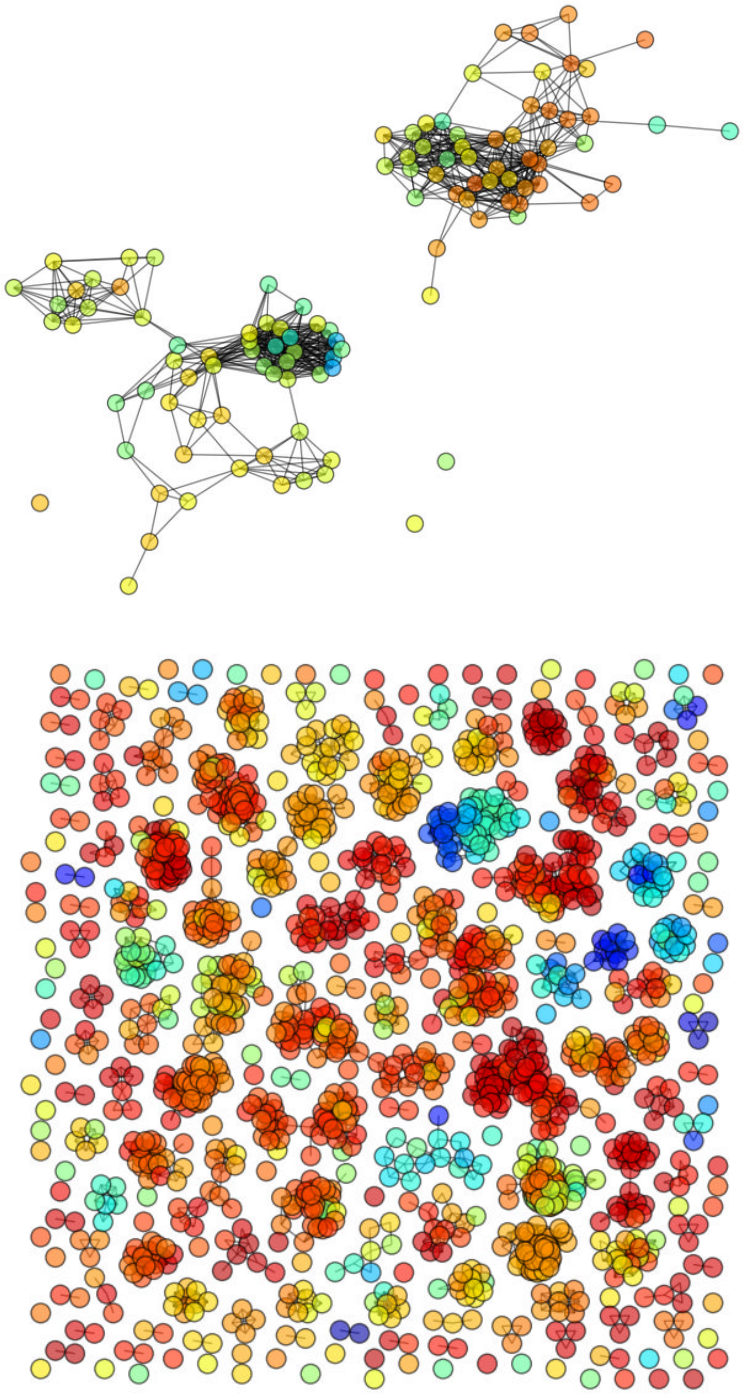


Figure 4. Structural diversity networks

In each of the networks compounds are represented as nodes and are coloured by their respective QED values. An edge connects nodes if they are structurally similar (defined by a Tanimoto threshold of >= 0.7). The networks provide a useful way of summarizing a large amount of data describing the published bioactivity data for a target in an intuitive and visually digestible form. The four targets were chosen as they each have a considerable number of associated compounds but illustrate the importance of considering druglikeness and chemical diversity when prioritizing targets. **(a)** Structural diversity network for matriptase, a target whose associated bioactive compounds are neither druglike nor diverse. **(b)** Structural diversity network for plasminogen, a target whose published bioactive compounds are diverse but not druglike. **(c)** Structural diversity network for 1-acylglycerol-3-phosphate O-acyltransferase beta, a target whose published bioactive compounds are druglike but not diverse. **(d)** Structural diversity network for norepinephrine transporter, a target whose published bioactive compounds are both druglike and diverse. The network images are generated by the open source graph visualization software GraphViz.

Table 1

Optimized desirability function weightings by Shannon entropy.

	Shannon Entropy	Rank	MW	ALOGP	HBD	HBA	PSA	ROTB*	AROM*	ALERTS
QED_w^{max}	293.42	1	0.50	0.25	0.50	0.00	0.00	0.50	0.25	1.00
QED_w^{mo}	293.03	1-1000	0.66	0.46	0.61	0.05	0.06	0.65	0.48	0.95
QED_w^{u}	283.08	81,657	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

* The relatively high weightings of the number of aromatic rings (AROM) and the number of rotatable bonds (ROTB) is consistent with recent analysis from AstraZeneca which found aromaticity and flatness to be important factor in attrition due to toxicology [53].

Table 2

Top human targets* by 3 different ranking schemes.

	TARGET (UNIPROT)	MEAN QED	TARGET (UNIPROT)	MEAN QED BEST LUSTER	TARGET (UNIPROT)	PROPORTION CLUSTERS MEAN QED >0.796
1	Free fatty acid receptor 2 (015552)	0.861	Neuropeptide Y receptor type 5 (Q15761)	0.935	Vesicular acetylcholine transporter (Q16572)	0.714
2	Sodium channel (Q9NY72, 060939, Q8IWT1, Q07699)	0.849	Serotonin transporter (P31645)	0.932	Phosphodiesterase 7A (Q13946)	0.667
3	Voltage-gated potassium channel (P15382, P51787)	0.835	Serotonin 1a (5-HT1a) receptor (P08908)	0.932	Melatonin receptor 1A (P48039)	0.5
4	Phosphodiesterase 9A (076083)	0.820	Norepinephrine transporter (P23975)	0.932	Melatonin receptor 1B (P49286)	0.462
5	Aldo-keto-reductase family 1 member C3 (P42330)	0.812	Dopamine transporter (Q01959)	0.931	Norepinephrine transporter (P23975)	0.453
6	Cholinergic receptor, nicotinic, beta 1 (Muscle) (Q8IZ46)	0.811	Histamine H1 receptor (P35367)	0.928	Histamine H4 receptor (Q9H3N8)	0.444
7	Sorbitol dehydrogenase (Q00796)	0.809	Dopamine D3 receptor (P35462)	0.927	Dopamine transporter (Q01959)	0.409
8	Sodium channel protein type IV alpha subunit (P35499)	0.809	Dopamine D4 receptor (P21917)	0.925	Serotonin 7 (5-HT7) receptor (P34969)	0.4
9	Endothelial lipase (Q9Y5X9)	0.804	Thromboxane-A synthase (P24557)	0.921	Neuronal acetylcholine receptor protein beta-4 subunit (P30926)	0.4
10	Vesicular acetylcholine transporter (Q16572)	0.798	Serotonin 2c (5-HT2c) receptor (P28335)	0.917	Neuronal acetylcholine receptor protein alpha-7 subunit (P36544)	0.4
11	Lipoxin A4 receptor (P25090)	0.794	Serotonin 2a (5-HT2a) receptor (P28223)	0.916	Hormone sensitive lipase (Q05469)	0.4
12	Small chondroitin/dermatan sulfate proteoglycan (Q99645)	0.790	Aldose reductase (P15121)	0.912	Serotonin 5a (5-HT5a) receptor (P47898)	0.4

	TARGET (UNIPROT)	MEAN QED	TARGET (UNIPROT)	MEAN QED BEST LUSTER	TARGET (UNIPROT)	PROPORTION CLUSTERS MEAN QED >0.796
13	Hypoxia-inducible factor prolyl 4-hydroxylase (Q9NXG6)	0.789	Alkaline phosphatase, tissue-nonspecific isozyme (P05186)	0.912	Histamine H1 receptor (P35367)	0.381
14	Huntingtin (P42858)	0.788	Dipeptidyl peptidase IV (P27487)	0.911	Metabotropic glutamate receptor 5 (P41594)	0.375
15	DNA-directed RNA polymerase (Q9Y2Y1, Q9H1D9, P05423)	0.786	Metabotropic glutamate receptor 5 (P41594)	0.911	Carbonic anhydrase VA (P35218)	0.375
16	Amine oxidase, copper containing (Q16853)	0.786	Neuronal acetylcholine receptor protein alpha-7 subunit	0.910	Quinone reductase 2 (P16083)	0.375
17	Nuclear factor of activated T-cells cytoplasmic (Q13469, 095644)	0.769	(P36544) Butyrylcholinesterase (P06276)	0.904	Serotonin transporter (P31645)	0.369
18	Sodium channel protein type VII alpha subunit (Q01118)	0.766	HERG (Q12809)	0.903	Dipeptidyl peptidase II (Q9UHL4)	0.333
19	Alkaline phosphatase, tissue-nonspecific isozyme (P05186)	0.764	Serine/threonine-protein kinase RAF (P04049)	0.902	Serotonin 3a (5-HT3a) receptor (P46098)	0.333
20	Serotonin 1f (5-HT1f) receptor (P30939)	0.762	Adenosine A1 receptor (P30542)	0.901	Steryl-sulfatase precursor (P08842)	0.333

[*] Only targets with at least 10 associated compounds are considered. Full data for the top ranking targets are provided in Supplementary Tables 5, 6 and 7.