J Comput Aided Mol Des (2013) 27:1–13
DOI 10.1007/s10822-012-9631-5

PERSPECTIVE

Inflation of correlation in the pursuit of drug-likeness

Peter W. Kenny · Carlos A. Montanari

Received: 29 October 2012 / Accepted: 28 December 2012 / Published online: 10 January 2013

Abstract Drug-likeness is a frequently invoked, although not always precisely defined, concept in drug discovery. Opinions on drug-likeness are to a large extent shaped by the relationships that are observed between surrogate measures of drug-likeness (e.g. aqueous solubility; permeability; pharmacological promiscuity) and fundamental physicochemical properties (e.g. lipophilicity; molecular size). This article draws on examples from the literature to highlight approaches to data analysis that exaggerate trends in data and the term correlation inflation is introduced in the context of drug discovery. Averaging groups of data points prior to analysis is a common cause of correlation inflation and results from analysis of binned continuous data should always be treated with caution.

Keywords ADMET · Correlation · Drug-likeness · Lipophilicity · Solubility · Promiscuity

Introduction

The concept of drug-likeness and the idea that the physicochemical properties of a compound determine its pharmacokinetic behaviour are integral to modern drug discovery. Data analysis provides much of the basis of drug-likeness and shapes the perceptions, beliefs and opinions of drug discovery scientists. In this article, we examine some of the misleading conclusions that can arise from partitioning data into subsets (usually termed bins) prior to analysis. Noisy data can be transformed into a smaller set of data points that do illustrate a trend. However, this will usually exaggerate the strength of the trend and one can say that transforming the primary data has led to correlation inflation. Averaging the data in each bin prior to analysis hides variation, creating the illusion that the observed trend explains more of the variation than it actually does. Correlation inflation can also arise from using standard error [1] rather than standard deviation to quantify intra-bin variation and to do so is to confuse a trend's strength with its statistical significance. Binning continuous data transforms it into categorical data and, to be useful in decision making, categorical data analysis needs to focus on effect size [1, 2] and not be distracted by statistical significance. We illustrate these points using simple, synthetic data and highlight a number of relevant studies from the literature of drug-likeness.

Much of the focus on drug-likeness is due to the difficulties that the pharmaceutical industry has been experiencing in recent years and reducing attrition in clinical development will be essential if Pharma is to be commercially viable. Projects are seen to fail because compounds are insufficiently drug-like and terms ADMET (absorption, disposition, metabolism, excretion, toxicity) and compound quality will often be encountered in this context. Not all attrition is equal and stopping a project after a Phase I study will always more be palatable than being forced to do so by toxicity during a Phase III study. It does not always make sense to blame attrition on poor drug-likeness of compounds. For example, a failure to demonstrate efficacy in Phase II may indicate that the target is simply not relevant

Electronic supplementary material The online version of this article (doi:10.1007/s10822-012-9631-5) contains supplementary material, which is available to authorized users.

P. W. Kenny (✉) · C. A. Montanari
Grupo de Estudos em Química Medicinal de Produtos Naturais, NEQUIMED-PN, Instituto de Química de São Carlos, Universidade de São Paulo, Av. Trabalhador Sancarlense, 400, São Carlos, SP 13560-970, Brazil
e-mail: pwk.pub.2008@gmail.com

to the human disease and a compound will not usually enter Phase II studies unless Phase I evaluation has concluded that exposure will be sufficient to engage the therapeutic target(s). However, it is not generally possible (at least in live humans) to measure free, intracellular concentration for an arbitrary compound and, for targets not in direct contact with the circulation, such as those protected by the Blood Brain Barrier, one can argue that a typical Phase I trial provides an incomplete description of distribution. Assessments of target 'druggability' [3] do need to take more account of target location.

No discussion of drug-likeness would be complete without reference to the influential Rule of 5 (Ro5) [4] which is essentially a statement of property distributions for compounds taken into Phase II clinical trials. The focus of Ro5 is oral absorption and the rule neither quantifies the risks of failure associated with non-compliance nor provides guidance as to how sub-optimal characteristics of compliant compounds might be improved. It also raises a number of questions. What is the physicochemical basis of Ro5′s asymmetry with respect to hydrogen bond donors and acceptors? Why is calculated octanol/water partition coefficient (ClogP) used to specify Ro5′s low polarity limit when the high polarity cut off is defined in terms of numbers of hydrogen bond donors and acceptors? It is possible that these characteristics reflect the relative inability [5] of the octanol/water partitioning system to 'see' donors (Fig. 1) and the likelihood that acceptors (especially as defined for Ro5) are more common than donors in pharmaceutically-relevant compounds. The importance of Ro5 is that it raised awareness across the pharmaceutical industry about the relevance of physicochemical properties. The wide acceptance of Ro5 provided other researchers with an incentive to publish analyses of their own data and those who have followed the drug discovery literature over the last decade or so will have become aware of a publication genre that can be described as 'retrospective data analysis of large proprietary data sets' or, more succinctly, as 'Ro5 envy'.

In pharmaceutical research, a number of ADMET assays (e.g. aqueous solubility, permeability, plasma protein binding, bioavailability in rats, hERG blockade) are run in order to assess the suitability of compounds for dosing in humans as therapeutic agents. Running these assays consumes significant resource and there is much interest in prediction of ADMET characteristics from fundamental physicochemical properties and calculated molecular descriptors. Although multivariate approaches can be used, a number of published studies assert the importance of individual descriptors, such as calculated octanol/water partition coefficient (ClogP) and molecular weight (MW) as determinants of drug-likeness and compound quality and it is these relationships which are the focus of this article. It

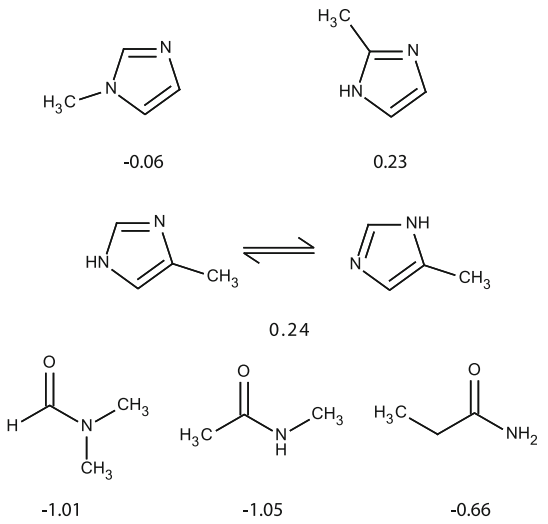


Fig. 1 Measured values of $\log P_{oct}$ suggest that the octanol/water partitioning system does not 'see' hydrogen bond donors. Data from LOGPOW database (http://logkow.cisti.nrc.ca/logkow/index.jsp)

is important to remember that the primary function of ADMET assays is to screen and dynamic range may be sacrificed for throughput. Some results will be out-of-range and this does not mean that these results are without value or that the assay is 'bad'. Out-of-range data values challenge current data-analytic capability and there is need for methodology that can include both in-range and out-of-range data within a single unified framework. One justification for binning data is to display mixed in-range and out-of-range data [6] and the case for binning is much weaker when all data points are in-range. Although beyond the scope of this study, it should also be remembered that it is not always clear how predictive ADMET assays are of clinical outcomes in live humans [7].

Despite widespread belief that control of fundamental physicochemical properties is important in pharmaceutical design, the correlations between these and ADMET properties may not actually be as strong as is often assumed. The mere existence of a trend is of no interest in drug discovery and strengths of trends must be known if decisions are to be accurately described as data-driven. Although data analysts frequently tout the statistical significance of the trends that their analysis has revealed, weak trends can be statistically significant without being remotely interesting. We might be confident that the coin that lands heads up for 51 % of a billion throws is biased but this knowledge provides little comfort for the person charged with predicting the result of the next throw. Weak trends can be beaten and when powered by enough data, even the feeblest of trends acquires statistical significance.

Although drug discovery is often described as a process of multiobjective optimization [8], it can be argued that minimization of therapeutic dose is its principal objective.

A drug functions by binding to its target(s) and no amount of drug-likeness, 'quality' or ligand efficiency [9] will rescue a compound from inadequate affinity. The essence of the drug design problem lies in the difficulty of finding compounds for which the relevant properties are all within acceptable ranges. Even defining acceptable ranges that are relevant is not trivial and the value of one property will in general affect the acceptability limits for others. For example, plasma protein binding should be seen in the broader context of distribution [7] and it can be argued that a compound that is poorly permeable or highly ionized needs to be more soluble than one that is highly permeable or neutral [4]. Adhering rigidly to a requirement that the hERG IC_{50} exceed 10 μM implies a high degree of confidence in the quality of dimethylsulfoxide stock solutions that may not always be justified.

Lipophilicity, usually defined as the logarithm of either the octanol/water partition coefficient (P) or distribution coefficient (D), is the most fundamental physicochemical property used in drug discovery [10]. To some extent, lipophilicity is to medicinal chemists what interest rates are to central bankers. While some characteristics of compounds tend to improve as lipophilicity is increased, most tend to deteriorate and lead optimization can be seen as searching for a 'sweet spot' as illustrated in Fig. 2. The separation and steepness of the curves will determine the characteristics (and existence) of the 'sweet spot' and the parallels with the molecular complexity model introduced by Hann and colleagues [11] will be obvious to many. With real projects, potency against the primary target(s) and the different ADMET characteristics will each respond differently to lipophilicity. Furthermore, real compounds will scatter about the response curves and it will be some time before drugs are designed in the way that aircraft are [12]. It can be helpful to think of molecular design as prediction-driven or hypothesis-driven and the latter can be seen as a framework for establishing structure–activity relationships as efficiently as possible [13]. It is not currently possible to predict affinity and ADMET properties with sufficient accuracy for prediction-driven molecular design and most medicinal chemistry design involves a large component that is hypothesis-driven. Although difficult to define objectively, the characteristic of a lead series that is most prized by medicinal chemists is room to maneuver. Prediction is very difficult, Niels Bohr is said to have observed, especially about the future and the human liver remains an effective antidote to the hubris of the drug designer.

Correlation inflation in analysis of drug-likeness is typically the result of partitioning data into bins and averaging quantities by bin prior to analysis. The data used in published analyses of drug-likeness are rarely made available and we have created synthetic data sets for this study to show how transforming primary data can exaggerate trends. Correlation inflation becomes an issue when the results of data analysis are used to make real decisions. To restrict values of properties such as lipophilicity more stringently than is justified by trends in the data is to deny one's own drug-hunting teams room to maneuver while yielding the initiative to hungrier, more agile competitors.

Data sets

While the data used to illustrate links between drug-likeness surrogates and physicochemical properties are rarely available, it is possible to create data sets to show how transforming primary data can exaggerate trends and inflate correlation. In this study the strengths of trends were quantified by the Pearson product-moment correlation coefficient, r [14]. The extent to which correlation is inflated typically increases with the size of the data set and two groups of data set were created for this study. The three data sets within each group can be thought of as random samples of different sizes that have been drawn from the same population. Generation of the data sets is illustrated in Fig. 3 and the starting point for each of these was the line of equality ($Y = X$) in the region $0 \leq X \leq 10$. The group 1 data sets were generated by first placing identical numbers (10, 100 or 1,000) of data points for integer values of X on this line. Data points (40, 400, 4,000) were distributed uniformly on the line of equality to provide the starting points for the group 2 data sets. Normally-distributed random numbers with zero mean were added to the Y coordinates of each starting data set. The standard deviation for the Normally-distributed noise was set to the difference (10 units) between the minimum and maximum values of Y in the data before the noise had been added. The use of the Pearson correlation coefficient to quantify the strength of relationship is appropriate because in each data set the underlying relationship is linear and values of X are uniformly distributed. The group 1 data sets were used both as generated and with the categories defined by the value of Y. The group 2 data sets were only used with their X-coordinates binned (each bin representing an equal number of data points). A published [15, 16] data set was used to explore the relationship between solubility and fraction of sp^3 carbon atoms and octanol/water log P values in Fig. 1 were obtained from the LOGKOW database [17]. The OEChem toolkit [18] was used to count carbon atoms and sp^3 carbon atoms using [C,c] and [CX4] as substructural targets defined in SMARTS notation [19]. The JMP software [20] was used for statistical analysis and all data sets used in this study are available as supplemental material.

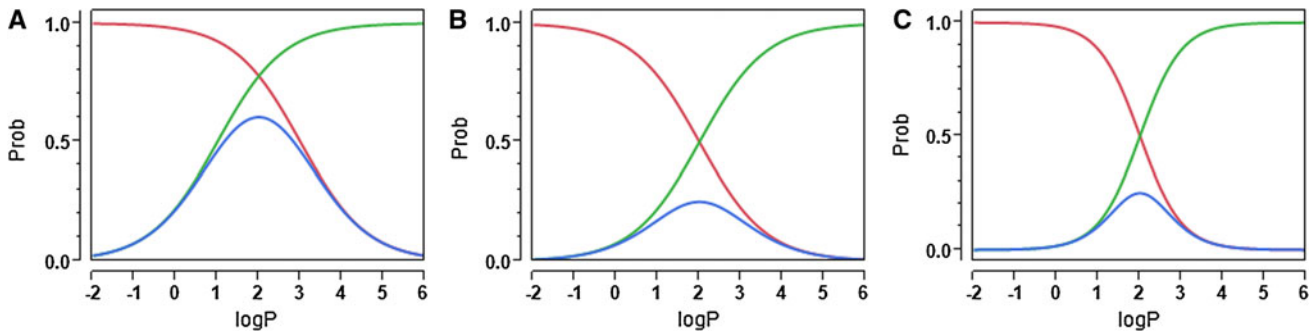


Fig. 2 The *red* and *green curves* represent probabilities of achieving satisfactory properties as a function of log P and in each case the maximum of the *blue curve*, corresponds to a log P value of 2. Displacement of the *red* and *green curves* in **a** relative to each other reduces the probability of success as given by the probability maximum in the *blue curve* in B. Increasing the steepness of the *red* and *green curves* in **b** reduces the extent of the 'sweet spot' (and, by implication, the project team's room to manoeuver) shown as the *blue curve* in **c**. Although this may represent a useful conceptual framework, the functional forms of the *red* and *green curves* (modelled here using the tanh function) will in general not be known and, even within series, there will be scatter about the curves

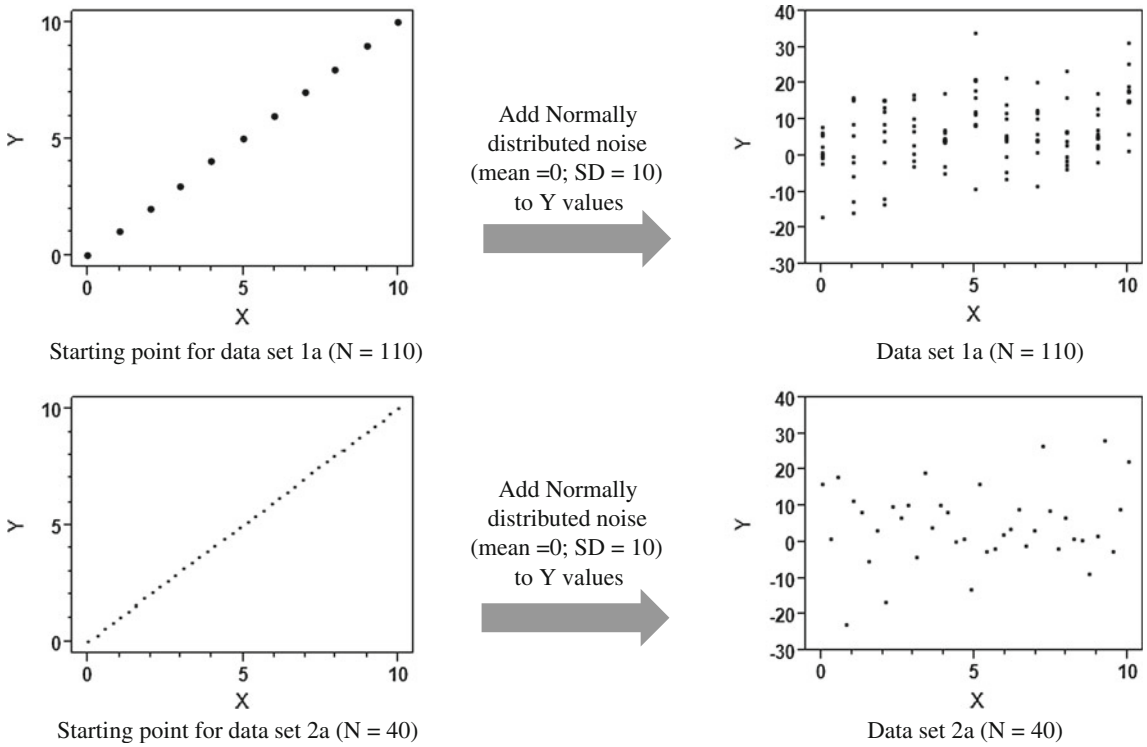


Fig. 3 Procedures for generating data sets 1 and 2 by adding Normally distributed random numbers to data points on line of unity

Inflating correlation by hiding variation

One way that correlation can be inflated is by hiding variation, for example, by representing groups of data points by average values. Table 1 illustrates the effects on correlation coefficient of applying this transformation to data sets 1a, 1b and 1c and it can be seen that the degree to which correlation is inflated increases with the size of the data set. While X explains <10 % of the variance in Y for data set 1c, it explains 99 % of the variance in mean (Y). The effects of correlation inflation can also be seen in Fig. 4. These results show that the observation of strong correlations between variables that have been transformed in this manner cannot be used to support assertions that the correlations between variables in the untransformed data are similarly strong.

Data analysis analogous to that presented in Fig. 4 and Table 1 has been used to explore relationships between pharmacological promiscuity and descriptors of lipophilicity and molecular size. In the HMO2006 study [21] promiscuity was defined as the number of targets against which an $IC_{50} \leq 10$ μM was observed. As is conventional when

Table 1 Correlation coefficients for primary and transformed data sets 1 and 2

Data set	Correlation	N	r[a]	Lower 95 %	Upper 95 %
1a	X and Y	110	0.3372	0.1600	0.4933
1a	X and mean (Y)	11	0.6727	0.1221	0.9067
1a	X and median(Y)	11	0.5352	−0.0952	0.8592
1b	X and Y	1,100	0.3045	0.2499	0.3572
1b	X and mean (Y)	11	0.9337	0.7588	0.9830
1b	X and median (Y)	11	0.9632	0.8607	0.9907
1c	X and Y	11,000	0.3074	0.2903	0.3242
1c	X and mean (Y)	11	0.9965	0.9860	0.9991
1c	X and median (Y)	11	0.9961	0.9847	0.9990
2a	X and Y	40	0.1241	−0.1950	0.4194
2b	X and Y	400	0.2936	0.2013	0.3807
2c	X and Y	4,000	0.2832	0.2544	0.3114

[a] Pearson product-moment correlation coefficient

fitting, the coefficient of determination, r^2, was quoted and for linear models this is simply the square of the correlation coefficient. The mean (MW) for each promiscuity level was plotted against promiscuity and an r^2 value of 0.93 was observed when a straight line was fit to this data:

> Analysis of this data showed a strong correlation between the mean molecular weight of compounds in each bin and the total number of targets that those compounds were found to be active against at the threshold level [21].

The analysis does indeed show that mean (MW) for compounds with activity exceeding the threshold for a given number of targets decreases with the number of targets and the r^2 value indicates a strong relationship. The HMO2006 study states:

> The standard deviation of the molecular weight is high for each activity bin, but an overall trend clearly emerges [21].

This observation is particularly relevant because the degree to which transforming the primary data inflates correlation generally increases with these standard deviations. The results in Table 2 suggest that the r^2 value for fitting Mean (MW) to number of targets is likely to be larger than that obtained from fitting MW to number of targets. There is no way of knowing the exact extent to which transforming the data has inflated correlation because intra-bin variation has been hidden. The results presented in Table 1 show that r^2 values obtained for fitting primary and transformed data cannot be assumed to be equal. Since the article does not actually provide a correlation coefficient for promiscuity and MW, it is incorrect to state that:

> The observed correlation between promiscuity and molecular weight fits very well with the hypothesis proposed by Hann et al. [30] on the complexity of biologically active compounds, assuming molecular weight is a gross proxy for molecular complexity [21].

Promiscuity was defined in the LS2007 study [22] as the number of targets for which >30 % inhibition was observed at a concentration of 10 μM. Promiscuity was fit to median(ClogP) and the resulting r^2 value was 0.69 (a value of r of 0.83 is what was actually quoted). The LS2007 study asserts that:

> Lipophilicity plays a dominant role in promoting binding to unwanted drug targets. [22].

The meaning of the term 'dominant' in this context is not clear. Even if it is accepted that the definition of 'activity' in this study is physiologically relevant, it would still be necessary to show a strong correlation between ClogP and promiscuity (as opposed to median promiscuity) to be able to back the claim of dominance. The LS2007 study [22] does not actually present a value of r^2 for fitting promiscuity to ClogP and, as the results in Table 2 show, this cannot be assumed to equal to the value of r^2 for fitting promiscuity to median (ClogP). On this basis, it can be argued that the analysis presented in LS2007 [22] does not support the assertion of a dominant role for lipophilicity in promoting binding to unwanted drug targets. It is important to stress that the issue is simply what (if any) inferences can be drawn from this published analysis of the strength of the relationship between promiscuity and lipophilicity.

The physiological relevance of promiscuity as defined in both articles far from clear as will become apparent from inspection of the tabulation in reference 7 of free drug concentrations in plasma. It is instructive to compare the conclusions of both articles as to the importance of MW as a determinant of promiscuity. HMO2006 [21] showed that Mean(MW) decreased with number of targets for which $IC_{50} \leq 10$ μM with r^2 value of 0.93. In contrast, the analysis in LS2007 [22] shows promiscuity (number of targets for which >30 % inhibition was observed at ≤ 10 μM) increasing with median (MW) although the relationship is much weaker ($r^2 = 0.21$; $r = 0.46$) than that reported in HMO2006 [21]. That two studies addressing the same question could come to such divergent conclusions raises additional questions about the data analysis in at least one of the studies.

Sometimes data is simply plotted without actually quoting correlation coefficients and it is left to the reader to judge how strongly the relevant quantities are correlated. If individuals conclude that trends are stronger in the binned data than in the primary data, the binning can be described as inflating correlation even though the degree of correlation is not explicitly quantified. The LBH2009 study [23]

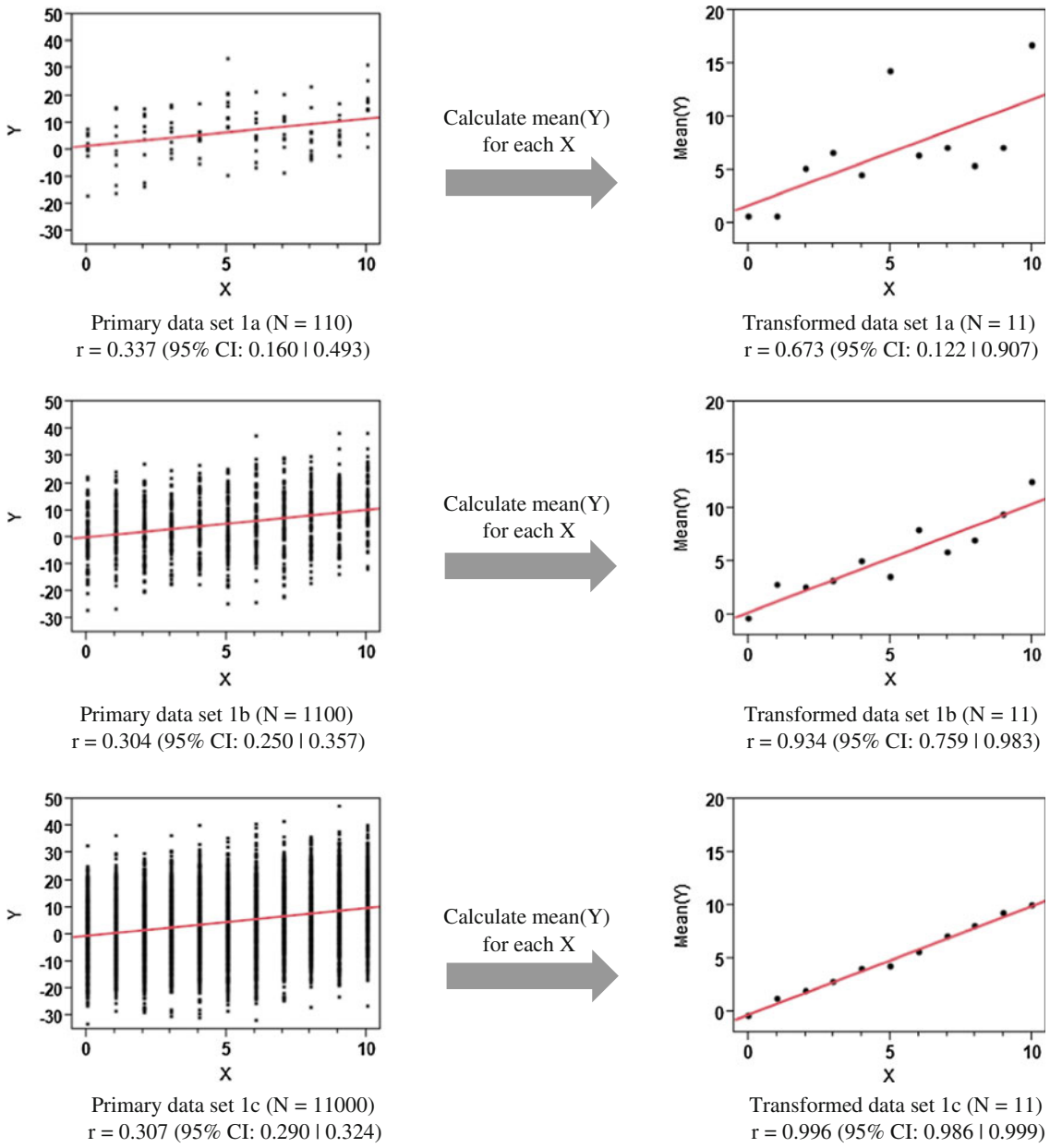


Fig. 4 Correlation inflation resulting from averaging values of Y for each value of X in the data sets 1. *Lines* (Y = A + BX) have been fit to data sets to provide a visual reference

Table 2 ANOVA for data sets 2

Data set	N	Bins	Degrees of freedom	F	*P*
2a	40	4	3	0.2596	0.8540
2b	400	4	3	12.855	<0.0001
2c	4,000	4	3	115.35	<0.0001
2c	4,000	2	1	270.91	<0.0001
2c	4,000	8	7	50.075	<0.0001

introduced carbon bond saturation as defined by fraction sp^3 (Fsp^3) as a molecular descriptor. The logarithm of solubility (log S) was partitioned into bins and average (probably mean rather than median) values of Fsp^3 for each log S bin were presented in bar chart format with an indication of significance of differences in average values for adjacent bins. No indication of the spread about these average values is given and this was also the case for comparisons of mean values of Fsp^3 for compounds in different stages of development. Although the chart is captioned *"Fsp^3 as a function of log S"* [23], it is actually average(Fsp^3) that is shown as a function of log S and the relationship these quantities appears to be very strong. Unusually for a study of this type, publicly available data [15, 16] was used in the analysis so it is possible to

quantify the extent to which binning data inflates correlation (Fig. 5). The correlation coefficient (r = 0.972; 95 % confidence interval for r: 0.846 | 0.995) for the binned data confirms the visual impression of a very strong relationship between log S and mean (Fsp^3). However the correlation coefficient for (r = 0.247; 95 % confidence interval for r: 0.193 | 0.299) for the primary data shows a much weaker relationship between log S and Fsp^3, indicating that the binning has inflated correlation. The correlation of log S with MW (r = −0.620; 95 % confidence interval for r: −0.653 | −0.584) is actually stronger than with Fsp^3. These observations call into question the value of Fsp^3, which explains just 6 % of the variance in log S, as a predictor of aqueous solubility.

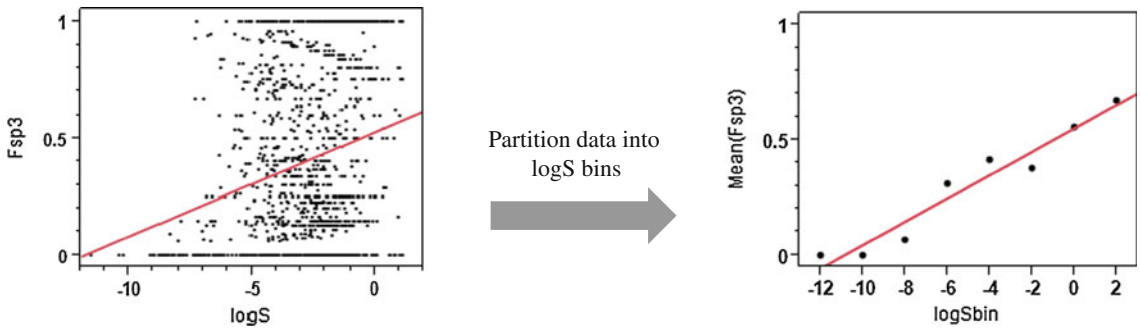
While the data analysis in LBH2009 [23] has been criticized, the suggestion that we need to look beyond aromatic rings is still valuable, although the rationale for doing so is based more on molecular recognition and diversity than aqueous solubility. One limitation of aromatic rings as components of drug molecules is that some regions above and below the plane defined by the atomic nuclear positions are not directly accessible to substituents. Molecular recognition considerations suggest a focus on achieving axial substitution in saturated rings with minimal steric footprint, for example by exploiting the anomeric effect or by substituting N-acylated cyclic amines at C2. Atoms have been known [24] for some years to behave as if they have volume and it would be more correct to say that benzene has constant or uniform thickness rather than describing it as flat or 2-dimensional.

Masking variation with standard error

The measure of promiscuity used in the HMO2006 [21] and LS2007 [22] studies is number of targets for which a specified level of activity was observed and analysis was performed for mean or median descriptor values for each number of targets. Although it is necessarily an integer, number of targets can be described as a continuous variable in the context of those studies because it can adopt many values and linear regression was used for the analysis. Different analyses, treating promiscuity as a categorical variable, could have also been performed. These data-analytic methods focus on differences observed between categories for the distributions of quantities of interest. For example, one might ask how aqueous solubility of carboxylic acids compares with the corresponding primary amides or whether pIC_{50} values observed for a set of compounds are more variable in assay A than in assay B.

Depending on the nature of subsequent analysis, binning can have the effect of transforming a continuous variable into a categorical one. The statistical analysis of the relationship between Fsp^3 and log S in LBH2009 [23] consisted of showing that the differences in average (Fsp^3) between adjacent bins were significant and this indicates that the log S bins were treated as categories for statistical analysis. When continuous variables are converted to categorical ones for statistical analysis, the ordering of the binned variables is lost. While the bins can still be placed in whatever order is desired for graphical display, statistical analysis that treats the bins as categories does not capture this information.

The issues associated with converting continuous variables into categorical ones can be illustrated more clearly with reference to the G2008 study [25]. The statistical analysis described in this study consisted of analysis of variance (ANOVA) and the continuous variables ClogP and MW were partitioned into three and four bins respectively. The question addressed by ANOVA is whether mean values of quantities are significantly different for different categories. Specifically ANOVA tests the null hypothesis that the mean values for all the categories are the same and the results are presented as the F-ratio and associated probability. Table 2 lists ANOVA results for



Primary data set (N = 1202)
r = 0.247 (95% CI: 0.193 | 0.299)

Transformed data set (N = 8)
r = 0.972 (95% CI: 0.846 | 0.995)

Fig. 5 Correlation inflation resulting from binning Fsp^3/solubility data. Data was binned as described in ref 23 and bin mid-points were used for the analysis and plot of the binned data. *Lines* (Y = A + BX) have been fit to data sets to provide a visual reference

data sets 2a, 2b and 2c and, as would be expected, the differences in mean(Y) values are found to be most significant for the largest data set (2c). It must be stressed that the ANOVA itself should not be regarded as inflating correlation because differences do indeed become more significant as larger samples are drawn from the relevant population. However, to interpret F-ratio as a measure of the strength (as opposed the significance) of a trend would be to inflate correlation. The ANOVA for data set 2c shows how statistics calculated from binned continuous data are in general dependent on the scheme (e.g. number of bins) used to partition the primary data. All data points used in the G2008 study were reported [25] to be in-range and this does weaken the case for binning.

Relationships between ADMET properties and the binned properties were presented in the G2008 study as ANOVA graphs in which mean values and associated 95 % confidence intervals were plotted for the bins. Similar plots can be generated for data sets 2a, 2b and 2c and these are shown in Fig. 6. The G2008 study asserts that:

> In each plot provided, the width of the errors bars and the difference in the mean values of the different categories are indicative of the strength of the relationship between the parameters. [25].

Whilst the correlations between X and Y in data sets 2a, 2b and 2c do not differ significantly, the confidence intervals for the mean values of Y decrease with size of the data set. If plots like these are interpreted as suggested in G2008 then the relationships between X and Y in data sets 2a, 2b and 2c would appear to become stronger as the size of the data set is increased. Either the assertion is incorrect or this presentation of the data has inflated correlation. Either way, it can be argued that these ANOVA plots do not accurately represent the degree of correlation between X and Y. This apparent paradox illustrates why one needs to think in terms of the effect size [2] (mean differences scaled by standard deviation) rather than significance (mean differences scaled by standard error).

The end point of the G2008 analysis is *"a set of simple interpretable ADMET rules of thumb"* [25] and it is instructive to examine these more closely. Two classifications (ClogP < 4 and MW < 400 Da; ClogP > 4 or MW > 400 Da) were created and these were combined with the four ionization state classifications to define eight classes of compound. Each combination of ADMET property and compound class was labeled according to whether the mean value of the ADMET property was lower than, higher than or not significantly different from the average for all compounds. Although the rules of thumb are indeed simple, it is not clear how useful they are in drug discovery. Firstly, the rules only say whether or not differences are significant and not how large they are. Secondly, the rules are irrelevant if the compounds of interest are all in the same class. Thirdly, the rules predict abrupt changes in ADMET properties going from one class to another. For example, the rules predict significantly different aqueous solubility for two neutral compounds with MW of 399 and 401 Da, provided that their ClogP values do not exceed 4. It is instructive to consider how the rules might have differed had values of logP and MW of 5 and 500 Da (or 3 and 300 Da) had been used to define them instead of 4 and 400 Da.

When looking at plots like these, there is a danger of thinking that most of the values of Y in a particular X-bin lie within the 95 % confidence interval for the mean. This creates the impression that distributions of Y for different bins are well separated even when they overlap to a significant extent. It should be stressed that no such suggestion is made in the G2008 study and in some cases confidence intervals for mean values are actually narrower than likely uncertainty in the assay. The TNK2012 study [26] uses a plot of mean values of two ligand efficiency metrics with their associated standard errors to assert separation between different groups of compounds:

> Pairwise comparison of corresponding LLE and LELP values separates leads, successful leads from P2 compounds, and marketed drugs (Fig. 3), suggesting that LLE- and LELP-based assessment of compound quality directs discovery programs toward the desirable drug space [26].

Since the standard error in the mean decreases with size of the sample drawn from a population it is incorrect to use standard error to assert that property distributions for two or more groups of compounds are separated. Using standard error to quantify separation implies that distributions become more separated as larger samples are drawn from them. Standard error helps to inform us whether or not we can base decisions on differences between mean values even when distributions overlap.

Correlation and graphical data representation of data

Drug discovery can be described as multivariate pursuit and it is common to see results of ADMET assays presented graphically. When done appropriately, this can reveal relationships between compounds and other patterns in the data:

> Human beings have an innate ability to quickly recognize and assimilate shapes, patterns and colours [27].

The flip side of this pattern recognition ability is that the humans will also see patterns in what is essentially random noise. For example, an element in a pie chart array

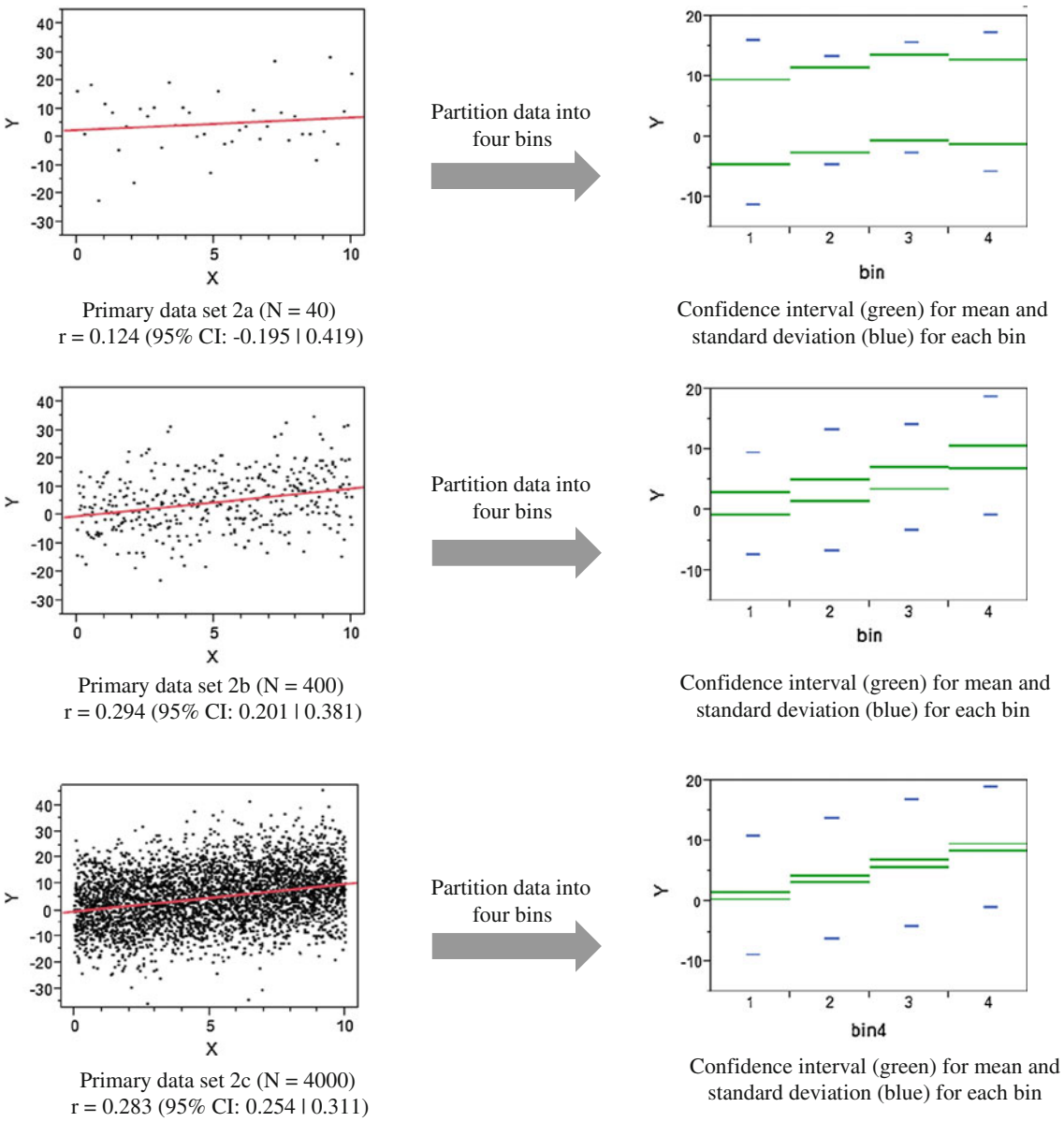


Fig. 6 Correlation coefficients for X and Y do not differ significantly for data sets 2a, 2b, 2c. Interpreting the range of the confidence interval for the mean as indicative of the strength of correlation leads to the erroneous conclusion that correlation between X and Y increases with the size of the primary data set. *Lines* (Y = A + BX) have been fit to data sets to provide a visual reference

representing a single compound may have greater visual impact than an element representing a thousand compounds. Graphical representation of ADMET data usually involves transforming continuous assay data into categorical or ranked (e.g. high, medium or low) data and the cut off values used for categorization or ranking will in general affect perception of the data. It is often assumed that limits of acceptability for one parameter (e.g. aqueous solubility) can be set independently of the values of another parameter (e.g. Caco-2 permeability) although this assumption would appear to be at odds with a multiobjective [8] view of lead optimization. The strength of graphical representations of ADMET data lies in their ability to reveal multivariate relationships between compounds but it is not appropriate to use them for quantifying the degree to which pairs of quantities are correlated. For example, how do we determine objectively whether the relationship shown by one pie chart array is stronger than that shown by another one?

We will focus on the HY2010 study [28] to illustrate problems associated with using graphical data representations to support statements about strengths of relationships between quantities. Both this study and the earlier RM2009 study [29] assert the importance of number of aromatic rings (#Ar) as a determinant of the ease with which a

compound can be developed. The HY2010 study used a bar chart format to show how the relative proportions of compounds in three solubility categories varied in response to (binned) lipophilicity and Solubility Forecast Index (SFI = #Ar + c $\log D_{pH7.4}$). Figure 7 shows the data sets 1a, 1b and 1c presented in this manner and it will be recalled that the correlations between X and Y do not differ significantly for these data sets. Can this be inferred from the bar chart representation of the data in Fig. 7?. If it is believed, for example, that the bar chart representations of data sets 1b and 1c indicate that the correlation between X and Y is stronger in the latter than in the former then it would be accurate to state that correlation inflation has occurred. The difficulties in using this data representation to quantify, or even rank, strengths of relationships are illustrated by the following statements in the HY2010 study:

> The clearer stepped differentiation within the bands is apparent when log $D_{pH7.4}$ rather than log P is used, which reflects the considerable contribution of ionization to solubility [28].
>
> This graded bar graph (Figure 9) can be compared with that shown in Figure 6b to show an increase in

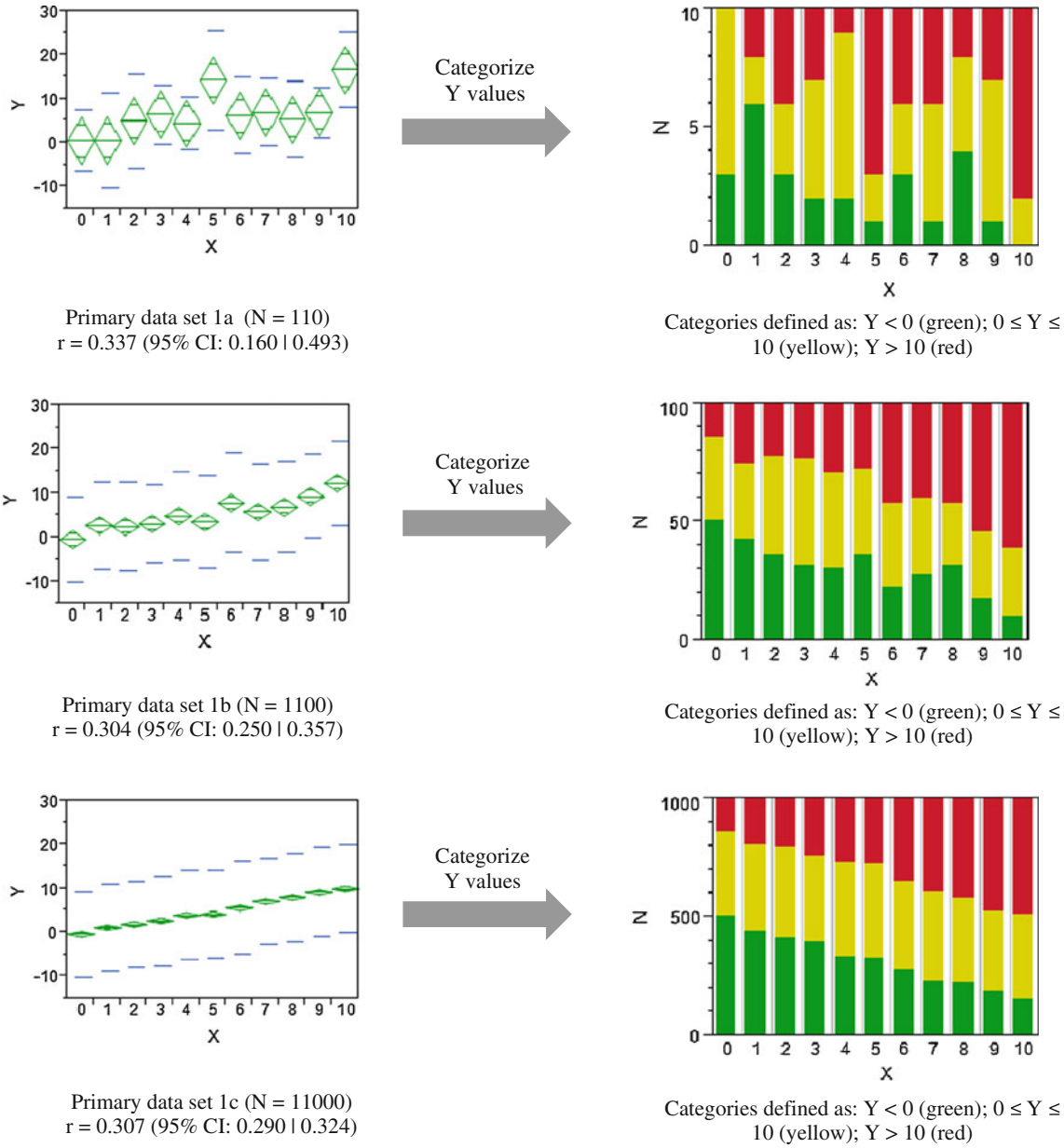


Fig. 7 Correlation coefficients for X and Y do not differ significantly for data sets 1a, 1b and 1c. Can this be inferred from the bar chart representation?

> resolution when considering binned SFI versus binned c log $D_{pH7.4}$ alone [28].

It is not clear why *"stepped differentiation"* was used to make one of the comparisons while *"resolution"* was used to make the other one and the definitions of both terms remain obscure in this context. The study does not address the key question of how much of the variance in measured solubility is explained by the different predictors of solubility. The definition of SFI is, in essence, a classification problem and a potential user of this metric could legitimately ask if #Ar + 0.8 × c $logD_{pH7.4}$ would have been a better predictor of solubility than SFI. Might not a predictor based on lipophilicity and MW (or any other measure of molecular size) be more effective than one based on lipophilicity and number of aromatic rings? Although both the HY2010 and MR2009 studies explore the relationship between aromatic ring count and lipophilicity, neither study appears to have examined the extent to which aromatic ring count is correlated with MW. The observation that MW explains more of the variance in log S in the data [15, 16] used in in the LBH2009 study [23] than does Fsp^3 suggests that correlation between MW and #Ar needs to be explored. Disregarding the influence of MW and other measures of molecular size in analyses like these is to risk equating correlation with causation by stating:

> These solubility differences are probably because of a combination of factors: increased molecular rigidity, melting point phenomena and the capacity for π–π stacking with increased aromaticity inter alia [29].

Both the MR2009 and HY2010 studies note the simplicity of the relationships that the analysis has revealed. Given that drug discovery would appear to be anything but simple, the simplicity of a drug-likeness model could actually be taken as evidence for its irrelevance to drug discovery. The number of aromatic rings in a molecule can be reduced by eliminating rings or by eliminating aromaticity and the two cases appear to be treated as equivalent in both the MR2009 and HY2010 studies. Using the mnemonic suggested in MR2009 [29] one might expect to make a compound more developable by replacing a benzene ring with cyclohexadiene or benzoquinone.

Conclusions

Although this study has highlighted examples of correlation inflation, it should be stressed that there is no suggestion that properties such as lipophilicity, molecular size or aromatic character are unimportant in drug discovery. However, it is still important to know strong the relevant correlations are because weaker correlations with individual descriptors strengthen the case for using multivariate techniques. This article has shown how averaging groups of data points prior to analysis can make trends in data appear to be stronger than they actually are. Binning continuous data reduces options for quantifying correlation and one should have a sound rationale, such as inclusion of out-of-range data, for transforming data in this manner. When continuous data is binned, the burden of proof is entirely on the data analyst to demonstrate that transformation of the primary data has not exaggerated trends. The extent to which correlation is inflated tends to increase with size of the data set and it can be useful to confirm that the strength of a trend observed in the transformed data is independent of the size of sample drawn from the primary data. Trends in data should be presented transparently and hiding or masking variation always reduces transparency. In the context of correlation, statistical significance should be seen a necessary, but not sufficient, requirement for relevance.

One reviewer of the manuscript noted *the "problem of analyzing unruly data"* and suggested that *we "elaborate on practical methods that drug discovery scientists use to tame the uncertainty"*. This is a good challenge and, in attempting to respond to it, we first re-iterate the need for data analysts to focus on effect size and not become ensnared by statistical significance. However, one must also remember that while statistical analysis can quantify uncertainty, it cannot make it go away. If relationships between outcomes and descriptors are too weak and noisy to allow useful prediction, the brutal reality is that one will need better descriptors and different prediction paradigms. For example, drug discovery scientists might look beyond the octanol/water partitioning system and consider modeling activity and properties in terms of relationships between structures [30]. Also, correlations between ADMET characteristics and properties like lipophilicity may well be stronger within structural series. Would it not be more useful to learn how to better perceive strong local correlations than to look for new ways to inflate weak global ones? What can the compounds that beat trends teach us? Might we not learn more about drug-likeness by trying to understand what distinguishes a drug from the compounds synthesized in the search for it? Should we still worry about lipophilicity and molecular size when ADMET assay results indicate that the profile of a compound is acceptable? Just how useful is it to treat marketed drugs as a separate and contiguous region of chemical space?

Despite the relatively narrow focus of this article, some of these approaches to analysis and presentation of data will be encountered elsewhere in the literature of drug-likeness (and drug discovery in general) and it is hoped that readers will now be better equipped to recognize correlation inflation when presented with it in its various guises. We hope that our study will stimulate debate and provide encouragement for

those who make the challenges to institutional wisdom that are essential for innovation. One key message of this article is that one needs to examine data analytic methodology carefully before accepting the conclusions drawn from the analysis or acting on the prescriptions of those who have performed it. Given a recent call to discipline in medicinal chemistry [31], it is perhaps appropriate to make a call to discipline in the data analysis that forms much of the basis of drug-likeness. The call is made not only to those who analyze data but also to those who review their manuscripts. The editors of the Journal of Medicinal Chemistry already provide guidelines for computational chemistry studies [32] and they, with editors of other journals, might consider providing similar guidelines for data analysis. We conclude our study with some suggestions:

1. Data sets should be made available as supplemental material.
2. A purely graphical representation of data is inadmissible as evidence that a relationship between one pair of variables is stronger than that between another pair of variables.
3. Provided that all data are in-range, a correlation coefficient for X and Y or coefficient of determination for the fit of Y to X should always be presented to support any assertion of a strong relationship between X and Y.
4. For data sets partitioned into bins, observation of strong relationship between average values of Y and X is inadmissible as evidence of a strong relationship between X and Y.
5. For data sets partitioned into bins, each average value should be accompanied by a measure (e.g. standard deviation; inter-quartile range) of the spread of the distribution that is independent of sample size.
6. For data sets partitioned into bins, it should be demonstrated that inferences drawn from analysis are independent of the binning scheme.

Acknowledgments We thank Anthony Nicholls for valuable advice and the reviewers of the manuscript for their helpful and constructive feedback. We are grateful to Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) and Conselho Nacional de Pesquisa (CNPq) for financial support and OpenEye Scientific Software for an academic software license.

References

1. Ziliak ST, McCloskey DN (2008) The cult of statistical significance: How the standard error costs us jobs, justice and lives. University of Michigan Press, Ann Arbor
2. Kelley K, Preacher KJ (2012) On effect size. Psychol Methods 17:137–152
3. Hajduk PJ, Huth JR, Fesik SW (2005) Druggability indices for protein targets derived from NMR-based screening data. J Med Chem 48:2518–2525
4. Lipinski CA, Lombardo F, Dominy BW, Feeney PJ (1997) Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. Adv Drug Deliv Rev 23:3–25
5. Abraham MH, Chadha HS, Whiting GS, Mitchell RC (1994) Hydrogen bonding. 32. An analysis of water-octanol and water-alkane partitioning and the Δlog P parameter of Seiler. J Pharm Sci 83:1085–1100
6. Colclough N, Hunter A, Kenny PW, Kittlety RS, Lobedan L, Tam KY, Timms MA (2008) High throughput solubility determination with application to selection of compounds for fragment screening. Bioorg Med Chem 16:6611–6616
7. Smith DA, Di L, Kerns EH (2010) The effect of plasma protein binding on in vivo efficacy: misconceptions in drug discovery. Nat Rev Drug Discov 9:929–939
8. Ekins S, Honeycutt JD, Metz JT (2010) Multiobjective optimization for drug discovery. In: Abraham DJ, Rotella DP (eds) Burger's medicinal chemistry, drug discovery and development, 7th edn. Wiley, New York
9. Hopkins AL, Groom CR, Alex A (2004) Ligand efficiency: a useful metric for lead selection. Drug Discov Today 9:430–431
10. van de Waterbeemd H, Smith DA, Jones BC (2001) Lipophilicity in PK design: methyl, ethyl, futile…. J Comput-Aided Mol Des 15:273–286
11. Hann MM, Leach AR, Harper G (2001) Molecular complexity and its impact on the probability of finding leads for drug discovery. J Chem Inf Comp Sci 41:856–864
12. Woltosz WS (2012) If we designed airplanes like we design drugs. J Comput-Aided Mol Des 26:159–163
13. Kenny PW (2009) Hydrogen bonding, electrostatic potential and molecular design. J Chem Inf Model 49:1234–1244
14. Rodgers JL, Nicewander WA (1988) Thirteen ways to look at the correlation coefficient. Am Stat 42:59–66
15. Hou TJ, Xia K, Zhang W, Xu XJ (2004) ADME evaluation in drug discovery. 4. prediction of aqueous solubility based on atom contribution approach. J Chem Inf Comp Sci 44:266–275
16. ADME/T prediction models and databases. http://modem.ucsd.edu/adme/databases/databases_logS.htm. Accessed 15 Oct 2012
17. LOGKOW, A databank of evaluated octanol-water partition coefficients. http://logkow.cisti.nrc.ca/logkow/index.jsp. Accessed 26 Oct 2012
18. OEChem Toolkit Manual, OpenEye Scientific Software, Santa Fe, NM 87508. http://www.eyesopen.com/docs/toolkits/current/html/OEChem_TK-c++/index.html. Accessed 26 Oct 2012
19. SMARTS Theory Manual, Daylight Chemical Information Systems, Inc., Laguna Niguel, CA 92677. http://www.daylight.com/dayhtml/doc/theory/theory.smarts.html. Accessed 16 Dec 2012
20. JMP version 10.0.0, SAS Institute, Cary, NC 27513. http://www.jmp.com. Accessed 16 Dec 2012
21. Hopkins AL, Mason JS, Overington JP (2006) Can we rationally design promiscuous drugs? Curr Opin Struct Biol 16:127–136
22. Leeson PD, Springthorpe B (2007) The influence of drug-like concepts on decision-making in medicinal chemistry. Nat Rev Drug Discov 6:881–890
23. Lovering F, Bikker J, Humblet C (2009) Escape from flatland: increasing saturation as an approach to improving clinical success. J Med Chem 52:6752–6756
24. Maxwell JC (1874) Van der Waals on the continuity of gaseous and liquid states. Nature 10:477–480
25. Gleeson MP (2008) Generation of a set of simple, interpretable ADMET rules of thumb. J Med Chem 51:817–834
26. Tarcsay A, Kinga N, Keserű GM (2012) Impact of lipophilic efficiency on compound quality. J Med Chem 55:1252–1260

27. Ritchie TJ, Ertl P, Lewis R (2011) The graphical representation of ADME-related molecule properties for medicinal chemists. Drug Discov Today 16:65–72
28. Hill AP, Young RJ (2010) Getting physical in drug discovery: a contemporary perspective on solubility and hydrophobicity. Drug Discov Today 15:648–655
29. Ritchie TJ, MacDonald SJF (2009) The impact of aromatic ring count on compound developability: are too many aromatic rings a liability in drug design? Drug Discov Today 14:1011–1020
30. Kenny PW (2012) Computation, experiment and molecular design. J Comput-Aided Mol Des 26:69–72
31. Johnstone C (2012) Medicinal chemistry matters—a call for discipline in our discipline. Drug Discov Today 17:538–543
32. Stahl M, Bajorath J (2011) Computational medicinal chemistry. J Med Chem 54:1–2