Ligand Efficiency Indices for an Effective Mapping of Chemico-Biological Space: The concept of an Atlas-like representation.

Authors:

Cele Abad-Zapatero¹,*, Ognjen Perišić¹,⁴, John Wass², A. Patrícia Bento³, John Overington³, Bissan Al-Lazikani³,⁵, Michael E. Johnson¹,

Affiliations:
¹Center for Pharmaceutical Biotechnology, University of Illinois at Chicago, College of Pharmacy.
²Independent statistics consultant, 602 Forest Hill, Rd. Lake Forest, IL. 60045.

*To whom correspondence should be addressed at:
¹Center for Pharmaceutical Biotechnology, MBRB building
University of Illinois at Chicago (MC870)
900 S. Ashland St. Room 3020
Chicago, IL, 60607
e-mail: caz@uic.edu

³European Bioinformatics Institute (EMBL-EBI)
Wellcome Trust Genome Campus
Hinxton, Cambridge
CB10, 1SD, UK

Current addresses:
⁴Department of Chemistry and Courant Institute of Mathematical Sciences
New York, University
251 Mercer Street
New York, NY. 10012.

⁵The Institute of Cancer Research
Haddow Laboratories
Belmont, Sutton, Surrey, UK

Version 3. 2/11/2010 5:02 AM
Abstract

We propose a numerical framework that permits an effective Atlas-like representation of Chemico-Biological Space (CBS) based on a series of Cartesian planes mapping the ligands with the corresponding targets connected by an affinity parameter (Kᵢ or related). The numerical framework is derived from the concept of ligand efficiency indices (LEIs), which provide a natural coordinate system combining the potency toward the target (biological space) with the physicochemical properties of the ligand (chemical space). This framework facilitates navigation in the multidimensional drug-discovery space using map-like representations based on pairs of combined variables related to the efficiency of the ligands per Dalton (MW or number of non-hydrogen atoms) and per unit of polar surface area (or number of polar atoms).
**Introduction**

Successful drug discovery requires the optimization of a large number of variables in two different domains: chemical and biological. The range of variables expands from the strictly physicochemical properties of the ligand such as MW to the more complex related to its affinity toward the target and to bioavailability and toxicity in the clinic. Navigating through this N-dimensional space using the potency of the ligand toward the target (i.e. Ki, IC50 or related quantities) alone often leads to good inhibitors but not successful drugs in the clinic. The impact of the advances in the related areas of molecular and structural biology, assay development, human genomics, proteomics, fragment-based drug discovery and other related technologies on this process, although noticeable, is still far from dramatic [1, 2].

Conceptually, the introduction of filtering criteria such as the Rule of 5 (Ro5) [3] a decade ago and related ones (such as the Rule of 3) in Fragment-Based strategies [4] was expected to expedite the discovery of chemical entities with effective potency in the clinic. Although this expectation has been fulfilled in certain cases, the effectiveness of the entire drug discovery process is under scrutiny [2, 5].

A more streamlined and effective framework is needed to facilitate the drug discovery process, making it more efficient. Toward that ultimate goal and for a number of years now, the number of databases dedicated to storing and annotating target:ligand complexes (SAR databases) has grown dramatically as a natural extension of the information contained at the Protein Data Bank [6]. Currently, PDBBind [7, 8], BindingDB [9, 10], DrugBank [11], MOAD [12, 13] and WOMBAT [14] and ChEMBL [15] (www.ebi.ac.uk/chemblpdb) among others have extended dramatically the number of observations in Chemico-Biological Space (CBS). These resources have amassed vast
amounts of high-quality data connecting ligands to their targets (PDBBind, BindingDB, MOAD, ChEMBL) complemented, when available, with structural information and relating them to relevant therapeutic entities (DrugBank, http://redpoll.pharmacy.ualberta.ca/drugbank).

The vastness of chemical space [16] and certain tools to navigate it [17] have already been presented and discussed. Briefly, chemical space has been considered before as being analogous to the cosmological universe with chemical compounds populating the space instead of stars [16]. However, the existence of this vast amount of data combining compounds and targets, without sound underlying principles, has been compared to the status of astronomy prior to the formulation of the Kepler Laws of planetary motion [18]. Is it possible to find certain generalizations that can effectively describe and represent the CBS? Can those insights be translated into a more efficient drug-discovery process?

Two conceptual trends in drug discovery have emerged in the last decade that could be combined to provide an effective mapping of CBS. One is the concept of ‘Chemogeography’ and the corresponding chemical global positioning system (ChemGPS) [17] as an aid to navigate chemical space. The other is the introduction of combined variables referred to as ligand efficiency (LE) or Ligand Efficiency Indices (LEI) to relate the potency of the compounds to their size and polarity [19],[20].

Any combination of chemical space with biological (or affinity) space raises many important issues related specifically to biological data that should be mentioned up front. The physiochemical properties associated with the chemical ligands are very well defined. In contrast, the data associated with the target, in particular the data related to
the binding affinity of the ligand and the target(s) toward which the chemical entity has
measureable activity, should be considered always with caution. Important issues that
always need to be kept in mind are: specific assay details, reproducibility within a
specific assay, variability among different assays, sample size and species variation,
among the most prominent variables that should be considered.

In spite of the above caveats, we propose that a combined representation of CBS
has merit. In science and technology, quite often, the introduction of new variables is a
first step in helping to provide a new framework that may be more effective to solve long
standing problems. We suggest that LEIs can play such a role in drug discovery. They
naturally connect the physicochemical properties of the ligand (i.e., MW, PSA and
others) with the biological target (typically a protein) via the affinity parameter Ki (or
related ones).

In this work, we briefly summarize the earlier concepts used to navigate chemical
space and present an algebraic framework based on LEIs that permits an intuitive,
graphical, representation of the combined CBS. We explain the appearance and overall
properties of this representation and provide some examples of the use of this
framework in drug discovery. It is presented as an initial concept for future application
and development. How this framework and graphical representation could facilitate
drug discovery in the future will require the application to specific examples in a
prospective and predictive manner.

1. Background concepts

1.1 Chemical Space. The introduction of combinatorial chemistry in drug
discovery made apparent the need for tools to navigate the vastness of chemical space.
The notions of ‘Chemogeography’ and a software navigating system termed ‘ChemGPS (chemical global positioning system) were introduced to relate the chemical space to drug space, i.e., subset of compounds that correspond to drugs in the market. The coordinates of ChemGPS maps were based on the t-scores derived from a principal component analysis (PCA) of 72 descriptors of various physicochemical and topological properties of the chemical entities [17]. From the drug-discovery perspective, the challenge is to identify regions of chemical space that contain biologically active compounds for certain biological targets [16]. Lipinski and Hopkins hypothesized that within the continuum and vastness of chemical space there would exist enclosed, discreet, regions occupied by compounds with specific affinity towards certain biological targets or even gene families (i.e., proteases, kinases, GPCRs and others). The question of what variables (or coordinate system) would allow this separation was left open.

1.2 Efficiency Indices. A first attempt to quantify the binding affinity of a ligand in relation to the number of non-hydrogen atoms was presented by Kuntz et al. [21]. An extension of their original concept was given an explicit formulation by Hopkins et al. [19]. They defined ‘Ligand efficiency’ (LE) numerically as the quotient between $\Delta G$ and the number of non-hydrogen atoms of the compound:

$$LE = \Delta g = \frac{\Delta G}{NHEA} \quad \text{eq. (1)}$$

where $\Delta G = -RT\ln K_i$ and NHEA is the number of non-hydrogen atoms (or heavy atom count, HAC). The units of LE ($\Delta g$) will be kcal/mol per non-hydrogen atom (Table 1). Variations and extensions of this idea continue to appear in the literature [22] and is gaining a wider acceptance among medicinal chemists. In particular, the concept of
‘group efficiency’ has been put forward as an extension of the original idea to estimate the binding efficiency of parts of a molecule, or of groups added to an existing fragment or lead molecule [23].

Since non-hydrogen atoms can be of many different types and a key property of a compound is its molecular weight (MW), a natural extension of the concept of LE (eq. 1) was introduced soon thereafter without using the energy units. Binding Efficiency Index (BEI) is a simpler index based on the ratio of the binding affinity given as $pK_i = -\log K_i$, and using MW as reference expressed in kiloDaltons (eq. (2), Table 1) (Box 1 in reference [20]). This concept of ligand efficiency in various definitions is gaining wider acceptance in the medicinal chemistry literature [24],[25],[23] (and references therein), especially in relation to guiding Fragment-Based Strategies to expedite drug discovery [2, 26]. It has also been used to dissect the most efficient fragment in the natural product Argifin, a cyclopentapeptide inhibitor of chitinase activity [27].

Although always considered to be important, much less direct use has been made of the concept of ligand efficiency as referred to the polarity of the compound. A specific definition of this concept (Surface Efficiency Index: SEI) has been introduced and defined using the affinity of the ligand ($pK_i$, defined as above) and the polar surface area (PSA) of the ligand scaled to 100 Å² ([28], [20]). We propose that an effective use of the concept of LEIs in the future should require the use of multiple non-redundant variables, within a numerical formulation that uses similar numerical scales. In the future of drug discovery, it is most likely the combination of several LEIs that will have the highest predictive value. An initial illustration of this concept combining SEI-BEI in a Cartesian plane for inhibitors of human PTP1B has been published [28].
2. Selecting the optimum variables to map CBS.

Table 1 lists several definitions of LEIs that we have explored and compared in proposing these concepts. As summarized, initial analyses favored formulations relating directly the potency (Ki, IC₅₀ or related measurements) to the MW (in kDa) and to the PSA of the ligands using BEI and SEI respectively [20]. This gave us the insight that the distribution of chemical compounds in the optimization plane was very easily understood in terms of a wedge of varying ratios of PSA/MW across the SEI-BEI plane (Fig. 5 of reference [28]). This result was confirmed using the PDBBind data as shown in the supplementary materials (section I, Fig. S1a).

However, there was no particular reason to suggest that this approach provided the best or only framework in the long term. Thus, we explored two formulations related to the number of atoms of the compounds, specially the number of heavy atoms (non-Hydrogens, NHEA also referred to as HAC or heavy atom count [19]) and the number of polar atoms (NPOL= number of O and N). This approach is reasonable since NHEA correlates well with the MW of the compounds (R²= 0.96 for the ligands in PDBBind refined set, n=1301; data not shown) and the second, NPOL, is related to the PSA (R²=0.91, same sample data not shown). More importantly also, this framework has the advantage that it relates directly to the mindset of the medicinal chemists working at the bench. It also relates better to previous work in the field [19, 22]. The working definitions of LEIs related to the number and type of atoms present in the ligand that have been explored in this work are shown in Table 1. An example of the distribution of target:ligand pairs in the 2007 version of PDBBind using variables NSEI-NBEI is shown
in the supplementary material (section I, Fig. S1b). This list cannot be considered to be exhaustive and it is indeed quite possible that additional LEIs will be defined in the future to optimize the drug-discovery process. However, it will be desirable that all can be related to one another in similar scales to provide additional dimensions for future numerical or statistical optimization.

Although the representation of CBS given by the planes SEI-BEI and NSEI-NBEI can be useful for certain applications, we would like to emphasize the representation in Cartesian planes defined by the variables NSEI-nBEI and NSEI-mBEI (x-y) (see Table 1 and Figs. 1a,b respectively). These figures depict the distribution of the PDBBind database compounds in the planes defined by variables nBEI vs.NSEI and mBEI vs. NSEI. The plots correspond to the distribution of lines in the plane described by equations (8) and (9), respectively, as shown in the supplementary material section II (Box S1). The change in the definition of the atom-related related variables (NBEI= -(log_{10}Ki)/NHEA as opposed to nBEI= -log_{10}([Ki/NHEA]), see Table 1) and their combination in the NSEI-nBEI plane (Eq. 8) has allowed a clear, fan-like, separation of all the compounds in the database in terms of the number of polar atoms (NPOL, slope) and number of heavy atoms (NHEA, intercept: log_{10}(NHEA). The regression analysis of all the lines ranging from NPOL=1 to NPOL=10 (R^2 > 0.99) are shown in Fig. 1a. The statistical analysis (data not shown) show that the model based on the slope of the lines alone accounts for over 99% of the variance in the mean value of the slopes due to the small differences in the intercept (log_{10} (NHEA)). As explained in the derivation (supplementary material section II Box S1), this is due to the fact that the intercept of the lines is given by logarithm of a relatively small number (log_{10}(NHEA)).
A similar result is obtained in the mBEI-NSEI representation, as depicted in Fig. 1b where the intercept is \( \log_{10}(MW) \) (Eq. 9). This alternative representation allows a better separation of the compounds along the intercept \( \log_{10}(\text{NHEA}) \) vs. \( \log_{10}(MW) \) Figs. 1a,b). Because of their fan-like nature, we refer to these plots as ‘fan-plots’ of the database, providing a rapid visual map of the content of any target:ligand database and their relative affinities (as given by \( K_i \) or related parameters).

It is important to emphasize that in either of these representations of the contents of PDBBind database, any small ligand can be placed along a specific line in the diagram given only its chemical composition: number of non-hydrogen atoms (NHEA or MW related to the intercept) and number of polar atoms (NPOL or PSA/MW ratio, related to the slope of the line). Within that line, the unique position of the target:ligand pair on the map is determined only when \( K_i \) values (experimental or calculated) are available with respect to a specific target. The same ligand with different affinities towards N separate targets, will be represented by N points located along the line; the higher the potency the farther away from the origin. In this way, the uncertainty in the affinity parameter(s) indicated before as a weakness in representing biological data, will be reflected by a ‘sliding’ position of the target:ligands pairs along the line, corresponding to the physicochemical properties of the ligand (PSA/MW, NPOL values defining the slope of the lines). Thus, the x (SEI-like) and y (BEI-like) coordinates of any target:ligand complex in the plane represent the polarity and size-related components respectively, of the efficiency of that ligand towards the total ligand efficiency index (LEI).

The previous analysis shows that it is possible to represent visually the connection between the chemical ligands and the biological targets in a two-dimensional, Cartesian plane. Any target: ligand complex in any database (as illustrated for PDBBind and WOMBAT in Figs. 1-2) can be represented by a point in the plane, given the physicochemical characteristics of the compound (i.e., PSA, MW or NPOL, NHEA) and its affinity towards a specified target (Ki). Obviously, this can be done by computing the corresponding variable pairs (SEI, BEI; NSEI, NBEI; or nBEI, NSEI or other combinations) and plotting them in a map-like fashion. The collection of maps of CBS in different coordinates (physical and atomic) for different regions of the plane and at different scales is what we would like to call AtlasCBS. This collection could be conceived as a series of pages (physical and/or electronic) that can be plotted, stored, searched, updated and used as a graphical representation of CBS and as a guide for drug discovery.

A selected combination of variables permits relating this graphical representation of CBS to more familiar concepts in medicinal chemistry and drug discovery. For instance, in the SEI-BEI plane, the ratio BEI/SEI (or NBEI/NSEI; i.e., the slope of the lines in the plane) is independent of the pKi values and is directly proportional to PSA/MW (or NPOL/NHEA) (see note on Table 1). This ratio (or NPOL/NHEA) decreases as the compounds move to the right, improving their SEI (NSEI) and shifting the compound towards more favorable drug-like properties. A linear regression analysis of ClogP vs. PSA/MW for the available data in WOMBAT shows that ClogP≈ -15.0*(PSA/MW) + 5.7 (R² =0.904)(data not shown).
Since PSA/MW is related to the solubility of the compounds, the decrease in values (counterclockwise) across the plane corresponds to the decrease of LogP from the most hydrophobic compounds (right) to the most polar (left). This is equivalent to an angular representation of the variable LogP of the ligands across the suggested ligand efficiency map. The importance of this variable in the drug-discovery process cannot be overemphasized as indicated by its prominent role in Lipinski’s Ro5 [3] and its variations (Ro3) [4].

In addition to the angular coordinate related exclusively to the properties of the ligand, the proposed two dimensional representations of CBS also contain information about the biological target, implicitly included in the definitions of the LEIs, via the affinity variable. The relationship to the target in the map is given by the radial distance of the point (representing the target:ligand complex) to the origin: analogous to its radial component (Fig. 2a). Thus, a compound or a series of compounds with nearly identical PSA/MW ratios can be placed anywhere along a certain line in the plane given its PSA/MW value. It is only when a pKi value is available that it is possible to place unambiguously the compound(s) at a specific point(s) within the line given their PSA/MW ratio. The most polar ligands within the different targets lay on the left and they become progressively less polar (more drug-like) as the PSA/MW ratio becomes smaller as it is shown for an extensive drug discovery effort directed towards human PTP1B (Fig.5 of reference [28]). How potent the compounds are for a specific target is proportional to the distance from the origin along the appropriate line. The power of this representation to illustrate a wide range of physicochemical properties of the ligands, as they relate to the targets, is shown in Fig. 2a. The content of the database WOMBAT
has been represented in the BEI-SEI plane using a wide range of values for both variables to illustrate the separation in chemical space between conventional (non-GPCR, upper left) and GPCR targets.

This **conceptual** separation is still valid for the representations of the optimization plane in efficiency indices related to the atomic composition of the ligands, namely NSEI, NBEI (supplementary material section I) and NSEI, nBEI (see Fig. 2b inset). In addition, in the latter plane, the separation of the different lines is more dramatic and more intuitive since the slopes are given by the number of polar atoms of the ligand, NPOL and varies dramatically for the low slope values (1,2,3..etc); it approaches a narrowly separated family of lines at high NPOL values.

A word of caution is appropriate here due to the wide variability of the biological data in any biological project as indicated above. LEIs can be calculated based on Ki’s, Kd’s and IC50’s as the most significant in measuring the affinity of the ligands towards the target. For the most part and preferably, the values should be experimental. Theoretical values could be considered only as a guide and approximation (see Box A). In any serious analysis involving LEIs the values should only be compared if the same variable is used to calculate the relative efficiencies. For comparison of internal compounds against compounds from different or competing groups the affinity should be measured under the same assay conditions. In addition, this representation of CBS is focused on ligands that inhibit the activity of the biological target (i.e., antagonists of receptors). Possibly, alternative efficiency indices can be devised in the future for agonists and a similar representation could be found. In the future, it might also be possible to add a third dimension related to the efficacy of the ligands in *in vivo* assays.
The Atlas-like representation of CBS can also be related to geographical maps in a different way. The macroscopic variables (PSA, MW) would correspond to exploring the physicochemical properties of the ligands (analogous to mountains, rivers and physical features of geographical maps), and the atomic properties (i.e., NPOL, NHEAV) would be analogous to more discrete features of the landscape such as cities, major highways, or even metro-stops or blocks at a magnified scale. We present some applications of the use of this Atlas-like representation (and related optimization planes) in two areas of drug discovery in the supplementary materials (section III).

4. Applications

Two examples of the application of the AtlasCBS representation to drug discovery are presented in the supplementary materials (section III).

1. Trajectories in drug-discovery space as illustrated for human Protein Tyrosine Phosphatase 1B (Figs. S2a-d). (Application 1)

2. Mapping of compounds with inhibitory and pharmacological activity for the HIV-I protease (Figs. S3a-c). (Application 2).

5. Conclusion: A natural representation of CBS

The affinity (or potency) between the ligand and the target (Kᵢ, IC₅₀ or related parameters) has been the dominant parameter in drug discovery probably ever since reliable estimates could be measured in vitro or in cell assays. The knowledge and experience acquired within the last decade or so, regarding the constraints imposed by the physicochemical properties of the ligands, for the success of preclinical drug-discovery programs has been incorporated into empirical rules-of-thumb that have
guided the discovery effort [3, 29] during the last decade. However, these guidelines do not consider the potency towards the target and thus have only limited predictive value and effectiveness in the tortuous road of drug discovery. We wish to suggest that LEIs offer a more natural and effective combination of variables that can graphically aid in mapping CBS. These combined variables encompass the more established variables (i.e., LogP, MW, PSA etc) and provide a graphical representation of the guiding criteria that are being used currently, for example Ro5. We suggest that these novel variables could, in the near future, make the drug discovery process more effective. How these ideas can be incorporated into the overall drug discovery process has been presented before [28].

We are suggesting that LEIs relating potency to two critical physicochemical properties of the ligand (i.e., size and polarity) can be defined in two different but related ways: i) relating to physicochemical properties such as MW, PSA (i.e., BEI, SEI; mBEI); and ii) referring to the atomic composition of the ligand as given by NHEA, NPOL (i.e., NBEI, nBEI, NSEI), respectively. We propose that suitable combinations of pairs of these variables (combining size and polarity-related variables in a Cartesian plane) can effectively aid in mapping the chemical space of available chemical matter and relate it to the targets towards which they have a measurable affinity. The representation is akin to an Atlas, representing how the different physicochemical properties of the ligands (angular coordinate) relate to the corresponding potency towards biological targets (radial coordinate).
It is anticipated that a wider use of this conceptual, numerical and graphical representation of the extensive public and proprietary target:ligand databases of inhibitors and commercial drugs at different stages along the approval pathway, could point the way towards a more effective and efficient drug discovery process. Possibly, this framework could be instrumental in identifying regions of CBS with higher probability of yielding clinical drugs for specific targets. This ‘drug-likelihood’ estimate (Drug Efficiency Index: DEI) could be incorporated into the proposed framework as an elevation coordinate (or additional coordinate) thus providing an effective mapping of the drug-discovery landscape.

Acknowledgments

Access to the data in WOMBAT granted by Prof. T. Oprea is fully appreciated. The authors appreciate the constructive discussions, criticism and support given to the project by Prof. H. Lu and his group in the Dept. of Bioinformatics and Bioengineering at UIC. Comments and discussions with Dr. Debbie Mulhearn and computer support by Dr. B. Santasiero are also greatly appreciated. Drs. Kent Stewart, Yvonne Martin and James Metz from Abbott Laboratories read previous versions of the manuscript and provided valuable criticism. Also Dr. Federico Gago from the University of Alcalá de Henares, Madrid, Spain. The contribution of Dr. P. Bento at EMBL-EBI in preparing the figures is greatly appreciated as well as the hospitality of the entire ChEMBL group at the Wellcome Trust campus, Hinxton, UK.

The medicinal chemistry and computational groups at Merck-Serono (Drs. Wolfgang Sauer, Serge Christman-Franck, Mireille Krier and colleagues) were exposed
first to these ideas and their positive response gave the impetus and motivation to continue.

Box A. Simple is beautiful

The information necessary to map compounds, chemical series and efficiency data on efficiency space is readily available using conventional software tools. The overall concept is summarized in Fig. 3.

- Given the empirical formula or SMILES string of a compound and a measure of affinity (Ki) or activity (IC₅₀), extract the following information. NPOL: number of N, O atoms; NHEA: number of non-hydrogen atoms; MW and tPSA (or any other estimate of the PSA). pKi = -log₁₀ Ki).

- Calculate the different LEIs as defined in Table 1. Create a spreadsheet with all the compounds including ‘reference’ compounds and/or compounds from other groups.

- Plot: NSEI, nBEI (x,y). Other LEIs planes (NSEI, NBEI; SEI, BEI) can also be used for fine tuning the analysis. In NSEI-nBEI, the compounds will appear along a series of lines with slopes equal to NPOL.

- If the reference compound has N polar atoms, it will appear on the corresponding NPOL line. Highly polar, very size-efficient compounds will plot to the left (in the upper left part) of the plot; compounds that are very potent and very hydrophobic will map to the lower right.
• The line NPOL=N basically divides the plane in two halves. Compounds with NPOL = N+1 will map to the left of the reference line; compounds with NPOL= N-1 will map to the right of the corresponding line.

• The position of a ‘hypothetical’ compound with a number of polar atoms different from N will map in the corresponding part of the map [30], in relation to the starting compound with N polar atoms (N + O). If no affinity data are available the compound cannot be placed uniquely along the new NPOL line; once a measurement (or theoretical estimate) has been obtained the target:ligand pair can be located unambiguously in the plane (Fig. 3 legend: solid arrows vs. dashed arrows).

• Strategy: try to optimize (maximize) both variables together in the plane insofar as possible. Compounds or series that move towards the diagonal of the plot optimize both variables simultaneously.

• Series with the same number of polar atoms (irrespective of where they occur in the structure of the ligand!) will map along the corresponding NPOL line as determined by the corresponding potency: more potent compounds will move up; less potent compounds will move down, along the line (Fig. 3). The most effective way to explore alternative regions of the plane is probably to change the number of polar atoms of the compound/series by using ‘replacement’ rules [30]. This is equivalent to jumping from ‘line-to-line’ in the NSEI-nBEI plane (‘line hopping’).
For certain targets and chemical series, when an optimum compromise has been found for both NSEI and nBEI and solubility issues are encountered, the best strategy may be to move to the left (increase solubility by adding polar atoms) but the efficiency per atom added (nBEI) should be retained or increased in so far as possible.
### Table 1

<table>
<thead>
<tr>
<th>Name</th>
<th>Definition</th>
<th>Example Value&lt;sup&gt;a&lt;/sup&gt;</th>
<th>Equation</th>
</tr>
</thead>
<tbody>
<tr>
<td>LE&lt;sup&gt;b,c&lt;/sup&gt;</td>
<td>ΔG/NHEA(non-hydrogen atoms)</td>
<td>-0.50&lt;sup&gt;c&lt;/sup&gt;</td>
<td>(1)</td>
</tr>
<tr>
<td>BEI</td>
<td>p(Ki), p(Kd), or p(IC&lt;sub&gt;50&lt;/sub&gt;)/MW(kiloDa)</td>
<td>27</td>
<td>(2)</td>
</tr>
<tr>
<td>SEI</td>
<td>p(K&lt;sub&gt;i&lt;/sub&gt;), p(K&lt;sub&gt;d&lt;/sub&gt;), or p(IC&lt;sub&gt;50&lt;/sub&gt;)/(PSA/100 Å²)</td>
<td>18</td>
<td>(3)</td>
</tr>
<tr>
<td>NSEI</td>
<td>NSEI = -log&lt;sub&gt;10&lt;/sub&gt; Ki/(NPOL) = pKi/NPOL(N,O)</td>
<td>1.5</td>
<td>(4)</td>
</tr>
<tr>
<td>NBEI</td>
<td>NBEI = -log&lt;sub&gt;10&lt;/sub&gt; Ki/(NHEA) = pKi/(NHEA)</td>
<td>0.36</td>
<td>(5)</td>
</tr>
<tr>
<td>nBEI&lt;sup&gt;d&lt;/sup&gt;</td>
<td>nBEI = -log&lt;sub&gt;10&lt;/sub&gt;[(Ki/NHEA)];</td>
<td>10.25</td>
<td>(6)</td>
</tr>
<tr>
<td>mBEI</td>
<td>mBEI = -log&lt;sub&gt;10&lt;/sub&gt;[(Ki/MW)]</td>
<td>11.5</td>
<td>(7)</td>
</tr>
</tbody>
</table>

<sup>a</sup>Reference values and definitions of BEI, SEI adapted from reference [28].

Ki or IC<sub>50</sub> = 1.0 nM; Molecular Weight = 333 Daltons (or 0.333 KiloDaltons)

This value of MW is near the mean value of MW for a large sample of marketed oral drugs [31] [32].

Van der Waals PSA = 50 Å²

<sup>b</sup>For the atom-related definitions, LEIs are calculated for each index using the following idealized values (units have been omitted in the table): K<sub>i</sub> or IC<sub>50</sub> = 1.0 nM; p(K<sub>i</sub>) = –log Ki = 9.00; MW = 0.333 kDa;

NHEA = Number of heavy atoms (non-Hydrogen in the compound) = 25; NPOL = Number of polar atoms (N,O) = 6;

<sup>c</sup>ΔG = -12.4 kcal/mol, ΔG = -RTlnKi, assuming Ki=1.0 nM, T=300º K; NHEA(non-hydrogen atoms) = 25; corresponding to a mean MW/atom of 13.3 Daltons.

<sup>d</sup>Hopkins et al [19].

<sup>d</sup>This small change in the definition of NBEI (taking the log<sub>10</sub> after the ratio of (Ki/NHEA) was found to be critical for the separation of the compounds in the nBEI-NSEI plane (see text). Similarly, for mBEI-NSEI pair (see equations (7),(9)).

**NOTE:** By definition, for any given compound the ratio of BEI/SEI is equal to 10(PSA/MW) or NBEI/NSEI = NPOL(N,O)/NHEA(non-H)= 0.36/1.5 = 6/25 = 0.24;
Figures

Fig. 1a

Polarity increases counter clockwise by NPOL

Fig. 1b

Potency, efficiency along the lines.
Fig. 1. Representation of the content of PDBBind in a Cartesian plane defined by LEIs related to the number of atoms of the ligand.

1a) Representation in a Cartesian (Atlas-like) map of the entries in the PDBBind (2007) refined set in the nBEI-NSEI plane. Each point corresponds to a target:ligand complex (one PDB access code, 1283 entries) and the slope of each line corresponds to a certain number of polar atoms in the ligand as indicated (see upper right-hand panel). Intercept is related to the number of heavy (non-hydrogen atom) in the ligand: $\log_{10}(\text{NHEA})$. NPOL range 1-28. (See Table 1 for definitions and Box A for the algebraic derivations). The chemical structures of the high efficiency ligands in the data set (PDB accession codes highlighted) have been shown below to show the progressive increase in the number of polar atoms and polarity of the ligands, as the slope increases counterclockwise.

1b) Representation in a Cartesian (Atlas-like) map of the entries in the PDBBind (2007) refined set in the mBEI-NSEI plane. The same set in the atom-related efficiency indices mBEI-NSEI. The intercept is related to the MW of the ligand as $\log_{10}(\text{MW})$. The compounds map along lines of slope defined by the number of polar (N and O: NPOL) atoms in the ligand ranging from 1 to 28 (see side panel). The statistical analysis was performed as indicated in the ensuing review. NPOL range 1-28 (see side panel). See Table 1 and Box A for the variable definitions and algebraic derivations.

Images prepared with JMP7 (SAS Institute) and Spotfire™
Fig. 2a.

Fig. 2b
Fig. 2. Representation of the content of WOMBAT drug database illustrating the separation between GPCR and non-GPCR targets.

Fig. 2a. Separation of the conventional and GPCR targets in an optimization plane using BEI-SEI as variables with a wide range of values SEI (0.925); BEI (0,100). The large values of the extreme target:ligand complexes corresponds to very potent and very small (top left, high BEI and low SEI values: glycine bound to NMDA/gly) or very potent and extremely hydrophobic (lower right, high SEI and low BEI: amitriptyline bound to $\alpha_1$ receptor). Data access courtesy of T. Oprea [14] as included in WOMBAT 2007. 'Geographically', the separation between the GPCRs and the more 'conventional' targets in this representation could be equivalent to a major ocean separating two different continents. Inset shows a magnified view of the region corresponding to the conventional targets (SEI (0,120), BEI (0,100)) using the same variables BEI-SEI and an adjusted scale to fit into the available space (see Table 1 for definitions). The color gradient in both images is related to 10*PSA/MW (slope of the lines) from minimum (0.04, dark green), 1st quartile (1.23), median (1.92), 3rd quartile (2.95) and maximum 11.38, red). The non-GPCR targets account for 90.5% of the sample size. The blue arrow indicates simply a scale change.

Fig. 2b. The same data and variable range as before with the various target:drugs complexes colored by target. Inset also represents the region corresponding to the conventional targets (as Fig. 2a) but now magnified in a different scale and represented with the variables mBEI-NSEI (NSEI: (0,12), mBEI: (2,16)) to emphasize the separation of the different lines by the number of polar atoms NPOL of the ligands. The colors of
the different squares correspond to different targets. The textured blue arrow indicates a change in scale and a representation in different variables. Images prepared with Spotfire™.

---

**Fig. 3.**

**Fig. 3. Schematic representation of the application of the AtlasCBS concepts in drug discovery (Box A).**

The concepts presented and illustrated in this work suggesting a graphical representation to guide drug-discovery have been summarized in a schematic representation in the nBEI-NSEI plane. The guiding strategies and concepts have been summarized in Box A. Given the empirical formula of a compound it can be placed anywhere along a specific line based on NPOL and \(\log_{10}(NHEA)\). An experimental affinity towards a target (Ki or related) allows placing the target:ligand complex on a
unique point in the plane (green dot). Strategies to modify this starting compound have been outlined in the diagram. Adding three polar atoms will move the compound to the left (yellow arrows, more polar). A solid arrow is meant to indicate an experimental affinity measurement; dashed arrows (three colors) mark the possible trajectory based on theoretical estimates of the binding affinity. Retaining the same number of polar atoms will move the trajectory along the corresponding NPOL line up or down depending on the resulting potency. A blue dot suggests the position of an alternative series or the position of a compound from the competition. The dashed red box outlines the region of the CBS space where very few successful drugs have been documented. Details on how to use the optimization planes related to LEIs as variables to optimize drug discovery have been published before [28]. The target:ligand complexes included in the refined PDBBind set were statistically modeled by NPOL (see supplementary material section II Box S2) and are included as background for the discussion to simulate other existing complexes.
Supplementary Material I: Additional figures and figure legends

Figs. S1a, b summarize the results of plotting the distribution of ligands in PDBBind (2007) in the SEI-BEI and NSEI-NBEI planes, using the corresponding variables for the abscissa and ordinate, respectively. Both exhibit a wedge-shaped distribution consistent with the one suggested before [28] based on a very limited set of targets and compounds. Density distributions can be calculated to estimate the probability of a certain ligand occupying a certain region of the two-dimensional plane based on its values for two sets SEI,BEI or NSEI,NBEI (data not shown). Based on their PSA/MW (for Fig. S1a) or NPOL/NHEA (Fig. S1b) ratios, the planes can be divided in wedges that distribute the compounds in different polarity ranges from left (most polar) to right (least polar; slope equal to zero will imply PSA=0, or no polar atoms) and more hydrophobic. Overall, the distribution obtained using NSEI, NBEI seems to be a bit more discrete (or 'grainier') allowing for a wider separation of the target:ligand pairs. In both cases, but particularly suitable for the analysis of smaller sets of compounds, distinct lines of slope NPOL/NHEAV can be modeled (with statistically significant differences) in the data to facilitate the analysis of certain chemical classes. By a suitable combination of variable ranges, this type of representation of CBS can be useful when a large number of compounds need to be examined varying only slightly in their PSA/MW ratios or when a wide range of compounds/targets needs to be examined (see below, Figs. 3a-b). This representation could be analogous to the pages representing physical features (mountains, rivers, etc) in a conventional atlas and displaying continents.
Fig. S1a

Fig. S1b

Fig. S1. Representation of the content of PDBBind in a Cartesian plane defined by ligand efficiencies related to polarity (abscissa) and size (ordinate).
**Fig. S1a.** Overall two-dimensional distribution of the entries in the PDBBind refined set (2007) in the BEI-SEI plane (1283 entries). Each point corresponds to a target:ligand complex and represents a PDB access code; they have not been included for clarity. The dashed lines vary clockwise in a traffic-light spectrum from the more polar (left, PSA/MW=0.90) to the least (PSA/MW=0.10). The blue line indicates very hydrophobic compounds (PSA/MW=0.045 corresponds to a LogP \( \sim 5 \)). The slope of the lines is \(10^*(PSA/MW)\) (See Table 1 for definitions).

**Fig. S1b.** The same PDBBind set in the atom-related efficiency indices plane NBEI-NSEI. The slope of the lines is NPOL/NHEA and the value ranges of the variables are different from **Fig. S1a**; however, the shape of the plots is essentially the same. The meaning of the polar coordinates (angular and radial) has been added for easy reference and emphasis. Reference lines for the polarity of the compounds have been added as for **Fig. S1a**. Slope of the lines: NBEI/NSEI=NPOL/NHEA.
Box S1. Algebraic derivations

Using the definitions in Table 1 for NSEI and nBEI, it can be shown that a linear relationship can be obtained relating nBEI and NSEI as follows.

From eq. (4): $-\log_{10} K_i = \text{NPOL} \times \text{NSEI}$;

From eq. (6): $-\log_{10} K_i = \text{nBEI} - \log_{10}(\text{NHEA})$;

Eliminating $-\log_{10} K_i$ between these two equations and rearranging terms:

$$\text{nBEI} = \text{NPOL} \times \text{NSEI} + \log_{10}(\text{NHEA}) \quad \text{eq. (8)}$$

This equation describes a family of lines and represent the geometrical locus of all the target:ligand complexes in the chemical database, independent of the value of the variable chosen to describe the affinity. This is a linear relationship between nBEI and NSEI, where the slope (NPOL) and the intercept ($\log_{10}(\text{NHEA})$) depend only on the chemical properties of the ligand; the dependence on the target has been eliminated by eliminating the variable $K_i$ (i.e., mathematical locus). The resulting distribution of the ligands in PDBbind in the Cartesian plane defined by nBEI vs. NSEI is presented in Fig. 1a.

Similarly, selecting equations (4) and (7) (Table 1) a linear equation can be also obtained relating mBEI and NSEI (Fig. 1b):
In equation (9), the intercept is \( \log_{10}(MW) \) that relates to the MW of the ligand. This representation is conceptually the same as eq. (8) but allows a wider separation of the compounds on the vertical axis based on MW and permits a better way of monitoring the size of the ligands. The corresponding distribution of the ligands in PDBbind in the Cartesian plane defined by mBEI vs. NSEI is presented in Fig. 1b.

As indicated previously [28], the definition of the variables BEI vs. SEI and NBEI vs. NSEI implies that in the corresponding Cartesian planes the complexes (target:ligand) map along lines whose slope is equal to 10(PSA/MW) (ratio of BEI/SEI) or to NPOL/NHEA (ratio of NBEI/NSEI), respectively, of the ligands alone (see supplementary Figs. S1a-b).

The representation with polarity-related variables (SEI, NSEI) in the abscissa is preferable because in this way the polarity of the compounds increases counterclockwise across the plane from the least polar (low angular coordinate, low polarity and nearly horizontal lines) to the most polar (high values of the angular coordinate or the slope of the corresponding lines) at the upper left. As shown, the best overall separation is obtained when the slope of the lines is given by NPOL. However, significant statistical separation is also possible in terms of the PSA/MW and NPOL/NHEAV ratios in the corresponding SEI-BEI and NSEI-NBEI plots by a suitable choice of the scale (see Box S2 in the supplementary materials, section II).
Box S2. Statistical analysis

The statistical analysis of the samples analyzed in this work was performed using the following general flow. Except where noted, all the statistical manipulations and analysis were done with JMP (SAS Institute) (JMP6 or JMP7).

Step 1 – Examine Distribution of NSEI and nBEI looking for normality, outliers, and trends.

Step 2 – Fit Model: nBEI (or mBEI) by NSEI, NPOL, NSEI*NPOL. Using standard least squares; examine: leverage and regression plots, actual by predict; ANOVA for model, lack of fit test, residual by row plot.

Step 3 – Compare the Slopes ANCOVA, GLM in SAS(9.1). Examine: $R^2$, ANOVA of Model, contrasts.

Step 4 – Compare Factors in NSEI, nBEI (or mBEI) for any equivalent lines; examine t-test and low power thresholds.

As the analysis is dependent upon the valuations of NSEI and mBEI, a simple distributional analysis was first done on these variables to examine them for outliers and trending. Both evidenced normality of distribution (N=89 for each; lack-of-fit test: $p=0.16$ (NSEI) and $p=0.72$ (mBEI)).

A linear model was fit for mBEI with the independent variables NSEI, NPOL number, and the interaction term NSEI*NPOL ($p<0.0001$). Multiple lines were fit, one for each NPOL number (adjusted $R^2 = 99.9$). Leverage plots and ‘actual by predicted’ plots were also generated to confirm the fit. Plotting residuals evidenced no abnormalities in distribution of error.
To demonstrate that the NPOL lines could be statistically differentiated on the mBEI (nBEI) by NSEI plot, an analysis of covariance was performed with SAS PROC GLM.

The NSEI*NPOL interaction term was significant (p<0.0001) and all contrasts were significant (largest p=0.0097) except for lines 13 vs 14 (p=0.2561). Factors that contribute to these variables were examined visually in t-test graphics to discover any factors that might drive the lack of difference. Most of these factors displayed a marked similarity between groups. No p-values were generated due to the low power driven by small numbers in each group. In certain groups, only one or two elements (groups of target:ligand complexes having only one or two representatives) were present and those were excluded to avoid singularities.

Note of clarification. The statistical analysis was not performed to prove the validity of the linear relationships that had been proven to be correct by the algebraic analysis presented in Box S1. The interest was mainly to see the effect of the intercepts in the total variance. In addition, it should be noted that the distribution of Ks in the PDBBind refined set is selected so that it is normally distributed [7, 8].
Supplementary Material III: Applications

1. Trajectories in drug-discovery space as illustrated for human Protein Tyrosine Phosphatase 1B (Figs. S2a-d).

Fig. S2a.

Fig. S2a. Mapping of compounds for hPTP1B in the NSEI-nBEI plane. Overview of the location and distribution of the hPTP1B:ligand complexes in the 2007 general release of PDBBind (63 compounds). Different regions of the map are outlined where different compound series from different laboratories map. The color coded lines correspond to compounds with 4-16 polar atoms as indicated in the upper right-hand panel. A few chemical structures have been illustrated for reference. Data from references [33-37] and PDBBind [7, 8]. Two hypothetical chemical series trajectories for two different chemical series from one laboratory are shown. Green corresponds to the
non-peptidic Isothiazolidinone (2cn\* PDB entries) series. Blue traces a possible path for the Isothiazolidinone phosphonate mimetics (2cm\* PDB entries) series of Ala et al [38, 39]. The compounds furthest to the right have been shown for reference. Direction of the arrows is purely hypothetical.

**Fig. S2b**
Figs. S2b-d. Examples of possible chemical-series trajectories for hPTP1B in various regions of the NSEI-nBEI map and different scales.
Fig. S2b. Benzotriazole series.

Close-up view of the region of the map corresponding to the benzotriazoles series of Scapin et al. SAR tables adapted from [33]. The chemical structures of the core compounds are also shown. The PDB entry codes corresponding to the hPTP1B:ligand structures are shown. The direction of the arrow heads is purely arbitrary and does not indicate order of design or synthesis. NPOL atoms of each line are indicated.

Fig. S2c. Oxalyl-aminobenzoic series.

As above for the oxalyl-aminobenzoic acid series. The arrows pointing in opposite directions after PDB entry 1ecv are shown to suggest alternative paths or different outcomes. Data from references [34, 35]

Fig. S2d. Sulfamic acid series.

As above for the sulfamic series data reported by Klopfenstein et al. [40]. The illustrated outcome of this hypothetical trajectory with a sulfamic acid scaffold would be an efficient compound per dalton but rather polar in nature (NPOL=10; low NSEI ~ 0.5 value).

2. Mapping of compounds with inhibitory and pharmacological activity for the HIV-I protease (Figs. S3a-c).
Fig. S3a

Fig. S3b.
Fig. S3c

Figs. 3. Mapping of the HIV protease:ligand complexes in the NSEI-nBEI plane.

Fig. S3a. The complexes available in the PDBBind general release (2007). Close up of the position of the complexes for the PDB entries corresponding to the most significant drugs in the market for the HIV protease. (Saquinavir\textsuperscript{TM}, Ritonavir\textsuperscript{TM}, Liponavir\textsuperscript{TM}, etc) and the additional milestones chemical series (Tripanavir\textsuperscript{TM}, Bicyclic Ureas: DMP323) are marked. The diagram illustrates the relative position of all the drugs in the market and the migration towards higher efficiency in size and polarity for the different drugs from the first peptide-mimetic Saquinavir\textsuperscript{TM} to the latest Darunavir\textsuperscript{TM}. The latter compound was designed to interact with the main chain of native HIV-I in an attempt to overcome the development of resistance by amino acid mutations. The different region of the space occupied by the non-peptido mimetics Tripanavir\textsuperscript{TM} and bicyclic ureas (DMP323) is apparent.
**Fig. S3b. View of the CBS related to the HIV-I target.**

A limited view of the vast area of chemical space showing some activity against the HIV-I protease target. The view is based on the 2,523 entries available from the ChEMBL database plotted using nBEI-NSEI. The database comprises literature reports for which some inhibitory activity against the target has been reported. It spans from the fullerenes analogs (NPOL=1) in the far right to the extended peptide (DP) in the far left with NPOL=34. The marketed drugs illustrated in **Fig. 3a** are marked by their initials with some structures shown. The different points with the same initial (i.e., S, D, A, etc) represent measurements of the same compound against different mutants of the HIV-I protease. Normally, the highest efficiency is against the native enzyme but certain variants might have been directed towards specific mutants. The different color for each target:ligand complex relates to the number of violations of Ro5 as indicated. As discussed, the polarity of the compounds varies counterclockwise from the right (very hydrophobic, predominantly blue-green colors) to the left (very polar, red). This trend is also related to the values of LogP of the Ro5; however, the additional factor of size (MW>500) also affects the number of violations noted (2-4). The fullerene compound is off the scale. A complete view of the same data is presented in **Fig. S3c**.

**Fig. S3c.**

As before showing the entire area of chemical structures showing some activity against the HIV-I protease target (from ChEMBL). The data show that i) the region of CBS where the marketed drugs for HIV-I are located is relatively small; ii) the marketed drugs do tend to optimize the values of the corresponding LEIs (NSEI:1-1.8; nBEI:11-13).
Haloperidol, a well known CNS drug, also showed some weak activity against this target and is marked for comparison. DA marks an example of variants of Darunavir that mapped in a separate ‘island’ (see Fig. S3b). Smet refers to position of a sulfomethoxazole inhibitor and DMP450 is a different bicyclic urea related to DMP323 highlighted in Fig. S3a.

**Fig. S3c inset. Optimization of the chemical series towards Tripanavir.** Analogs and variants of Tripanavir available from BindingDB are shown in their location in the plane in relation to their chemical compositions (NPOL atoms: 6-8). TPVA# refers to the analog number; TPV# refers to variants. Arbitrarily, TPV0 (upper right-hand corner) labels Tripanavir at the position where the two variables are maximized. Inset data from BindingDB [41].
REFERENCES


