DOI: 10.1002/cmdc.200600168

In Silico ADMET Traffic Lights as a Tool for the Prioritization of HTS Hits

Mario Lobell,* Martin Hendrix, Berthold Hinzen, Jörg Keldenich, Heinrich Meier, Carsten Schmeck, Rudolf Schohe-Loop, Tobias Wunberg, and Alexander Hillisch[a]

The need for in silico characterization of HTS hit structures as part of a data-driven hit-selection process is demonstrated. A solution is described in the form of an in silico ADMET traffic light and PhysChem scoring system. This has been extensively validated with in-house data at Bayer, published data, and a collection of launched small-molecule oral drugs.

Introduction

Table 1. Analysis of the hit-to-lead process for Bayer screening projects: similarity analysis.

	Similarity[b] (Tanimoto)	Δ MW	Δ CLOGP	Δ PSA	Δ (# Rot Bonds)	Δ (# Rings)
57 hit lead pairs[a]	82.3 %	+12.4	+0.18	+0.09	+0.03	+0.18

[a] Including identical pairs. [b] Calculated with UNITY fingerprints (TRIPOS Inc.).

The introduction of ultrahigh-throughput screening (uHTS) assays that generate >100000 data points per day has created an enormous flood of primary biological data. Furthermore, compound collections in the pharmaceutical industry are steadily growing. A survey[1] conducted in 2003 covering 51 screening departments in the pharmaceutical industry revealed an average library size of nearly 1 million compounds. At Bayer, routinely more than 1.7 million compounds are screened against 30–40 targets per year, resulting in comprehensive hit lists that frequently require prioritization for subsequent processing.

The smaller hit lists of the pre-uHTS era were typically evaluated by medicinal chemists based on their experience and intuition, in addition to evaluations of potency and in some cases selectivity against other related targets. A recent systematic study[2] has confirmed the lingering suspicion that such decisions might lack consistency between different medicinal chemists. In many cases the most potent compound, but not necessarily the one with the best overall profile, is selected for further optimization. Therefore, many projects have failed because of problems inherent to the selected chemotype. Thus, despite its importance for assessing hits, "potency often turned out to be a false prophet".[3] Recognizing the modern hit-to-lead process as the multidimensional optimization problem that it is in times of heightened awareness regarding patient safety and convenience, a balanced prioritization that considers in vitro pharmacological and ADMET-related data offers the best chance for the identification of those hits that can truly be turned into drugs.

The need for an improved hit evaluation process is highlighted further by the observation that hits, leads, and corresponding development candidates are structurally closely related.[4] Addressing the early discovery phase, 57 Bayer in-house screening hits and the resulting leads were analyzed with regard to similarity and changes in calculated physicochemical parameters resulting from the optimization (Table 1). Notably, for 24 (42 %) of these projects, the hit and corresponding lead were identical (that is, either truly identical, an active enantiomer, or a salt modification). Furthermore, even for those projects in which a hit optimization was performed, only minor changes were made, and calculated physicochemical parameters remained in the same range.

The relationship between leads and the corresponding development candidates, development compounds, or drugs has also been studied at Bayer and discussed by other authors[5,6] (Table 2). From these studies, a consensus has emerged that observed changes in calculated physicochemical parameters are mostly small and structural similarity is mostly high.

Based on the above findings of structural similarities between hits, leads, and development candidates, hit prioritization and selection becomes a very critical stage in the drug-discovery process. Liabilities and chances-of-development can-

[a] *Dr. M. Lobell, Dr. M. Hendrix, Dr. B. Hinzen, Dr. J. Keldenich, Dr. H. Meier, Dr. C. Schmeck, Dr. R. Schohe-Loop, Dr. T. Wunberg, Dr. A. Hillisch*
Bayer HealthCare AG, PH-GDD-EURC-CR
Aprather Weg, 42096 Wuppertal (Germany)
Fax: (+49) 202-364490
E-mail: mario.lobell@bayerhealthcare.com

Supporting information for this article is available on the WWW under http://www.chemmedchem.org or from the author: includes a file with 812 chemical entities in SDF format.

Table 2. Comparison of leads and resulting candidates for clinical development (I): similarity analysis.

Source	Sample Size (Pairs)	Similarity[e] (Tanimoto)	Ø MW Lead	Δ MW	Ø CLOGP Lead	Δ CLOGP	Δ PSA	Δ (# Rot Bonds)	Δ (# Rings)
Bayer	45	63.6%	421	35.1	2.49	0.06	7.90	0.41	0.20
Proudfoot[5] (1)	25[a]	73.4%	313	20.1	2.47	0.01	−1.52	0.40	0.28
Proudfoot[5] (2)	22[b]	69.7%	318	22.8	2.38	0.01	−1.73	0.45	0.32
Oprea et al.[6] (1)	30[c]	57.9%	324	44.3	2.50	0.64	0.10	1.67	−0.07
Oprea et al.[6] (2)	60[d]	58.7%	300	79.8	0.53	0.53	11.62	2.00	0.55

[a] All data. [b] Same data set but excluding identical pairs of lead/drug. [c] Lead is a drug. [d] Lead is no drug itself. [e] Calculated with UNITY fingerprints (TRIPOS Inc.).

didates in many cases already reside in the corresponding hits and leads.

Several issues have to be addressed in this context:

- With the screening of growing compound collections, greater numbers of structural classes will be identified and are considered worth following up. What are the criteria to select the most interesting compound class?
- ADMET liabilities are more difficult to address in lead-optimization projects than is insufficient potency.[7] Which molecular features are needed to qualify a weak hit for further optimization?
- In a traditional hit-to-lead process, only few compound classes were followed up in a rather sequential fashion. Will a more parallel process in hit selection be superior to the traditional approach? Will such a workflow better exploit the investments made into compound repositories and uHTS technologies?

Prerequisite for addressing all these issues optimally is a data-driven hit-selection process that replaces the traditional potency- and intuition-dependent selection with a multifactorial analysis that covers ADMET-related parameters as well. This prioritization and selection process has the form of a screening cascade that involves in silico, in vitro, and in vivo aspects (Figure 1).[4] An important first step in this cascade is the in silico characterization of validated HTS hit structures. For this purpose, a system of in silico ADMET traffic lights (TLs) and the in silico oral PhysChem score have been developed.

Other useful applications of these in silico tools include the profiling of real and virtual compound libraries and the filtering and prioritization of compounds for purchasing. Herein we present a detailed description of this traffic light and score system and its validation with appropriate data sets.

Results and Discussion

A full set of nine TLs (including metabolic clearance TL, hERG inhibition TL, CYP inhibition TL, and undesirable groups TL) was introduced[4] previously. The focus of this paper is on the first five TLs and their assembly into an in silico oral PhysChem score (Tables 3 and 4) that addresses oral absorption from the gastrointestinal tract. The selected five TLs can be reliably calculated with little error. They are directly linked to physicochemical compound properties, hence the name PhysChem score. The remaining TLs, on the other hand, are more limited with regard to their prediction accuracy and coverage of chemical space. They are intended as individual alerts, highlighting the need for experimental verification (microsomal metabolism, hERG inhibition, CYP inhibition, stability, reactivity, toxicity) or

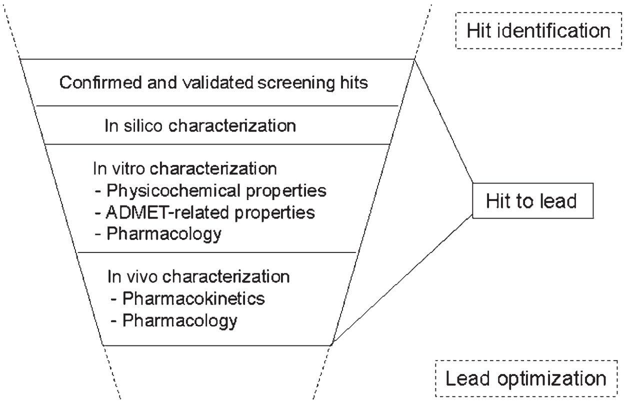
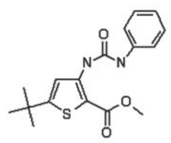
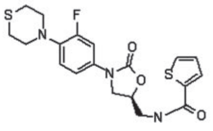

Hit identification
Confirmed and validated screening hits
In silico characterization
In vitro characterization
- Physicochemical properties
- ADMET-related properties
- Pharmacology
In vivo characterization
- Pharmacokinetics
- Pharmacology
Hit to lead
Lead optimization

Figure 1. Evaluation cascade addressing pharmacological, physicochemical, and ADMET parameters.

Table 3. In silico ADMET TLs and oral PhysChem score for two Bayer lead compounds that have been successfully optimized into a marketed drug and a clinical candidate.[8,9]

Structure	Oral PhysChem Score	TL Solubility [mg L−1]	TL CLOGP	TL MWcorr	TL PSA [Å2]	TL Rot Bonds
	4	🟥 <10	🟥 5.1	🟩 332.4	🟩 67.4	🟩 5
	2	🟨 10-50	🟩 2.6	🟨 407.7	🟩 61.9	🟩 5

replacement (sufficiently undesirable groups). In silico ADMET TLs are valuable for identifying possible liabilities of prototypic hits very early in the hit-to-lead process. This allows the focused experimental assessment of these putative drawbacks, thus substantiating the evaluation of early hits significantly. The primary purpose of the PhysChem scoring system is to assist medicinal chemistry experts in the prioritization of HTS hit sets. These in silico tools are thus not supposed to be automatic filters to remove compounds before evaluation by experts.

For a compound to be transcellularly absorbed from the gastrointestinal tract into systemic circulation it has to fulfill several requirements. It must be reasonably soluble in aqueous solution (TL solubility) but should not be too polar[10] (TL PSA), too lipophilic[11] (TL CLOGP), too large[11] (TL MWcorr) or too flexible[12] (TL rotatable bonds) to pass cellular membranes. These five aspects are well known to medicinal chemists and can be addressed by structural modifications. We have thus built our scoring system on these parameters. Each traffic light color (green, yellow, and red) corresponds to a TL value, as indicated in Table 4. The TL values for all TLs are summed up to derive the in silico oral PhysChem score (Table 3). This can adopt a minimum value of 0 and a maximum value of 10. The lower the value, the more favorable the in silico evaluation of a compound's physicochemical properties in serving as a lead for the discovery of an orally administered drug. This in silico oral PhysChem score fulfils its purpose if it differentiates between oral drugs and the average HTS hit, thus providing a means to select those hits with the best chances to be developed into a drug. Ideally such differentiation should be better than what is delivered by current and established in silico methods such as Lipinski's "rule of five",[11] which is probably the best-known in silico method for labeling compounds that have a low probability of serving as successful leads for an orally administered drug. The rule of five comprises four sub-rules (MW < 500, CLOGP < 5, sum of NH and OH groups ≤ 5, sum of N and O atoms ≤ 10). A compound violates Lipinski's rule of five if at least two of these sub-rules are violated.

Table 4. In silico ADMET TL values/colors and corresponding property values.

TL Value	TL Solubility [mg L^{-1}]	TL CLOGP	TL MWcorr	TL PSA [Å^2]	TL Rot Bonds
0	≥50	≤3	≤400	≤120	≤7
1	10-50	3-5	400-500	120-140	8-10
2	<10	>5	>500	>140	≥11

Figure 2 shows that 812 orally formulated, launched, small-molecule drug chemical entities (CEs, see Computational Methods below) have an in silico oral PhysChem score distribution entirely different from that of confirmed HTS hits. Launched oral drug CEs show an exponential decline of the population for increasing scores (average score = 1.9), whereas HTS hits show a Gaussian-like distribution of scores peaking at a score of 5 (average score = 4.1). Only 9.6% of launched oral drug CEs, but as many as 45.8% of confirmed HTS hits have an in silico oral PhysChem score of 5 or higher. However, applying Lipinski's rule of five, 9.1% of launched oral drug CEs, yet only 23.7% of confirmed HTS hits violate this latter rule. Thus the "in silico oral PhysChem score rule of five" allows better differentiation than Lipinski's rule of five. The in silico oral PhysChem score, with its range from 0 to 10, also gives a finer grading for compound prioritization than does Lipinski's rule of five, which gives a simple yes or no answer.

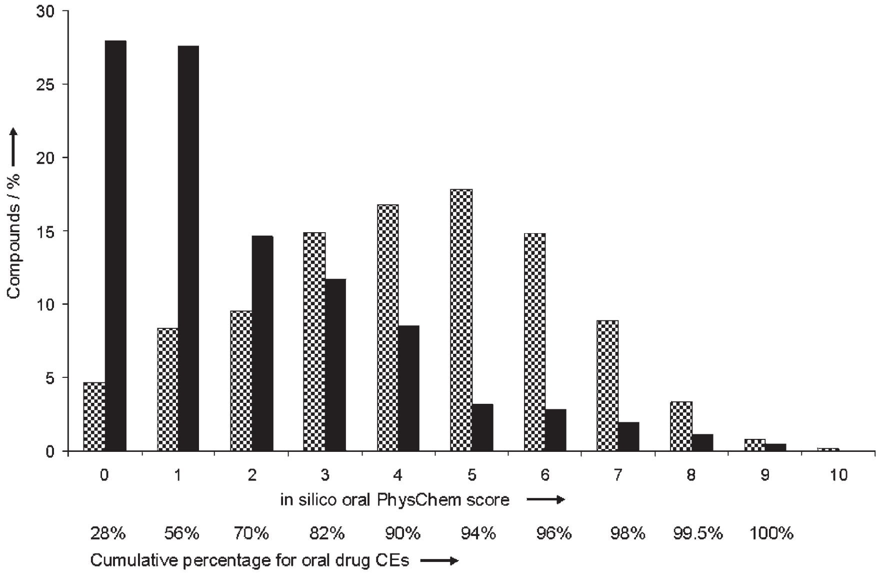


Figure 2. In silico oral PhysChem score distribution (in percent) of 13 775 confirmed Bayer HTS hits (checkered bars) from 20 screens (11 cell-based, 9 biochemical) compared with 812 unique small-molecule chemical entities (CEs), derived by chemical synthesis and present in launched drugs with oral formulation (black bars). Orally formulated, launched drugs show an exponential decline of the population for increasing scores (average score = 1.9). HTS hits show a Gaussian-like distribution of scores peaking at a score of 5 (average score = 4.1). 90% of oral drug CEs have an in silico oral PhysChem score of < 5.

It is clear from Figure 2 that a low in silico oral PhysChem score should be preferable for an oral drug candidate and enhance its chances to make it all the way to market approval. While arguably a lead-optimization process can significantly improve physicochemical and ADMET properties, more typically, parameters such as molecular size and lipophilicity, and consequently oral PhysChem scores tend to increase during lead

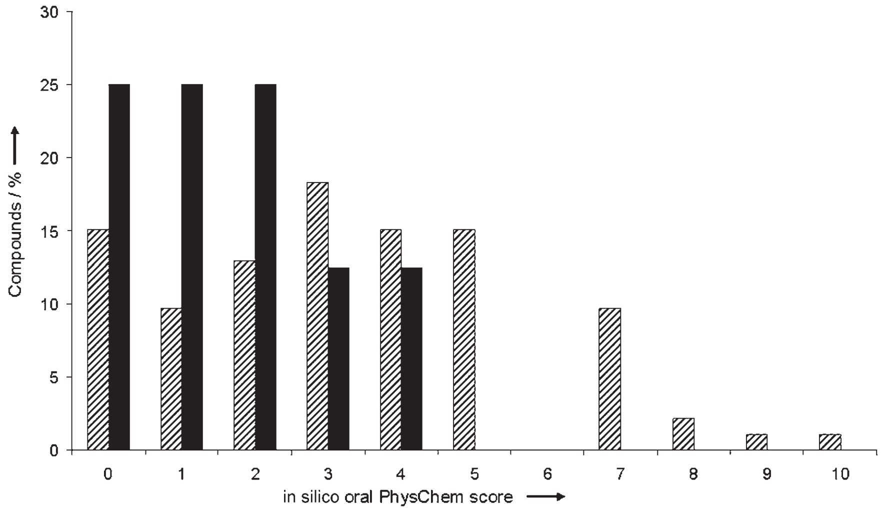


Figure 3. In silico oral PhysChem score distribution (in percent) of eight Bayer drugs (black bars; vardenafil, moxifloxacin, ramatroban, ciprofloxacin, clotrimazole, emiglitate, vibunazole, nifedipine) compared with 93 development compounds (striped bars) which had reached development compound status during the same time frame but did not achieve market approval. Low scores dominate for eight drugs (average score = 1.6), whereas low scores are less frequent; medium to high scores are more frequent for the 93 development compounds without market approval (average score = 3.3).

optimization, as has been documented in the literature[3,6] and which is further demonstrated in Tables 1 and 2. An analysis of in-house small-molecule development compounds (Figure 3) also indicates that low in silico oral PhysChem scores confer a higher probability to ultimately succeed as a launched drug.

As always, there are exceptions to the rule, and it is wise to consider the particular circumstances of a disease area and target protein class.[13] Prime examples of an exception are HIV protease inhibitors. The outlier drug with an oral score of 8 in Figure 4 is lopinavir, which is marketed as Kaletra (a co-formulation of lopinavir and ritonavir (oral score = 9)). Kaletra ranked 124th among the 200 top selling prescription drugs in the US in 2002. The average in silico oral PhysChem score of all nine launched oral HIV protease inhibitor CEs is 7.3 (indinavir (oral score = 5, 60 % bioavailability), nelfinavir (score = 5, 20–80 % bioavailability), saquinavir (score = 7, 4 % bioavailability), amprenavir (score = 7, 35–90 % bioavailability), fosamprenavir (score = 7), tipranavir (score = 8, limited absorption), lopinavir (score = 8), ritonavir (score = 9, 70 % bioavailability), and atazanavir (score = 10)). HIV protease, with its large lipophilic tunnel-shaped peptide binding site, requires relatively large, lipophilic compounds for efficient inhibition. This is reflected by the high molecular weight and log *P* values (and thus higher in silico oral PhysChem scores) of these compounds. As HIV protease inhibitors are life-saving drugs, suboptimal pharmacokinetic properties can be tolerated.

Other outlier classes include macrolide antibiotics (ricamycin (oral score = 7) and josamycin (score = 7)), and oral cephalosporin antibiotics (cefteram pivoxil (score = 7) and cefditoren pivoxil (score = 7)). These compounds are absorbed through active-transport processes and can therefore compensate for negligible passive intestinal absorption. The oral anti-obesity drug orlistat (oral score = 7) is not absorbed into the bloodstream; it inhibits lipase in the intestinal tract, thereby decreasing the generation and absorption of fatty acids.

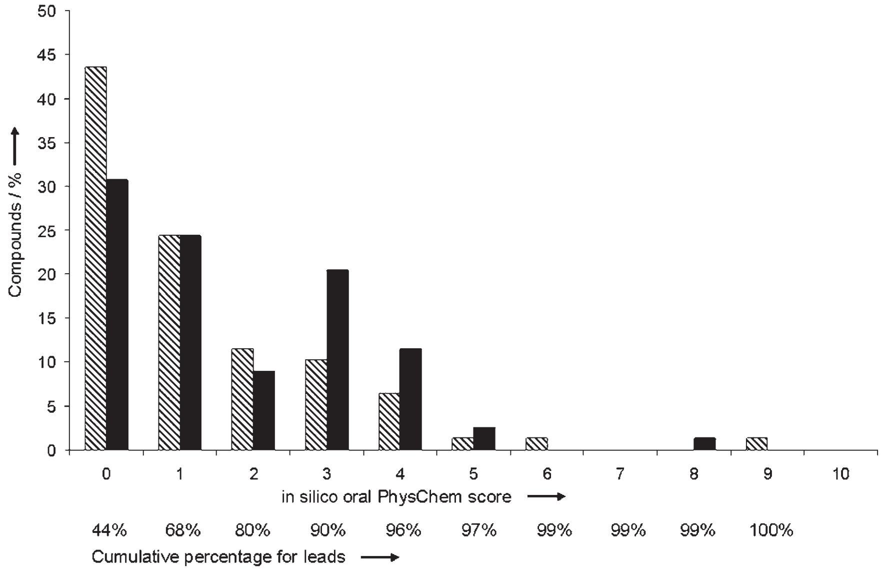


Figure 4. in silico oral PhysChem score distribution (in percent) of leads[5,6] (striped bars) compared with their corresponding launched drugs (black bars). Low scores dominate for both. Leads have a slightly lower average score (1.3) than their corresponding drugs (1.7). 90 % of leads have an in silico oral PhysChem score of < 4.

Conclusions

A system for the in silico ADMET characterization of structures was developed, tested, and validated as a tool for HTS hit prioritization. These in silico predictions are easy to interpret owing to the intuitive traffic light color scheme and a scoring system in the convenient range of 0–10. The selected five traffic lights contribute equally to the in silico oral PhysChem score. This makes the scoring system transparent and easily understood, and contributes to a high acceptance rate by medicinal chemists. Intestinal metabolism and efflux are absorption-relevant factors that are not addressed by the traffic light system presented herein. Although it is desirable to have TLs that predict liability to intestinal metabolism and efflux, we do not presently have sufficient experimental data to serve as a training set for method development. Moreover, the effect of intestinal metabolism and efflux on the overall absorption is highly variable, as these processes have a limited capacity and can be easily saturated and overrun if the concentration of dissolved compound in the gut and the rate of passive absorption are sufficiently high. The primary purpose of the in silico ADMET traffic light and scoring system is to aid medicinal chemistry experts in the analysis and prioritization of HTS hit sets with regard to potential ADMET-related liabilities. Other important applications include the profiling of real and virtual compound libraries and the filtering and prioritization of commercially available compounds intended as additions to the HTS compound repository. For validation, the score distributions of a representative collection of Bayer HTS hits was compared with collections of successful leads and launched drugs. The majority of successful leads and launched drugs have low in silico oral PhysChem scores, whereas HTS hits peak at medium scores. At Bayer Pharma, the in silico ADMET traffic light and scoring system is fully integrated into a data-driven hit-selection process that includes in silico and experimental procedures (see Figure 1). The success of this data-driven process will ultimately have to be measured by the future output of development compounds and their success in the clinic.

Computational Methods

A detailed description of the traffic lights is given in the following paragraphs.

TL MWcorr

The diffusion of molecules in water or membrane phases is dependent on their molecular volume. In the same phase, molecules with a large volume have diffusion rates lower than those of molecules with a smaller volume. Molecular volume therefore influences the rate and amount of oral absorption and tissue distribution. Molecular volume can be calculated from a 2D structure as McGowan volume (MV).[14] However, chemists are accustomed to working with molecular weight (MW) as a surrogate description of molecular size. MW correlates very well with MV ($MW = 1.336 \times MV$, $R^2 = 0.943$) if only organic molecules devoid of halogen atoms are taken into account. Halogen atoms are exceptional because they have a nonproportional small volume for their atomic mass. Molecules with halogen atoms can be included if appropriate mass corrections are introduced to yield the corrected molecular weight (MWcorr). The optimal mass corrections for halogen atoms were determined as follows:

MW, MV, and the number of halogen atoms were calculated for a training set of 3291 Bayer compounds (a subset of neutral compounds in a solubility training set of 4806 Bayer compounds). 1677 compounds did not contain any halogen atoms. These 1677 halogen-free compounds were used to perform a linear regression (Microsoft Excel). The regression line was forced through the origin (0,0) to obtain the linear equation $MW = 1.3361 \times MV$ ($R^2 = 0.943$). The MV contribution of a fluorine atom is 3.92. This was derived from the MV difference of monofluorobenzene (73.41) and a phenyl radical (69.49). Similarly, the MV contributions of a chlorine (14.39), bromine (19.65), and iodine atom (27.97) were calculated. To conserve the above linear relationship with MW (obtained for halogen-free compounds) the following corresponding MW contributions should ideally apply: $MW(F) = 1.3361 \times 3.92 = 5.24$, $MW(Cl) = 1.3361 \times 14.39 = 19.23$, $MW(Br) = 1.3361 \times 19.65 = 26.25$, and $MW(I) = 1.3361 \times 27.97 = 37.37$. The corresponding real MW contributions are 19.01 (F), 35.46 (Cl), 79.92 (Br), and 126.91 (I). Therefore, for each fluorine atom, one needs to subtract $19.01 - 5.24 = 13.8$ MW units from the real MW to derive MWcorr. Similarly, one must subtract $35.46 - 19.23 = 16.2$ MW units for each chlorine atom, $79.92 - 26.25 = 53.7$ MW units for each bromine atom, and $126.91 - 37.37 = 89.5$ MW units for each iodine atom. Figure 5 illustrates the effect of the applied corrections. The margin $MWcorr = 500$ for the red TL color was inspired by Lipinski's rule of five.[11]

TL CLOGP

Lipophilicity is often expressed as log *P*, the logarithmic partition coefficient of the molar distribution of a compound between equal volumes of *n*-octanol and water. The TL CLOGP calculates log *P* by the fragment summation method CLOGP (version 4.2 with version 22 of its associated fragment database, as implemented in Sybyl version 7.0, Tripos Inc.). The TL CLOGP reflects the association of increases in log *P* value with decreased aqueous solubility, increased plasma protein binding, and increased susceptibility to oxidative metabolism and CYP inhibition. The margin $CLOGP = 5$ for the red TL color was again inspired by Lipinski's rule of five.[11]

TL solubility

Aqueous solubility is of major importance for the ability to develop appropriate oral and intravenous formulations for drugs and drug candidates. As a first approximation, conditions in the small intestines of humans and laboratory animals can be mimicked by a buffer solution at pH 6.5. Sufficient solubility in this medium is therefore usually required to avoid solubility-limited absorption of drug candidates which, as a consequence, cannot reach the required oral exposure in safety studies or during clinical development. At our laboratories, the solubility of compounds in aqueous buffer at pH 6.5 (containing 1% DMSO) is investigated by a precipitation method.[15] A training set of 4806 Bayer in-house compounds had been assembled for which accurate solubility measurements could be obtained.[15] Solubility *S* is expressed in $mol\,L^{-1}$, which is standard procedure for the development of in silico solubility prediction methods. A set of descriptors (described in Table 5 and Figure 6) was calculated with Sybyl (version 7.0, Tripos Inc.) by

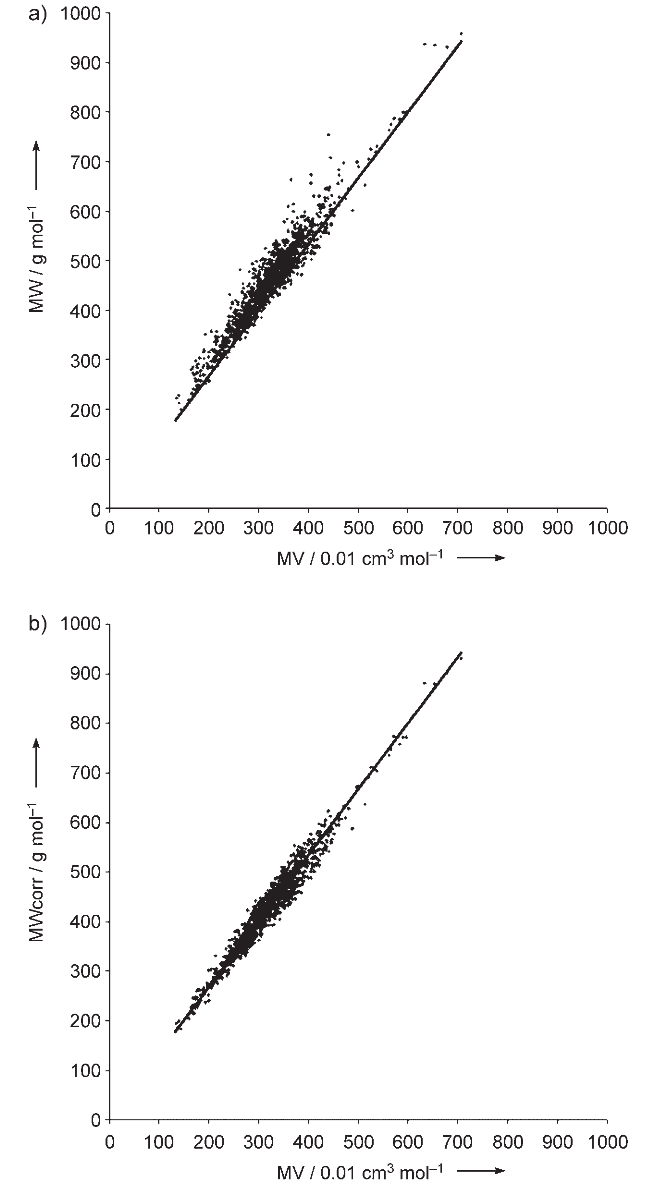


Figure 5. a) MW and b) MWcorr are plotted versus McGowan molecular volume (MV) for 1614 compounds that contain one or more halogen atoms; the regression line MWcorr = 1.336 × MV is shown.

performing a combination of property calculations (CLOGP, MW), substructure queries (SLN (Sybyl Line Notation) queries), mathematical and logical operations (NOT, AND, OR, IF THEN ELSE) implemented through custom written scripts in SPL (Sybyl Programming Language).

Based on the training set, a computational prediction algorithm for log *S* was developed by stepwise multiple linear regression analysis applied to the calculated descriptors using Cerius2 (version 4.8, Accelrys Inc.). Prob(F) is the statistical probability that the observed improvement of the model by inclusion of the descriptor is derived by chance. The standard Prob(F) threshold of 0.05 was applied (equivalent to the 95% confidence level). The largest actual Prob(F) value for an included descriptor was 4.2×10^{-10}. Table 5 shows the final descriptor components, which form part of the derived model equation, and Table 6 shows the statistical parameters.

Table 6. Statistical parameters from stepwise linear regression analysis.

Parameter	Value
R	0.62
R^2	0.39
F-test	430
mean absolute error	0.63

The first four descriptors of the derived solubility equation (Table 5) relate to the predicted charge state of the compounds at pH 6.5. As a rule of thumb, the derived coefficients indicate that a solubility increase of one log unit is observed if a compound is mostly ionized at pH 6.5, as opposed to uncharged. The descriptor "aromatic proportion" is related to crystal stability. The aromatic parts of the molecular structure are flat and easy to stack up in a dense crystal lattice, often without the need to adopt a specific conformation for stacking, as is the case for the flexible, nonaromatic parts of the structure. Highly aromatic compounds therefore tend to form very stable crystals with high melting points. The predictive capacity of the descriptor "aromatic proportion" for aqueous solubility was previously demonstrated by Delaney.[16] The inclu-

Table 5. Solubility equation descriptor components derived from stepwise linear regression analysis.

Descriptor Included	Descriptor Name	Definition
positive coefficient	Zwitterionic	= 1 if compound is predicted to be zwitterionic at pH 6.5, otherwise = 0
positive coefficient	Positively charged not ZW	= 1 if compound is predicted to be positively charged at pH 6.5 but not zwitterionic, otherwise = 0
positive coefficient	Any COOH, PO_3H, SO_3H, not ZW	= 1 if compound has a COOH, PO_3H, or SO_3H acidic group but is not predicted to be zwitterionic at pH 6.5, otherwise = 0
positive coefficient	Any Tetrazole not COOH, PO_3H, SO_3H, ZW	= 1 if compound has a tetrazole acidic group but not a COOH, PO_3H, or SO_3H acidic group and is not predicted to be zwitterionic at pH 6.5, otherwise = 0
negative coefficient	Aromatic proportion	(number of aromatic atoms)/(number of heavy atoms)
negative coefficient	CLOGP	CLOGP (as described above)
negative coefficient	MWcorr	MW corrected for halogen atoms (as described above)
not included	Vinylogous acid	= 1 if compound has a vinylogous acidic group but not a tetrazole, COOH, PO_3H, or SO_3H acidic group and is not predicted to be zwitterionic at pH 6.5, otherwise = 0
not included	Rot Bonds	number of freely rotatable bonds
not included	MW	molecular weight

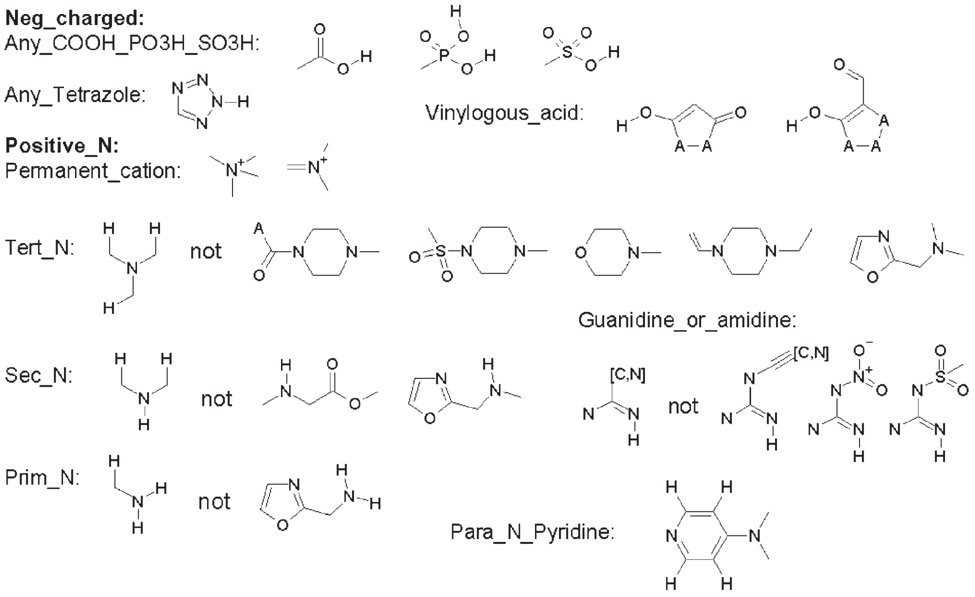


Figure 6. Substructure queries for the calculation of charge descriptor flags; a compound is predicted to be zwitterionic if Neg_charged = 1 and Positive_N = 1.

sion of MWcorr as a relevant descriptor might be related to the fact that larger compounds also tend to have higher melting points and form more stable crystals. The inverse relationship of the lipophilicity descriptor CLOGP to aqueous solubility is intuitive and has been observed and reported many times.[16,17]

80% of the training set compounds are predicted with an error below one log unit, and 99% with an error below two log units. The mean absolute error is 0.63 log units. For a validation set of 7990 Bayer in-house compounds, 68% are predicted with an error below one log unit and 98% with an error below two log units. The mean absolute error for the validation set is 0.76 log units. The 7990 compounds of the validation set have also been binned according to the three solubility traffic light colors (green: ≥50 mg$\,L^{-1}$, yellow: 10–50 mg$\,L^{-1}$, and red: <10 mg$\,L^{-1}$) to compare the bin distributions from predicted and experimental solubility values. Table 7 demonstrates that the accuracy of the developed solubility prediction algorithm is sufficient for the purpose of the classification into three traffic light colors. 77% of highly soluble compounds are correctly identified as green, whereas 92% of the low-solubility compounds are correctly classified as red. The solubility margin of 50 mg$\,L^{-1}$ for the green traffic light was inspired by our own in-house experience with the development of oral drugs, but is also in line with the comments made by Lipinski et al.[11]

Table 7. Validation of TL solubility with 7990 Bayer compounds.

No. Compds	Solubility [mg$\,L^{-1}$][a]	Compds Predicted [%]		
		≥50 mg$\,L^{-1}$	10–50 mg$\,L^{-1}$	<10 mg$\,L^{-1}$
2790	≥50	77	9	14
924	10–50	37	19	44
4276	<10	5	3	92

[a] Experimentally determined.

TL PSA

PSA (polar surface area) is defined as the sum of the topological surface area in Å^2 of all nitrogen and oxygen atoms plus any attached hydrogen atoms. It is known that excessive hydrogen bonding of a compound's polar atoms to water molecules can generate strongly bound hydration shells that impair the ability of the compound to enter the lipophilic phase of cell membranes. Palm et al.[10] and later Clark[18] were the first to demonstrate the usefulness of this descriptor for the prediction of poor intestinal absorption via the dominant passive absorption route. Both derived a PSA threshold value of 140 Å^2, which should not be surpassed if poor passive absorption is to be avoided. This threshold value was later confirmed by Veber et al.[12] in a study on oral bioavailability in rats. Kelder et al.[19] analyzed the PSA distribution of 1590 orally administered non-CNS drug candidates that have reached at least Phase II efficacy studies. They found that orally active drugs which are transported passively by the transcellular route should not exceed a PSA of about 120 Å^2. We defined the color thresholds of the TL PSA (Table 4) in accordance with the above published findings. Although PSA is a 3D property, it has been shown that it can also be calculated from the 2D structure with little loss of accuracy.[20] The widely used PSA calculation method of Ertl et al.[20] was implemented by using SPL in Sybyl. In contrast to the original method of Ertl et al., we did not include sulfur and phosphorus as polar atoms.

TL rotatable bonds

The number of rotational bonds is a reflection of the conformational flexibility of a molecule, which in turn influences the oral bioavailability of drug candidates.[12] Problems with high conformational flexibility result from higher energy penalties (in the form of lost entropy) to be paid upon adoption of a specific conformation required to pass lipid bilayers or to bind to the intended pharmacological target. Furthermore, this energy penalty decreases the rate of transport across cell membranes and limits the achievable binding affinity to the pharmacological target (in vitro potency). The color thresholds of the TL rotatable bonds (Table 4) were defined in accordance with the results of Veber et al.[12] obtained in a study on oral bioavailability in rats. Recently, Blake[21] analyzed 6467 compounds that had been selected for clinical development and confirmed a disadvantage for compounds with more than 10 rotatable bonds. The number of rotatable bonds can be easily calculated from a compound's structure. An appropriate calculation algorithm was implemented by using SPL in Sybyl.

Collection of 1205 unique small-molecule CEs present in launched drugs

The Pharmaproject Database (PJB Publications, T&F Informa UK Ltd., London) was searched for all drug entries (up to the end of 2004) labeled as "launched". This search yielded 2838 structure entries with attached information fields. Drug CEs are often marketed in various formulations and also in combination formulations with other drug CEs. This leads to the occurrence of the same drug CE in multiple database entries. There are also many drug CEs which are not derived by conventional chemical synthesis, such as protein drugs and other drugs which are at least partly derived by biological fermentation or extraction from plant and animal material. Furthermore, there are many drug CEs that are exclusively administered through routes other than oral or injectable formulations (for example, topical skin application, inhalants, nasal sprays, eye drops). Therefore, a subsequent deconvolution step had to be included to yield a collection of unique small-molecule CEs derived by chemical synthesis and present in launched drugs with oral or injectable formulations. Details on this deconvolution step are provided in the Supporting Information.

One ambiguity is derived from the fact that oral or injectable formulation is regarded as the default and therefore not indicated unless it is a special-release formulation (for example, retarded release). Formulations other than oral and injectable formulation (topical, inhalable, nasal, etc.), on the other hand, are always indicated in the database. 1205 unique small-molecule CEs (formulation label distribution: 503 oral, 63 oral (gastrointestinal drugs), 11 oral + iv, 42 iv, 586 not indicated and assumed oral and/or injectable) were left after all the filtering steps had been performed. Oral formulation could therefore be unambiguously established for a subgroup of 577 CEs. To increase this number, we cross-referenced all CE trivial names with a database of oral launched drugs, which is available from the supporting information from a publication by Leeson and Davis[22] on time-related differences in the physical property profiles of oral drugs. This step yielded a subgroup of 812 CEs with confirmed oral formulation. The Supporting Information herein contains a file with these 812 CEs in SDF format.

Acknowledgements

We thank Gertrud Ahr for the retrieval of data on a subset of Bayer's development compounds and Hans-Georg Rohbeck for the retrieval of raw data and structures on launched drugs. Finally, we thank J. R. Proudfoot for providing us with electronic copies of the chemical structures used in his work.[5]

Keywords: computer chemistry · hit to lead · in silico ADMET · medicinal chemistry

[1] Industry Report "HighTech Business Decisions (HTBD)—High-Throughput Screening 2003: Improving Strategies, Technologies, and Productivity", http://www.hightechdecisions.com.
[2] M. S. Lajiness, G. M. Maggiora, V. Shanmugasundaram, *J. Med. Chem.* **2004**, *47*, 4891–4896.
[3] A. L. Hopkins, C. R. Groom, A. Alex, *Drug Discovery Today* **2004**, *9*, 430–431.
[4] T. Wunberg, M. Hendrix, A. Hillisch, M. Lobell, H. Meier, C. Schmeck, H. Wild, B. Hinzen, *Drug Discovery Today* **2006**, *11*, 175–180.
[5] J. R. Proudfoot, *Bioorg. Med. Chem. Lett.* **2002**, *12*, 1647–1650.
[6] T. I. Oprea, A. M. Davis, S. J. Teague, P. D. Leeson, *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1308–1315.
[7] M. MacCoss, T. A. Baillie, *Science* **2004**, *303*, 1810–1813.
[8] T. B. Lowinger, B. Riedl, J. Dumas, R. A. Smith, *Curr. Pharm. Des.* **2002**, *8*, 2269–2278.
[9] S. Roehrig, A. Straub, J. Pohlmann, T. Lampe, J. Pernerstorfer, K.-H. Schlemmer, P. Reinemer, E. Perzborn, *J. Med. Chem.* **2005**, *48*, 5900–5908.
[10] K. Palm, P. Sternberg, K. Luthman, P. Artursson, *Pharm. Res.* **1997**, *14*, 568–571.
[11] C. A. Lipinski, F. Lombardo, B. W. Dominy, P. J. Feeney, *Adv. Drug Delivery Rev.* **1997**, *23*, 3–25.
[12] D. F. Veber, S. R. Johnson, H. Y. Cheng, B. R. Smith, K. W. Ward, K. D. Kopple, *J. Med. Chem.* **2002**, *45*, 2615–2623.
[13] M. Vieth, J. J. Sutherland, *J. Med. Chem.* **2006**, *49*, 3451–3453.
[14] J. A. Platts, D. Butina, M. H. Abraham, A. Hersey, *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 835–845.
[15] A. H. Göller, M. Hennemann, J. Keldenich, T. Clark, *J. Chem. Inf. Mod.* **2006**, *46*, 648–658.
[16] J. S. Delaney, *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1000–1005.
[17] M. Lobell, V. Sivarajah, *Mol. Diversity* **2003**, **7**, 69–87.
[18] D. E. Clark, *J. Pharm. Sci.* **1999**, *88*, 807–814.
[19] J. Kelder, P. D. J. Grootenhuis, D. M. Bayada, L. P. C. Delbressine, J.-P. Ploemen, *Pharm. Res.* **1999**, *16*, 1514–1519.
[20] P. Ertl, B. Rohde, P. Selzer, *J. Med. Chem.* **2000**, *43*, 3714–3717.
[21] J. F. Blake, *BioTechniques* **2003**, *34*, S16-S20.
[22] P. D. Leeson, A. M. Davis, *J. Med. Chem.* **2004**, *47*, 6338–6348.

Received: July 10, 2006
Published online on September 22, 2006