# Simple Size-Independent Measure of Ligand Efficiency

J. Willem M. Nissink*

Computational Chemistry, Astrazeneca, 50S27 Mereside, Alderley Park, SK10 4TG, United Kingdom

Received March 10, 2009

This paper suggests a functional form for a size-independent ligand efficiency (LE) measure of the form SILE = affinity/$N^{0.3}$, where $N$ is the number of heavy atoms. The size-dependency effect of the commonly used LE measure (LE = affinity/$N$) is apparent in experimental settings like hit-to-lead optimization, but equally well applies to in-silico binding-energy-based estimates. It is therefore expected that the correction is useful in experimental drug optimization, as well as in in-silico applications like protein−ligand docking.

## INTRODUCTION

The dependency of the commonly used ligand efficiency (LE) measure on molecule size has been noted by many workers in the field,[1,2] and recently, corrections have been suggested by Reynolds et al.[1] Their work, based on a large set of ligand $K_i$ and $IC_{50}$ data across a range of targets, proposes an empirical, size-independent ligand efficiency score that can be used to track progress during the drug-design optimization process. In this short communication, we propose a simpler size-independent ligand efficiency measure, and explain a rationale for its functional form.

## THEORY

**Dependency of LE on Heavy Atom Count.** The LE measure as proposed in the literature and used for optimizing compounds from hit-finding to lead-optimization stage is defined as[2]

$$\mathrm{LE} = \frac{\text{affinity}}{N} \tag{1a}$$

where $N$ denotes the number of heavy atoms in the molecule. Affinity is generally approximated by $RT\ln IC_{50}$, $pIC_{50}$, $pK_i$, or experimental $\Delta G$.[1,2] It is well-known that the LE depends strongly on molecule size.[2] The measure tends to decrease with increasing number of atoms, something which can be observed particularly well for small molecules. This has implications for application of eq 1a in drug design and fragment-based drug design in particular, as drug size tends to increase during hit-to-lead optimization.

Reynolds et al.[1] recently proposed a size-independency correction of LE using a generic polynomial form. For this purpose, they extracted $IC_{50}$ and $K_i$ data from a public database and fitted a curve through those observations with *maximum* LE, as a function of size. Our analysis of the curve shape of the maximum-LE data reveals that it follows the form

$$\mathrm{LE}_{max} = \frac{\text{optimal affinity}}{N} \propto \frac{1}{N^x} \tag{1b}$$

closely. Taking into account this size-dependency of LE, we arrive at the following form for a simple size-independent ligand efficiency measure

$$\mathrm{SILE} = \frac{\text{affinity}}{N^{1-x}} \tag{2}$$

**Value of Exponent Term 1 − *x* in eq 2.** With $LE_{max}$ showing a dependency on $N$ of the form $LE_{max} = f(N) = a/N^x$, we can obtain an estimate of $x$ (and scaling factor $a$) by fitting maximum-observed LE values across a range of $N$. A large set of $K_i$ values was collated from the BindingDB database[3,4] for this purpose. The following functional form fits this data well ($R^2 = 0.995$, based on a set of 6945 data points; Figure 1).

$$\ln(\mathrm{LE}_{max}) = 1.40 - 0.73 \ln(N) \tag{3a}$$

which can be rewritten as

$$\mathrm{LE}_{max} = e^{1.40}/N^{0.73} \tag{3b}$$

This yields a value for $x$ of 0.73, and combination of eqs 1, 2, and 3b yields

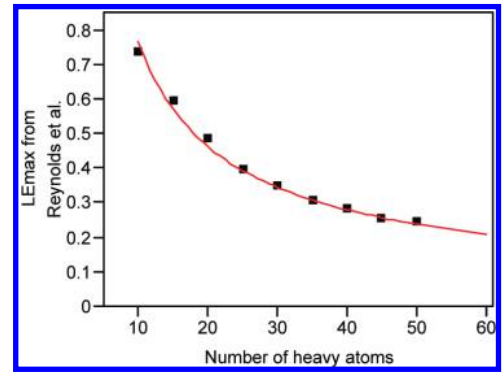


**Figure 1.** Selected maximum-LE data points (based on $pK_i$) taken from Figure 7 in ref 1 and fitted with the functional form from eq 3b (curve).

* To whom correspondence should be addressed. E-mail: willem.nissink@astrazeneca.com.

10.1021/ci900094m CCC: $40.75 
Published on Web 05/13/2009

$$\mathrm{SILE}_{\max} = \mathrm{LE}_{\max} \cdot N^{0.73} = \frac{\mathrm{p}K_{i\max}}{N^{0.27}} \approx 4.1 \qquad (4)$$

The value of 4.1 (i.e., $e^{1.40}$) for $\mathrm{SILE}_{\max}$ follows from the fitted function shown in eq 3b. A similar fitting of $\mathrm{pIC}_{50}$-based ligand efficiency values (analogous to Figure 3 in ref 1, using 16 384 data points) yields a similar fit ($R^2 = 0.996$, eqs 5a and 5b), and corresponding estimate of $1 - x$=0.21

$$\mathrm{LE}_{\max} = e^{1.54}/N^{0.79} \qquad (5a)$$

which implies

$$\mathrm{SILE}_{\max} = \frac{\mathrm{pIC}_{50_{\max}}}{N^{0.21}} \approx 4.7 \qquad (5b)$$

The maximum SILE value of 4.7 obtained for the $\mathrm{pIC}_{50}$-based approximation is, again, derived from the fit parameters (eq 5a). It should be kept in mind that these maximum values are approximations only, and are limited by the simple mathematical form that is assumed. Still, the values obtained are reasonably similar for both the $\mathrm{p}K_i$ and $\mathrm{pIC}_{50}$ data sets. Above results suggest a value of $1 - x$ in the range of 0.3 to 0.2 for use with eq 2 to produce a size-independent ligand efficiency value.

**What is the Expected Dependency of LE on *N*?** Several workers in the field have remarked on the correspondence between the solvent-accessible-surface area per atom of bound ligands and LE curve shapes (e.g., Kuntz[2] and Reynolds[1]). This makes sense intuitively but is surprising nonetheless. Considering the affinity of a ligand binding to a protein, both enthalpic and entropic terms will contribute. The enthalpic contribution to affinity is largely proportional to the ligand-to-protein contact area, to a first approximation. This principle is applied in the various protein−ligand docking scores, where, for example, short-range Lennard-Jones-type energy contributions dominate over long-range interactions. The enthalpic contribution to LE is therefore expected to be roughly proportional to area-per-atom.

Yet, the nature of the entropic contribution to the protein−ligand affinity is much more complex, and results from both changes in ligand and protein conformation, as well as water displacement and rearrangement. Assuming that for *optimal* inhibitors, any protein and ligand conformational changes are relatively small upon binding, the largest part of the entropic contribution would derive from water reorganization in these cases. This contribution will be both proportional to ligand volume (bulk water displacement) and, again, contact area, because of displacement or rearrangement of waters at the interaction surface. The entropic contribution is therefore expected to vary with both ligand area, and ligand volume.

As a result of the above, optimal affinity, and hence, optimal LE will correlate with both ligand volume and protein−ligand contact area

$$\text{optimal affinity} \propto \text{volume} + s \times \text{area} \qquad (6)$$

where $s$ is a scale factor. A more detailed analysis that suggests eq 6 holds in first approximation is detailed in the Appendix. The number of heavy atoms $N$ is observed to be linearly proportional to ligand volume (Figure 2a). Therefore, the optimal ligand efficiency value $\mathrm{LE}_{\max}$ can be expressed

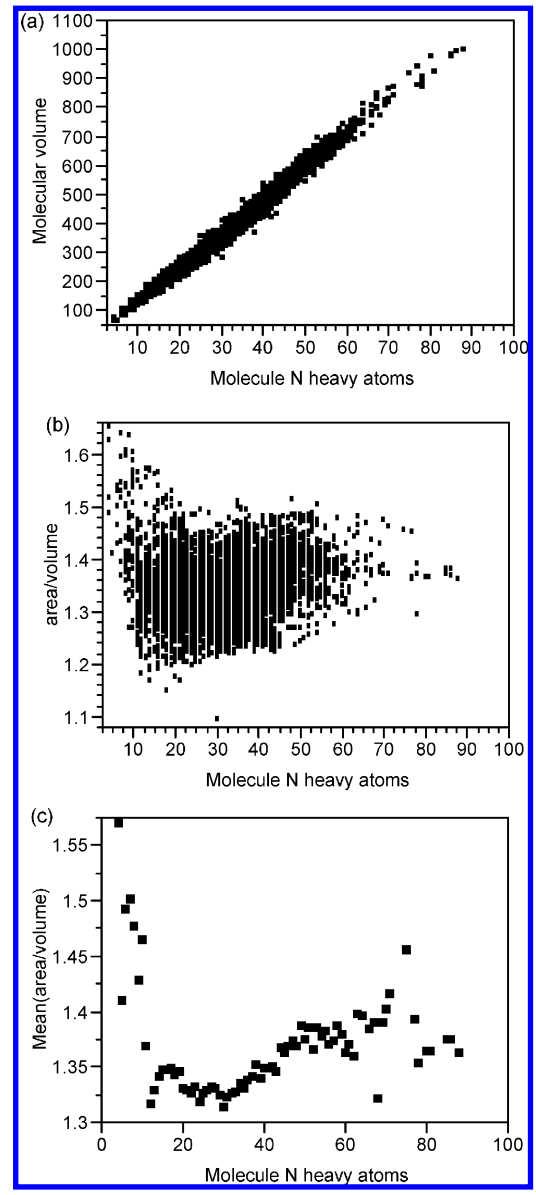


**Figure 2.** (a) Molecular volume as function of number of heavy atoms for a set of ligands from BindingDB (as in Figure 1); (b) molecule area/molecular volume; (c) mean area divided by volume averaged per size. Area and volume were calculated analogous to the method described in ref 14.

as follows (using eqs 1 and 6 and $N \propto$ volume; $s$ is the scale factor from eq 6)

$$\mathrm{LE}_{\max} = \frac{\text{optimal affinity}}{\text{volume}} \propto 1 + s \times \frac{\text{area}}{\text{volume}} \propto \frac{\text{area}}{\text{volume}} \quad \text{for } s \gg 1 \propto \frac{\text{area}}{N} \qquad (7)$$

that is, the ligand efficiency value of eq 1 is expected to display a trend in line with the ligand's area-per-atom, in approximation. A very simple spherical approximation of ligand size would suggest a $1/N^{1/3}$proportionality for the area-to-volume ratio. Area-to-volume ratios calculated for a large set of molecules using 2D descriptors are shown in Figure 2b (all data) and c (averaged per ligand size). These areas and volumes depend on interaction radii chosen for the atoms

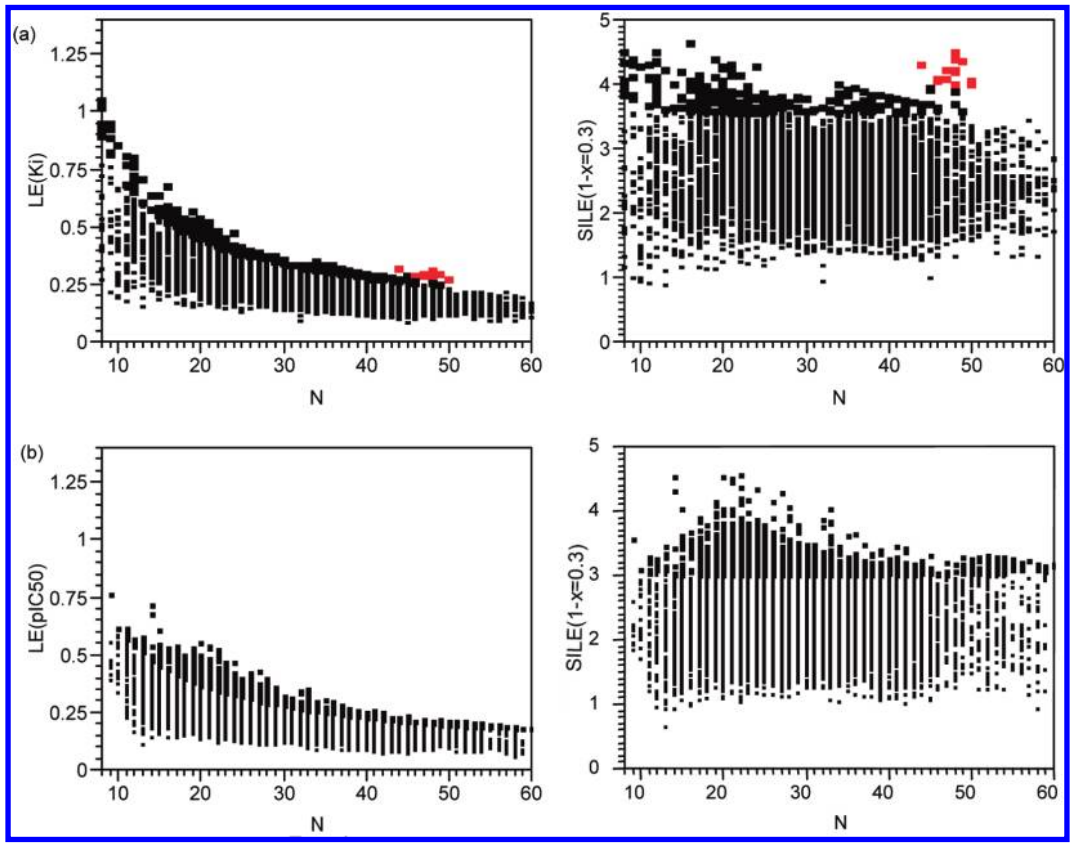


**Figure 3.** Ligand efficiencies and corrected ligand efficiencies for data taken from the BindingDB database[2] (both $K_i$ and $IC_{50}$ data), as function of number of heavy atoms (N). Left: LE (eq 1); right: corresponding SILE values (eq 2, $1 - x = 0.3$). (a) $K_i$-based ligand efficiencies (6945 data points). Outlier compounds in red are all from an arylsulfonamide HIV-1 inhibitor series with subpicomolar HIV-1 protease activities. All compounds in bold black have been selected as SILE ($1 - x = 0.3$) > 3.5. (b) $IC_{50}$-based ligand efficiencies (16 831 data points). Data points highlighted in bold have been selected as SILE ($1 - x = 0.3$) > 3.0. The SILE thresholds were chosen solely to provide a guide to the eye.

and are calculated over a conformational ensemble rather than a single binding mode, so they are approximate only. Figure 2b and c show that an area-overvolume term thus calculated does only partially reflect the behavior of LE as a function of *N*. We tend to see a large range of area-to-volume values for every size of molecule (Figure 2b), but interestingly, the trend shows that small molecules with less than ca. 25 atoms yield very different ratios on average (Figure 2c), indicating that shape for small molecules is very different from shape observed in larger ones. This is not unexpected: shapes of large, fairly rigid molecules are in part stick-like, and surface area of stick-like groups increases proportionally with the number of heavy atoms. Small molecules tend to be more globular or flat than large ones, and this is expressed through their high area-to-volume ratios. This observation suggests that the composition of enthalpic and entropic components to binding energy may be intrinsically different for small and large molecules, and it would explain the dependency of LE on molecular size.

## RESULTS AND DISCUSSION

**Application to $K_i$ and $IC_{50}$ Data.** LE and SILE ligand efficiencies were calculated for $K_i$ data and $IC_{50}$ data downloaded from the BindingDB database[2] (on Jan 12, 2009). The data cover a wide selection of targets, 147 for the $K_i$ data set, and 207 different targets for the $IC_{50}$ data set. Results are shown in Figure 3. Figure 4 shows LE and SILE values for data from the NCI AIDS antiviral cell screen (AIDS screen data May 2004) for a large set of compounds.[10] LE and SILE data were generated using both $EC_{50}$ data for protection of infected cells and IC50 data for growth inhibition of uninfected cells.

From a side-by-side comparison of LE and SILE distributions it is clear that the maximum-observed values for the latter (using $1 - x = 0.3$) are much less-dependent on ligand size (expressed as number of heavy atoms *N*) than the LE value calculated according to eq 1. Similar results were obtained using in-house affinity data (not shown). We note, however, that the assay setup can have an influence on the affinity value measured, and hence, on the SILE value, with maximum values observed in a range of 3−4. In Figure 3, we have highlighted a small set of highly active compounds in red ($K_i$ plot). These arylsulfonamide compounds originate from one publication[9] and exhibit picomolar activities measured in an assay specifically developed to track such potent compounds. As assay conditions may well have an influence on results, it is recommended that LE and SILE values for compounds are only compared for compounds measured in specific assays, though maximum-SILE results across a wide range of assays appear remarkably consistent across a wide range of *N*.

We note that for high-potency compounds both LE and SILE may be affected by an artificial ceiling to the measured $pIC_{50}$ (usually around 9, but depending on assay setup; in high-affinity cases e.g. the concentration of protein plays a role), in such cases, LE (and SILE) may artificially decrease

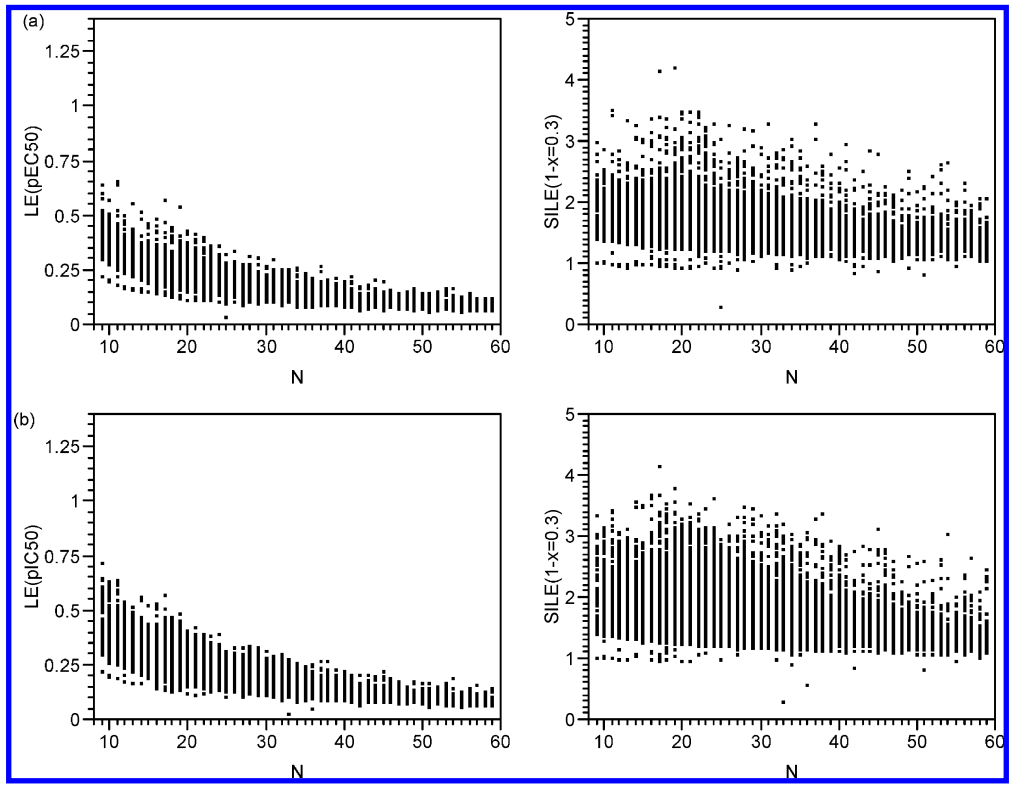


**Figure 4.** Ligand efficiencies and corrected ligand efficiencies for cell data taken from the NCI antiviral screen (AIDS; May 2004) database,[10] as function of number of heavy atoms (*N*). Left: LE (eq 1); right: corresponding SILE (a) based on $EC_{50}$ indicating 50% protection of infected cells and (b) based on IC50 indicating 50% growth inhibition of uninfected cells.

with size when $pIC_{50}$ reaches the maximum-observable value, where it does not reflect increased affinity anymore. We further note that the correction of the form shown in eq 2 is less prone to overfitting than the multiparameter polynomial factor suggested by Reynolds et al.[1] The parametrization of the latter will be more susceptible to outliers especially at the low end of the scale, which may not be desirable.

**Application to Fragment-Based Drug Design.** In Figure 5, the SILE metric and LE are compared side by side for six fragment-based design trajectories, where potent inhibitor compounds have been deconstructed into small, fragment-like start points. All examples shown are taken from recent literature.[11−13] Examples A−D constitute a retrospective analysis, and E, F, and G result from fragment-based design examples. The SILE values increase with size for all trajectories, which suggests that all additions to the preceding smaller compound yield a more ligand-efficient inhibitor in each step. This is not seen for the LE metric, where values tend to remain approximately the same or decrease slightly, which can make interpretation difficult (see, e.g., ref 5 for a discussion). It is surprising that the SILE increases in value for these trajectories with increasing ligand size, as it suggests that final ligand-efficiency of the optimized lead does not necessarily correlate with the initial ligand efficiency of the start point. One should keep in mind, however, that deconstruction of a potent compound into a hypothetical 'lead fragment' does not necessarily lead to the most optimal lead possible. Still, fragment-based design examples F and G in Figure 5 are cases where the SILE and LE values of the unoptimized starting fragments are moderate, and substantial gains are seen in subsequent optimizations of the fragment.

All except two of the trajectories end with final compounds displaying SILE values greater than 2.5. The actual value that is obtained will in part result from experimental settings of the various assays used being different. Still, the values of the final SILE are remarkably consistent, and provide a usable end-point for optimization.

**Application to Analysis of Compound Properties.** The SILE value offers an easy way to separate out those compounds with good ligand efficiency regardless of size, opening up the way to analyze properties of good binders across a wide range of molecular size. Interesting is, for instance, that a weak tendency is observed for the average number of rotational bonds for large good binders (SILE > 3.0; $N > 35$) to be slightly lower than those with lower ligand efficiencies in the $K_i$ data set (Figure 6). The effect is in line with expectation, as the change in entropy $\Delta S$ resulting from diminished ligand flexibility is expected to improve affinity when the latter is largely driven by enthalpy.

## CONCLUSIONS

Ligand efficiency (LE) as defined according to Kuntz et al.[3] has been observed to show a dependency on ligand size in numerous publications. This dependency poses problems when using an LE measure to guide the design of drugs, as size of the ligands is likely to increase during this process. Various schemes have been designed to correct for this effect (see, e.g., Verdonk and Rees[5] and Reynolds et al.[1]). Following recent publication of a scaled, size-independent ligand efficiency measure by Reynolds et al.,[1] we propose a different size-independent ligand efficiency of the form

$$\mathrm{SILE} = \frac{\mathrm{affinity}}{N^{0.3}} \tag{8}$$

and show that the size-dependency of the classic LE metric follows the area-per-atom trend.

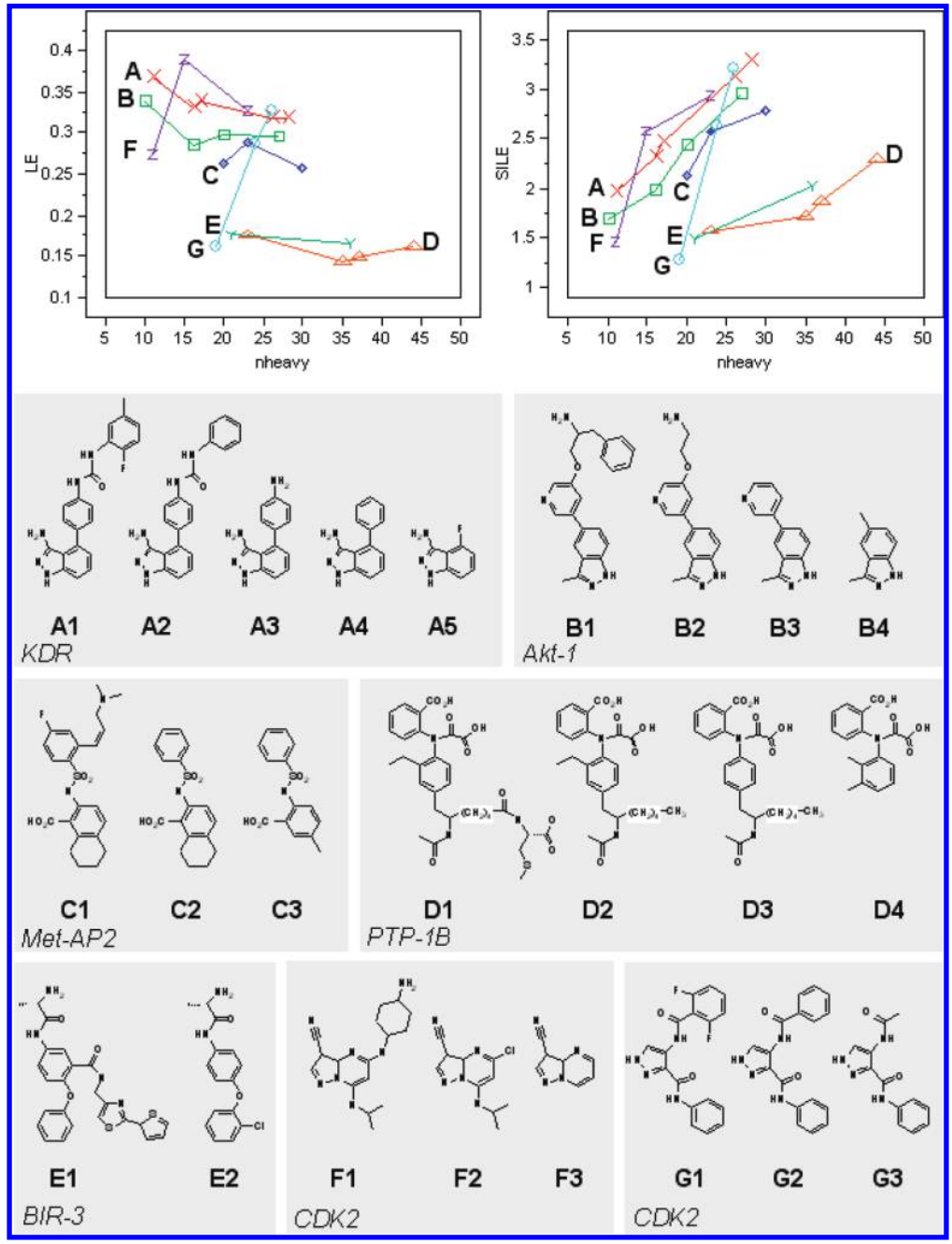


**Figure 5.** LE (top left) and SILE (top right) values for six examples of fragment-based design trajectories. Compounds are shown below. (A) KDR;[11] (B) Akt-1;[11] (C) MetAP2;[11] (D) PTP-1B;[11] (E) BIR-3;[12] (F) CDK2;[13] (G) CDK2.[13] LE and SILE values are based on p$K_d$ for examples A−E, and on p$IC_{50}$ for F and G.

The size-dependency effect is apparent in experimental settings like hit-to-lead optimization but equally well applies to in-silico binding-energy estimates. We show that the correction is useful in experimental hit-to-lead optimization settings, but it is also expected to be of use in applications like protein−ligand docking, where calculated affinities and scores tend to show a strong dependency on compound size. The SILE value offers a pragmatic way to identify compounds with high ligand efficiency across a wide range of ligand sizes, and track their progress during the drug-design cycle, for example in fragment-based drug design trajectories.

## ACKNOWLEDGMENT

The author thanks Ed J. Griffen, Peter W. Kenny, and Thorsten Nowak for useful discussion.

## APPENDIX

The dependency of enthalpic and entropic contributions on ligand area and volume can be rationalized as follows. Affinity is generally taken as the Gibbs energy change upon binding

$$\Delta G = \Delta H - T \times \Delta S \quad \text{(A1)}$$

The ligand-specific enthalpic contribution upon binding can be described as

$$\Delta H_{bind} = \Delta H_{conf} + \Delta h_{polar} \times \Delta ASA_{polar} + \Delta h_{apolar} \times \Delta ASA \quad \text{(A2)}$$

where $\Delta H_{conf}$ describes effects caused by conformational changes in protein and ligand; $\Delta ASA$ and $\Delta ASA_{polar}$ describe the change in contact area and polar contact area upon binding;

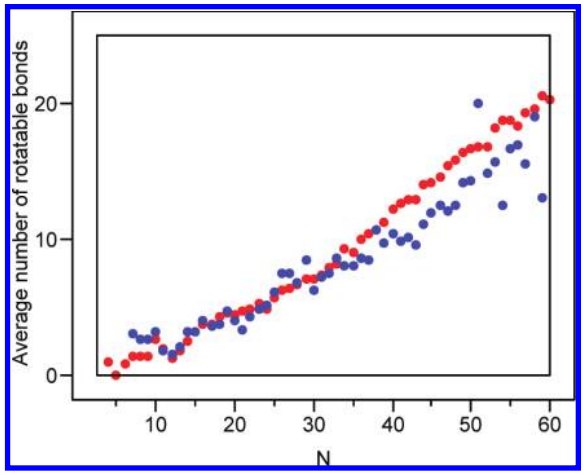


**Figure 6.** Average number of rotatable bonds as function of ligand size for ligands with SILE > 3 (blue) and those with SILE ≤ 3 (red) in the $K_i$ data set.

$\Delta h_{apolar}$ and $\Delta h_{polar}$ are scale factors describing energetic contributions per area gained.[6] In this paper we have modeled *optimal* affinity, and for ligands displaying maximum affinity it is expected that the conformational term $\Delta H_{conf}$ will be small compared to the sum of other enthalpic and entropic contributions (ie. an optimal ligand will likely adopt a bioactive conformation easily). ΔASA and $\Delta ASA_{polar}$ will show a strong linear correlation with the total ligand area available, assuming that a large proportion of the ligand interacts with the protein. As a result, the enthalpic contribution arising from an optimally binding ligand is expected to be proportional to the change in contact area between ligand and protein upon binding, to a first approximation.

The entropic contribution to binding is much more complex, and has been described as follows in the literature[6]

$$\Delta S_{tot} = \Delta S_{solv} + \Delta S_{conf} + \Delta S_{r/t} \qquad \text{(A3)}$$

The term $\Delta S_{solv}$ describes the change in entropy resulting from solvent release upon binding of the ligand; $\Delta S_{r/t}$ entails the loss of translational and rotational freedom upon binding from solution; $\Delta S_{conf}$ is a configurational term that captures the change in rotational degrees of freedom in ligand, water molecules, and protein side chains. The most important contribution to $\Delta S_{tot}$ arises from $\Delta S_{solv}$, which can be approximated using the observed change in heat capacity ($\Delta C_p$)

$$\Delta S_{solv} \propto \Delta C_p \qquad \text{(A4)}$$

The change in heat capacity $\Delta C_p$ can be modeled as a linear combination of changes in polar and nonpolar surfaces buried upon binding,[7,6] and as a result, $\Delta S_{solv}$ is expected to be proportional to ligand surface area in a first approximation.

Equation A5 has been proposed to model the configurational entropy contribution for nonpeptidic ligands[8]

$$\Delta S_{conf} = k_1 \times N_{rb} + k_2 \times N \qquad \text{(A5)}$$

Here, $N_{rb}$ is the number of rotatable bonds, and $N$ is the number of atoms. The first component in this equation models the change in conformational flexibility upon burial of the ligand, the second component (often omitted from docking and scoring functions) reflects the effect of water expulsion resulting from replacement by ligand bulk (with ligand excluded volume assumed to be proportional to the number of atoms). Coefficients $k_1$ and $k_2$ were estimated as $-1.76$ cal(K mol)$^{-1}$ and 0.414 cal(K mol)$^{-1}$, respectively,[8] and oppose each other. The number of rotatable bonds can be shown to vary linearly with the number of atoms in a molecule, on average (results not shown, $N_{rb} \approx -3 + 0.4N$, $R^2 = 0.95$).

With the notion that number of atoms and volume correlate, and that both enthalpy and entropy contain terms that are largely area-dependent, above observations lead to our approximation in eq 2, namely, that optimal affinity ∝ volume + $s \times$ area ($s$ is a scale factor; eq 6), and the assumption that the volume-dependent terms in the entropic contribution are relatively small compared to the area-dependent terms in eq 7.

## REFERENCES AND NOTES

(1) Reynolds, C. H.; Tounge, B. A.; Bembenek, S. D. Ligand Binding Efficiency: Trends, Physical Bias, and Implications. *J. Med. Chem.* **2008**, *51*, 2432–2438.
(2) Kuntz, I. D.; Chen, K.; Sharp, K. A.; Kollman, P. A. The Maximal Affinity of Ligands. *Proc Natl. Acad. Sci. U.S.A.* **1999**, *96*, 9997–10002.
(3) Chen, X.; Lin, Y.; Liu, M.; Gilson, M. K. *The Binding Database: Data Management and Interface Design Bioinformatics* **2002**, *18*, 130–139.
(4) Liu, T.; Lin, Y.; Wen, X.; Jorissen, R. N.; Gilson, M. K. BindingDB: A Web-Accessible Database of Experimentally Determined Protein-ligand Binding Affinities. *Nucleic Acids Res.* **2007**, *35*, D198–D201.
(5) Verdonk, M. L.; Rees, D. C. Group Efficiency: A Guideline for Hits-to-Leads Chemistry. *ChemMedChem* **2008**, *3*, 1179–1180.
(6) Perozzo, R.; Folkers, G.; Scapozzo, L. Thermodynamics of Protein-Ligand Interactions: History, Present, and Future Aspects. *J. Recept. Signal Transduction* **2004**, *24*, 1–52.
(7) Murphy, K. P.; Freire, E. Thermodynamics of Structural Stability and Cooperative folding Behaviour in Proteins. *Adv. Protein Chem.* **1992**, *43*, 313–361.
(8) Bardi, J. S.; Luque, I.; Freire, E. Structure-Based Thermodynamic Analysis of HIV-1 Protease Inhibitors. *Biochemistry* **1997**, *36*, 6588–6596.
(9) Miller, J. F.; Furfine, E. S.; Hanlon, M. H.; Hazen, R. J.; Ray, J. A.; Robinson, L.; Samano, V.; Spaltenstein, A. Novel Arylsulfonamides Possessing Sub-picomolar HIV Protease Activities and Potent Anti-HIV Activity against Wild-Type and Drug-Resistant Viral Strains. *Bioorg. Med. Chem. Lett.* **2004**, *14*, 959–63.
(10) Weislow, O. S.; Kiser, R.; Fine, D. L.; Bader, J. P.; Shoemaker, R. H.; Boyd, M. R. New Soluble Formazan Assay for HIV-1 Cytopathic Effects: Application to High Flux Screening of Synthetic and Natural Products for Aids Antiviral Activity. *J. Nat. Cancer Inst.* **1989**, *81*, 577–586.
(11) Hajduk, P. J. Fragment-Based Drug Design: How Big Is Too Big? *J. Med. Chem.* **2006**, *49*, 6972–6976.
(12) Wyatt, P. G.; Woodhead, A. J.; Berdini, V.; Boulstridge, J. A.; Carr, M. G.; Cross, D. M.; Davis, D. J.; Devine, L. A.; Early, T. R.; Feltell, R. E. E.; Lewis, E. J.; McMenamin, R. L.; Navarro, E. F.; O'Brien, M. A.; O'Reilly, M.; Reule, M.; Saxty, G.; Seavers, L. C. A.; Smith, D. S.; Squires, M. S.; Trewartha, G.; Walker, M. T.; Woolford, A. J. A. Identification of *N*-(4-Piperidinyl)-4-(2,6-dichlorobenzoylamino)-1*H*-pyrazole-3-carboxamide (AT7519), A Novel Cyclin Dependent Kinase Inhibitor Using Fragment-Based X-ray Crystallography and Structure Based Drug Design. *J. Med. Chem.* **2008**, *51*, 4986–4999.
(13) Huang, J. W.; Zhang, Z.; Wu, B.; Cellitti, J. F.; Zhang, X.; Dahl, R.; Shiau, C. W.; Welsh, K.; Emdadi, A.; Stebbins, J. L.; Reed, J. C.; Pellecchia, M. Fragment-Based Design of Small Molecule X-Linked Inhibitor of Apoptosis Protein Inhibitors. *J. Med. Chem.* **2008**, *51*, 7111–7118.
(14) Labute, P. A Widely Applicable Set of Descriptors. *J. Mol. Graph. Model.* **2000**, *18*, 464–477.

CI900094M