JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

iLOGP: A Simple, Robust, and Efficient Description of *n*-Octanol/Water Partition Coefficient for Drug Design Using the GB/SA Approach

Antoine Daina,[†] Olivier Michielin,*[,†,‡,§] and Vincent Zoete*[,†]

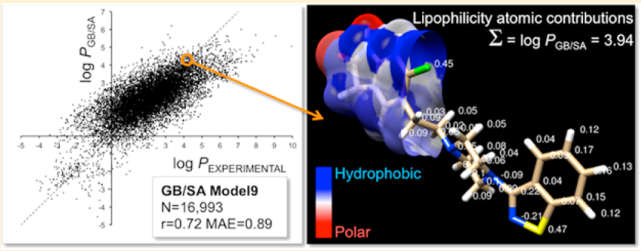
[†]Molecular Modeling Group, SIB Swiss Institute of Bioinformatics, Quartier Sorge, Bâtiment Génopode, CH-1015 Lausanne, Switzerland

[‡]Ludwig Center for Cancer Research of the University of Lausanne, CH-1015 Lausanne, Switzerland

[§]Department of Oncology, University of Lausanne and Centre Hospitalier Universitaire Vaudois (CHUV), CH-1011 Lausanne, Switzerland

Supporting Information

ABSTRACT: The *n*-octanol/water partition coefficient (log $P_{o/w}$) is a key physicochemical parameter for drug discovery, design, and development. Here, we present a physics-based approach that shows a strong linear correlation between the computed solvation free energy in implicit solvents and the experimental log $P_{o/w}$ on a cleansed data set of more than 17,500 molecules. After internal validation by five-fold cross-validation and data randomization, the predictive power of the most interesting multiple linear model, based on two GB/SA parameters solely, was tested on two different external sets of molecules. On the *Martel* druglike test set, the predictive power of the best model (N = 706, r = 0.64, MAE = 1.18, and RMSE = 1.40) is similar to six well-established empirical methods. On the *17-drug* test set, our model outperformed all compared empirical methodologies (N = 17, r = 0.94, MAE = 0.38, and RMSE = 0.52). The physical basis of our original GB/SA approach together with its predictive capacity, computational efficiency (1 to 2 s per molecule), and tridimensional molecular graphics capability lay the foundations for a promising predictor, the implicit log *P* method (iLOGP), to complement the portfolio of drug design tools developed and provided by the SIB Swiss Institute of Bioinformatics.



INTRODUCTION

Among other physicochemical properties, lipophilicity plays an important role for molecular discovery activities in a variety of domains including, but not exclusively, cosmetics, agrochemicals, material sciences, food chemistry, environmental chemistry, and especially medicinal chemistry.[1] The typical quantitative descriptor of lipophilicity is the partition coefficient *P* of a given molecule between two immiscible solvents. The *n*-octanol/water system is traditionally employed in biomedical and pharmaceutical research. Because of its amphiphilic nature, *n*-octanol is considered a good mimic of phospholipid membrane characteristics.[2] Experimentally, the molecule under investigation is solubilized in the two-solvent system at a pH assuring the neutral form of the compound. This enables to measure the equilibrium concentrations in water and in *n*-octanol, respectively. The parameter log $P_{o/w}$ is defined as the decimal logarithm of the ratio of the molar concentrations of the neutral form in *n*-octanol, C_o, and in water, C_w

$$\log P_{o/w} = \log \frac{C_o}{C_w} \quad (1)$$

Actually, log $P_{o/w}$ is not an absolute description of lipophilicity because it quantifies the partition of the neutral species of molecules solely, in contrast with the distribution coefficient (log *D*) that accounts for all electric species at a given pH. Nevertheless an accurate estimation of log $P_{o/w}$ is central for the discovery and development of efficacious therapeutic molecules.[3−5] Whereas lipophilicity cannot characterize the whole physicochemical nature of a compound, properties governing lipophilicity have a fundamental effect on the behavior of organic molecules, such as drugs or drug candidates. This is true when the compound is interacting with phospholipid membranes but also with proteins. It follows that log $P_{o/w}$, in combination with other parameters describing various physicochemical or molecular properties, was successfully used in *in silico* models to predict pharmacokinetics, ADME, and druglikeness profiles,[6−10] as well as pharmacodynamics and target affinity of new chemical entities.[11−14]

However, despite the good solubility of most organic compounds in *n*-octanol and ease in lab handling, the experimental determination of log $P_{o/w}$ remains a resource- and time-consuming procedure. Methods to estimate log *P* are mainly dedicated to medicinal chemistry and molecular design

Received: July 29, 2014

   dx.doi.org/10.1021/ci500467k | *J. Chem. Inf. Model.* XXXX, XXX, XXX−XXX

activities. At early steps of typical medicinal chemistry projects, the compounds are synthesized at the milligram scale and employed primarily for structure validation and activity assays. Furthermore, during computer-assisted molecular design, some molecules are only virtual because a physical sample does not necessarily exist.[15] In those cases, experimental measurement of *n*-octanol/water is not feasible, and thus, models based on chemical structure are crucial to get a log $P_{o/w}$ estimation.

Many algorithms to compute log $P_{o/w}$ have been developed and made available. The work of Mannhold and colleagues[16,17] provides a thorough description and an exhaustive benchmarking of existing log *P* predictors. The vast majority of these predictive tools rely on empirical methodologies trained on a data set of measured log $P_{o/w}$ values. Typical log *P* predictors can be divided into two categories. The first ones split molecular structures into molecular fragments (*fragmental* methods, e.g., KLOGP,[18] KOWWIN[19]) or atoms (*atomic* methods, e.g., ALOGP,[20,21] XLOGP,[22,23] and Wildman and Crippen's SLOGP[24]). The log *P* value is obtained by summing fragmental or atomic contributions; some methods add corrective factors to refine the structural description (for more details, refer to Use of Empirical Log P Predictors in the Method section). The second category gathers the *topological* methods in which the molecule is defined by descriptors related to its topology, such as the count of or flags for specific atoms, groups, or structural properties (e.g., MLOGP[25,26] or a simple count of carbon atoms and heteroatoms[16]). The prediction is given by statistical equations, usually obtained by multiple linear regressions trained on large molecular data sets. A hybrid approach proposed by Silicos-it includes both molecular fragments and topological parameters.

The main drawback of the *fragmental* and *atomic* methods to compute log $P_{o/w}$ lies in their additive nature prone to overestimate the lipophilicity of large molecules and to prevent straightforward chemical interpretation for designing or discovering molecules with desired lipohilicity. This yields an overall reduced performance as well as limited validity domain. Moreover, no fragmental system can cover the entire chemical space, which leads to the frustrating situation faced by many cheminformatician and described by Reynolds,[27] when a fragment of the molecule under investigation is not present in the fragmental system. In that case, an unsatisfactory dummy value is attributed, which lowers the confidence on the prediction as well as any prospect of mechanistic understanding. On the other hand, the most significant problem of current *topological* methods is the use of parameters lacking any physical meaning. This prevents mechanistic insights into the structural, electronic, or other properties that govern lipophilicity. Although the increasing size of available data sets—up to multiple tens of thousands of data points—enables decent predictive capacity, especially for druglike structures, the intrinsic issue of limited applicability domains remains. The predictive power of empirical models for log $P_{o/w}$ may be high for molecules similar to those used for training, yet not satisfactory for essentially new chemical structures because the prediction is largely impacted by extrapolation. The strong dependence on measured data is another important limitation shared by all empirically derived models. As stated by Arnott,[4] the common assumption that calculation of small-molecule lipophilicity is straightforward and can clearly forecast the success of compounds promoted to later stages of drug discovery programs is basically wrong.

This suggests that calculating log $P_{o/w}$ values based on the physics governing solvation could produce models (i) with increased robustness and more prone to generalization outside of applicability domain and (ii) that enable mechanistic insights into the partition effect and thus add a strong rationale to the chemical modifications required during the optimization steps of drug design processes.

For the *n*-octanol/water system, log *P* is related to the Gibbs free energy of transfer between both solvents $\Delta G^{o/w}_{transfer}$. The latter can be considered as the difference $\Delta\Delta G^{o/w}_{solv}$ of the solvation free energies in *n*-octanol, ΔG^{o}_{solv}, and in water, ΔG^{w}_{solv}

$$\begin{aligned} -RT \times \log P_{o/w} &= \Delta G^{o/w}_{transfer} \\ &= \Delta\Delta G^{o/w}_{solv} \\ &= \Delta G^{o}_{solv} - \Delta G^{w}_{solv} \end{aligned} \quad (2)$$

The most accurate method to compute the partition coefficient is theoretically the determination of ΔG^{o}_{solv} and ΔG^{w}_{solv} directly $\Delta G^{o/w}_{transfer}$ by simulations of the solute in explicit water and *n*-octanol. In practice, the pioneering works of Jorgensen,[28] Kollman,[29,30] Richards,[31] and more recent ones (e.g., ref 32) have proven excellent capacities to describe the interaction forces between solute and diverse solvents, in particular within the *n*-octanol/water system. However, all-atom molecular dynamics (MD) or Monte Carlo (MC) simulations are computationally very demanding and can hardly compete with an experimental determination of log $P_{o/w}$ when a physical sample is available. Therefore, these methods are currently not applicable to the large number of molecular structures to be handled during drug design projects.

A powerful alternative is the use of implicit solvation models that simplify the calculation of the solvation free energy by considering the solvent as an ensemble averaged continuous medium bearing the properties of the real solvent. In the PB/SA approach, the solvation free energy ΔG^{x}_{solv} for a given solvent *x* is split into electrostatic, $\Delta G^{x}_{solv,elec}$, and nonpolar, $\Delta G^{x}_{solv,np}$, contributions.

$$\Delta G^{x}_{solv} = \Delta G^{x}_{solv,elec} + \Delta G^{x}_{solv,np} \quad (3)$$

This procedure speeds up the computation to a fraction of what is required for an explicit solvent representation thanks to the nonpolar contribution generally approximated as proportional to the solvent accessible surface area (SASA) and the polar contribution defined by the Poisson−Boltzman (PB) equation. This has been extensively validated for macromolecules.[33] However, the use of PB is still too computationally demanding for very large systems or when numerous small systems have to be treated, which is indeed the case for drug design activities involving large chemical libraries. One efficient way to avoid too long computation is to employ the much faster generalized Born (GB) method[34] to approximate the solution of the PB equation by considering pairwise atomic interactions and self-contribution terms.[35] The GB model has proven to be a reasonable estimate of the electrostatic contribution to the solvation free energy.[36] Approaches combining GB with SASA to take into account electrostatic and nonpolar contributions to the Gibb's energy are termed GB/SA methods.

Only few attempts to correlate GB/SA calculations and partition coefficients have been published. Back in 1995, the chloroform/water partition was predicted for a set of 30 rigid and structurally related small organic molecules.[27] After *in vacuo* minimization within the OPLS-AA force field,[37,38] a single conformation was submitted to GB/SA computations (probably following the original description of Still[35] for the electrostatic part) in both solvents to obtain the difference of solvation free

dx.doi.org/10.1021/ci500467k | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

energies $\Delta\Delta G^{c/w}_{solv}$. A strong linear correlation ($r = 0.96$) between experimental log $P_{c/w}$ and computed $\Delta\Delta G^{c/w}_{solv}$ was found. This study validated the concept of GB/SA approach to predict log P and showed how useful for interpretation a physics-based method can be. The same research group developed two years later a GB/SA model to predict solvation free energy in *n*-octanol[39] making use of the Merck Molecular force field (MMFF).[40−44] The *n*-octanol continuum ensemble was defined with a 2 Å radius and a 10.3 dielectric constant. The computed solvation energies in *n*-octanol ΔG^{o}_{solv} were strongly correlated with experiment ($r = 0.96$) for 66 simple organic molecules; the root mean squared error (RMSE) was 0.72 kcal/mol. Submitting a single conformer to this model, and to a similar model for water, enabled the estimation of log $P_{o/w}$ as the difference of solvation free energies in both solvents. The log $P_{o/w}$ values for the set of 66 organic molecules were computed using eq 2 with a mean absolute error (MAE) of 1.16 versus experimental values. This slightly disappointing statistics were attributed mainly to the water continuum model setup (discrepancy between MMFF and OPLS parameters). By replacing calculated ΔG^{w}_{solv} with experimentally determined water solvation free energies, the correlation between computed and measured *n*-octanol/water partition coefficients was much stronger ($r = 0.95$ and MAE = 0.37 log unit). In a later paper,[45] this GB/SA model was compared with a computationally intensive approach involving free energy perturbation (FEP) and MD in explicit solvents within the AMBER environment.[46] The predictive power of the continuum solvation models was better than that of the all-atom simulations for a set of 12 small and rigid molecules (MAE being 0.50 and 0.74 log unit for GB/SA and FEP, respectively) at a tiny fraction of computer time, demonstrating the potential of the GB/SA model.

A direct calculation of log $P_{o/w}$ making use of an original way to estimate GB was proposed.[47] The study argues for the use of an improved correction term to the Coulomb law to solve an intrinsic issue of continuum solvent models that hardly account for solute−solvent hydrogen bonding. Born radii were calculated from the solvent exposure. In practice, two spheres are used in a five-parameter equation. The nonpolar contribution to the solvation free energy was computed by the SASA modulated by eight corrective parameters, which were fitted on a training set of 81 simple organic molecules by multiple linear regression. The computation of log $P_{o/w}$ on this small training set was as accurate as 0.23 log unit RMSE, while the external prediction on a test set of 19 drugs was 0.96 log unit RMSE. This hybrid physics-based/empirical method was called GBLOGP in benchmark studies.[16,17]

These extremely valuable methods linking partition coefficients to calculated solvation free energies all share the limitation of being generated on a few experimental data points only, including a majority of chemical structures with properties differing greatly from druglike molecules. As a response and taking advantage of the recent impressive improvements in solvation models, we reinvestigated the relationship between GB/SA solvation free energies and log $P_{o/w}$. The use of the analytical generalized Born molecular volume 2 model (GBMV2)[48,49] for electrostatic contributions together with a simple description of the SASA for the nonpolar term within the CHARMM environment[50] enabled to compute the Gibbs solvation free energies in *n*-octanol and water on an extensive cleansed data set of more than 17,500 molecules with trustworthy experimental log $P_{o/w}$ values. Different levels of methodological approximation were applied and are discussed to find the optimal balance between statistical relevance, robustness, predictive power, physical meaning, chemical interpretation, and computational speed in order to make this approach an efficient tool for drug discovery and development purposes.

METHODS

Building of Molecular Data Sets. *Training Sets and Subsets.* A cleansed data set consisting of more than 17,500 organic molecules with experimentally determined *n*-octanol/water partition coefficients was built. Four freely accessible sources of log $P_{o/w}$ values were considered: (i) LOGKOW database (Sangster Research Laboratories, http://logkow.cisti.nrc.ca/logkow/intro.html), (ii) data set of KOWWIN[51] provided through the EPI Suite (version 4.11, 2012, http://www.epa.gov), (iii) DrugBank database (version 3.0, 2011, http://www.drugbank.ca), and (iv) Enhanced NCI Database Browser (version 2.2, 2013, http://cactus.nci.nih.gov).[53] For the sake of molecular description homogenization, all chemical structures were converted to canonical SMILES using OpenBabel (version 2.3.0, 2012, http://openbabel.org).[54] In case of duplicated molecules identified by the Obgrep program (OpenBabel, version 2.3.0) or the JChem Search utility (version 6.1.0, 2013, http://www.chemaxon.com), the log $P_{o/w}$ value was attributed according to the following priority order: LOGKOW "recommended" values, then EPI, then DrugBank, and finally NCI. The highest priority was given to the LOGKOW database because it contains "recommended" values selected by experts after meticulous curation and then to the data from EPI that were purposely compiled to train the KOWWIN log *P* predictor. Lowest priority was attributed to DrugBank and NCI because these databases are not specifically dedicated to chemistry or physicochemistry purposes.

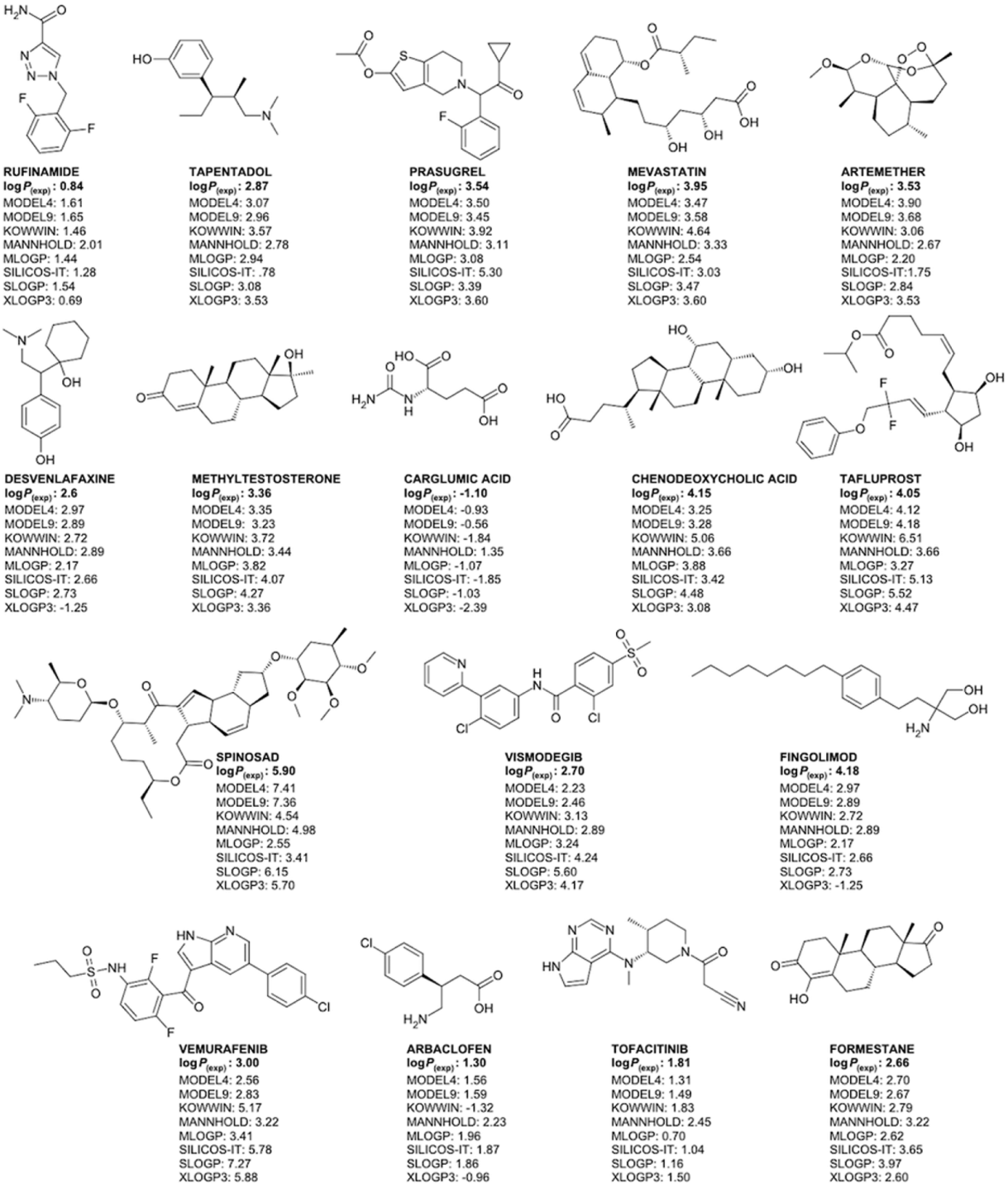
Some information was missing or wrong in the data sources. In particular, this prevented us to define the stereochemistry for numerous entries of the data set. ChemAxon's Standardizer (version 6.1.0, 2013, http://www.chemaxon.com) was employed to delete all stereochemical information, to remove counterions, salts and solvent, and finally to neutralize the molecule. A subsequent substantial data cleansing effort was necessary. Problematic entries were removed either because of a structural mistake/ambiguity or unrealistic experimental log $P_{o/w}$ values.

When a structural mistake or ambiguity was detected because of the impossibility to process further the SMILES, or of the generation of charged or very uncommon chemical functions, the CAS number (or a chemical name) was retrieved. Additional resources, i.e., Pubchem (U.S. National Institute of Health, https://pubchem.ncbi.nlm.nih.gov),[55] Chemspider (Royal Society of Chemistry, http://www.chemspider.com), or chemIDplus (U.S. National Library of Medicine, http://chem.sis.nlm.nih.gov/chemidplus/), were then queried with the CAS number (or chemical name). In case output of such searches converged to a common structural description, the corresponding SMILES (canonicalized by OpenBabel) replaced the erroneous one. We removed all entries for which the searches within external resources did not lead to a unique compound or when the CAS number is not available or not corresponding to the original SMILES.

Then, experimental partition values that were suspected wrong (extreme or not matching the structure *a priori*) were checked directly from the original literature, if possible. Otherwise the log $P_{o/w}$ was estimated by six typical empirical predictive methods (described in Use of Empirical Log P Predictors in the Methods section). As it is unlikely that all the predictive methods share the

dx.doi.org/10.1021/ci500467k | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

Chart 1. Chemical Structures of the *17-Drug* External Test Set[a]



[a]Log $P_{o/w}$ predictions by GB/SA models (Model 4 and Model 5) and six empirical methods (KOWWIN, MANNHOLD; KOWWIN; SILICOS-IT, SLOGP, and XLOGP3) are compared to the experimental *n*-octanol/water partition coefficients (log $P_{(exp)}$).

same bias, the experimental value was kept only if close to a consensus prediction of the six computed values. If not, then the entry was withdrawn from the data set.

Data cleansing led to a reliable training set of 17,511 unique SMILES linked with log $P_{o/w}$ values to assess the relationship between computed solvation free energies and experimentally determined *n*-octanol/water partition.

For thorough analyses of the link between implicit solvent free energy and partition, several subsets were created. Two subsets relate to the flexibility of the molecules; the "RB10" and "RB5" sets contain 16,572 and 12,647 molecules, respectively, bearing not more than 10 and 5 rotatable bonds, respectively. Two other subsets were built by making use of OpenEye's FILTER program (version 2.4.6, 2013, http://www.eyesopen.com). The "BLOCKBUSTER" set involves 11,993 molecules passing the permissive druglike criteria as defined in the "BlockBuster" settings of FILTER (for filtering parameters, please refer to the online documentation (http://www.eyesopen.com/docs/filter/

dx.doi.org/10.1021/ci500467k | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

current/html/filter_files.html?highlight=blockbuster). The "DRUG" set consists of 3754 molecules passing the strict seminal druglike criteria as established by Oprea.[56]

Test Sets. Finding a significant number of compounds associated with reliable experimental partition values and external to our large training set was made possible thanks to the recent publication of Martel et al., which presents log $P_{o/w}$ measurements for 707 molecules performed in an unique laboratory following a liquid chromatography (LC) standardized procedure.[57] The compounds were chosen to extensively cover the chemical space starting from a vendor-driven selection of 4.5 million structures from the ZINC database[58] mined by global and tridimensional descriptors. The strong advantages of this data set are the uniform experimental procedure, which assures homogeneous measurements, and its broad chemical diversity resulting from the vast variety of the structural sources. A drawback stems from the use of a unique LC column with C-18 reversed phase properties for all samples, which implies that only relatively hydrophobic molecules could be measured. As a consequence, the lipophilicity range of this molecular set is rather narrow and hydrophobic. According to the Obgrep program (OpenBabel, version 2.3.0) and the JChem Search utility (version 6.1.0, 2013, http://www.chemaxon.com), only one structure was part of our 17,511-molecule training set. The 706 remaining molecules with a log $P_{o/w}$ between 0.30 and 6.96 represent the *Martel* external test set. The probability to find any of these molecules in the training set of any broadly employed predictor is very low (though not null). Overall, we consider the *Martel* test set as a fair evaluation instrument to compare the predictive power of our GB/SA-based approach with that of typical fragmental and topological log *P* methods.

In order to create an additional test set with an even lower probability of overlap with any training set and hence to ascertain real unbiased comparison of predictive capacities, we retrieved the experimental lipophilicity of 17 recent entries in the DrugBank 4.1 database[59] (termed *17-drug* test set, Chart 1). The selection was made on entries added after June 2014, which is posterior to the setup of our 17,511-molecule training set and to those of the empirical methods compared (refer to the Use of Empirical Log P Predictors section). The externality of this test set with our 17,511-molecule training set was confirmed by queries with the Obgrep program (OpenBabel, version 2.3.0) and the JChem Search utility (version 6.1.0, 2013, http://www.chemaxon.com). This verification could not be achieved for the six empirical methods because their training sets are not disclosed or not freely available. However, the recent addition of *17-drug* entries in the DrugBank database establishes the externality of this test set to any of the training sets involved in this comparative study. Moreover, the selection focused on molecules with clearly referenced log $P_{o/w}$, and macromolecules were discarded. Fifteen compounds are FDA-approved drugs, whereas the two others, mevastatin and arbaclofen, are a natural precursor of hypolipidemic statins and the *R*-enantiomer of the muscle relaxant baclofen, respectively. All *n*-octanol/water partition measurements were taken from the DrugBank database, except for vemurafenib, whose experimental log $P_{o/w}$ was found in the new drug application review of the FDA (http://www.accessdata.fda.gov/drugsatfda_docs/nda/2011/202429Orig1s000ClinPharmR.pdf, accessed June 2014). The experimental lipophilicity values of desvenlafaxine and spinosad are referring to distribution coefficient at physiological pH (log $D_{7.4}$). The correpsonding partition coefficient was obtained by using the ionization constants given by Marvin's Calculator plugin (version 6.3.1, 2014, http://www.chemaxon.com).

Both the *Martel* and *17-drug* external test sets were submitted to log $P_{o/w}$ predictions using the six widely employed empirical predictors described below as well as two of our GB/SA models most appropriate for drug design purpose (because statistically relevant and robust although one single conformer per molecules is generated). Similarly to the training procedure, the computational workflow, described in the next paragraph, started from SMILES that were retrieved from the respective database Web sites (http://zinc.docking.org, accessed on March 2014, and http://www.drugbank.ca, accessed June 2014).

Computation of GB/SA Parameters and Link with Log $P_{o/w}$. The solvation free energies in water and *n*-octanol for the molecules under study were obtained through the following computational workflow.

Conformers, Topology, and Molecular Mechanics Parameters Generation. Tridimensional geometries were generated from SMILES[60] using OpenEye's OMEGA program[61,62] (version 2.4.6, 2013, http://www.eyesopen.com). In order to analyze in detail the impact of the conformation on the calculated solvation free energy, 1, 20, or 200 conformers containing all hydrogen atoms were generated in MOL2 format. CHARMM27 all-atom force field compatible topologies and parameters for these tridimensional molecular structures were generated with the SwissParam tool (SIB Swiss Institute of Bioinformatics, 2014, http://www.swissparam.ch). This technique calculates topologies and parameters relying on the Merck molecular force field,[40−44] yet in a functional form that can directly be input in the CHARMM force field computation.[63] For our needs, charges and force constants were taken from MMFF and van der Waals radii from CHARMM27 all-atom force field.

Geometry Optimizations. A short optimization was performed on every conformer of every molecule within the CHARMM environment (version c36b1)[50] using the parameters obtained from SwissParam and CHARMM27 all-atom formalism. Each starting conformation was submitted to two different 250-step steepest descent minimizations in parallel. One optimization was done in a continuous medium mimicking the aqueous phase (dielectric constant, ε_{wat} set to 80), while the other minimization took place in an octanol-like phase with a lower dielectric constant ε_{oct} = 10.3.[39] The internal dielectric constant of the solute was ε_{solut} = 1.

Solvent Accessible Surface Areas (SASA). The solvent accessible surface areas of every solvent-dependent optimized conformer ($SASA^w$ and $SASA^o$, respectively) were computed to evaluate the nonpolar contributions to the free energies of solvation ($\Delta G^o_{solv,np}$ and $\Delta G^w_{solv,np}$, respectively). The calculations were performed analytically within CHARMM version c36b1. The radii of the solvents (RPROBE) were set to 1.4 and 2.0 Å for water and *n*-octanol, respectively.[39]

Generalized Born (GB) Parameters. The electrostatic contributions to the solvation free energies ($\Delta G^o_{elec,solv}$ and $\Delta G^w_{elec,solv}$, for water and *n*-octanol phases, respectively) were evaluated by the GB approach through the analytical model of GBMV2,[48,49] which was shown to greatly reduce computational time without compromising the accuracy.[64] Single-point energy calculations with infinite cutoffs were performed with the analytical method II of the GBMV module as implemented in CHARMM (version c36b1) to obtain $\Delta G^o_{elec,solv}$ and $\Delta G^w_{elec,solv}$ values for every conformer of all molecules. For conformations generated in water, the dielectric constant (ε_{wat}) was set to 80 (EPSILON default value), whereas for conformations generated

dx.doi.org/10.1021/ci500467k | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

in *n*-octanol, the dielectric constant (ε_{oct}) was set to 10.3. All other parameters were kept as default. Centering the molecule over the grid was required to obtain reproducible values from GB computation.

Multiple Linear Regression (MLR) Analysis. Solvation (i.e., GB/SA) values generated on the whole data set were used to investigate the correlation with the *n*-octanol/water partition coefficient. The electrostatic solvation energy difference $\Delta\Delta G_{solv,elec} = \Delta G^{w}_{solv,elec} - \Delta G^{o}_{solv,elec}$ together with the nonpolar solvation energy difference $\Delta SASA = SASA^{w} - SASA^{o}$, or $SASA^{w}$ alone, which are meant to model the total solvation free energy difference between both solvents, were submitted to multiple linear regression (MLR) analyses against experimental log $P_{o/w}$ values for the 17,511 compounds of the entire data set or for the four subsets described above (RB10, RB5, BLOCKBUSTER, and DRUG). MLR was preferred to a strict application of eq 2 in order to account for the unavoidable approximations regarding the description of the system. When multiple conformations were generated from a single molecular structure (20 or 200 conformers in the present study), the Boltzmann average of every GB/SA value was considered for correlation. Hence, two types of multilinear relationship were built and evaluated: the models taking the difference of SASA in both solvents (ΔSASA, eq 4) as nonpolar parameter or the ones accounting only for SASA from geometries obtained in water ($SASA^{w}$, eq 5):

$$\log P_{o/w} = \alpha_1 \times \Delta\Delta G_{solv,elec} + \beta_1 \times \Delta SASA + C_1 \quad (4)$$

$$\log P_{o/w} = \alpha_2 \times \Delta\Delta G_{solv,elec} + \beta_2 \times SASA^{w} + C_2 \quad (5)$$

where α_1, α_2, β_1, and β_2 are the regression coefficients, and C_1 and C_2 are the regression constants. These models have in common to involve only three parameters, which were trained on multiple thousands of data points, thus avoiding overfitting biases. As described below, the coefficients and constants as well as the regression statistics vary according to the number of geometries generated (1, 20, or 200) and the data set on which the MLR model is trained (entire, RB10, RB5, BLOCKBUSTER, or DRUG). The various models obtained by training eqs 4 or 5 on the different training sets were submitted to internal validation by performing cross-validation and data randomization. Subsequently, the most suitable models for drug design purposes, validated this way, were used to predict log $P_{o/w}$ on two external test sets. The external predictive capacity of our GB/SA-based models was compared to six empirical log *P* predictors widely employed by the medicinal chemists and the cheminformatics community, relying on either fragmental or topological methodologies and described below.

Molecular Dynamics. Two hundred molecules from the entire data set, regularly distributed over the whole range of experimental log $P_{o/w}$ values, were submitted to molecular dynamics (MD) simulations in order to further scrutinize the relationship between geometry, solvation, and partition.

For each molecule, a MD simulation was started from the single conformation generated by OMEGA and minimized in CHARMM using the SwissParam parameters. The *leapfrog* dynamic integrator of the CHARMM package (version c36b1) and the GBMV2 implicit aqueous and octanol-like solvation parameters were used. First, the system was heated to 300 K over 6 ps. The SHAKE algorithm[65] was applied to restrain the length of the bonds involving a hydrogen atom. Trajectories of 7.5 ns were produced with a 1.5 fs integration step in both solvents in parallel. The temperature was maintained at 300 K by coupling heavy atoms to a Langevin thermostat with a 10 ps^{-1} frictional coefficient. GB and SASA values were computed on conformations from both solvents trajectories at snapshots taken every 1000 steps during the production phase. MLR models based on our GB/SA approach using either eqs 4 or 5 were employed to calculate log $P_{o/w}$ along the simulations.

Internal Validation and Data Randomization. In order to evaluate internal quality and robustness, we submitted the most interesting GB/SA models to five-fold cross-validation and *Y*-scrambling procedures.

Cross-validation consists in selecting some entries of the entire data set in order to gather a validation set. The model is retrained on the remaining entries (considered as training set) and the log $P_{o/w}$ calculated for the molecules belonging to the validation set. In the present case, a five-fold cross-validation was performed. We attributed every fifth entry of the list of molecules ranked by experimental log $P_{o/w}$ values to the validation set. This was repeated five times starting from entries 1, 2, 3, 4, or 5 to obtain five different pairs of training/validation sets covering the broadest possible range of experimental log *P*. The MLR models were trained on each training set and log *P* computed for the molecules belonging to the corresponding validation set. The cross-validated correlation coefficient, q^2_{CV}, was calculated, and error metrics were averaged over the five validation sets (refer to the Statistical Evaluation of Computed Values section).

Y-scramblings were performed to check that multilinear correlations are not mere coincidence. This test consisted in randomizing the observations between all entries, while keeping their $\Delta\Delta G_{solv,elec}$ and SASA values. For this study, we tested the statistical relevance of 25,000 "fake" models built by training the MLR on scrambled experimental log $P_{o/w}$ values and compared them to the statistical parameters of the genuine models.

Use of Empirical Log *P* Predictors. The external predictive power of GB/SA models most suitable for drug design purposes was compared to that of typical empirical log *P* prediction methods on two external test sets (*Martel* and *17-drug* sets as described above). We have limited our work to free-for-academics programs allowing batch computation.

KOWWIN. KOWWIN is a reductionist fragmental log *P* predictor trained on 2473 simple molecules and tested on 10,589 molecules of diverse complexity. It involves 150 fragments and 250 correction factors to take into account steric forces, hydrogen-bonding, and polar effects of specific moieties.[19] We have used the SMILES batch mode calculation of KOWWIN (version 1.68, 2000, http://www.epa.gov/oppt/exposure/pubs/episuite.htm) as embedded in the EPI Suite.

MANNHOLD. In an extensive benchmark study,[16] Mannhold proposed an very simple topological approach based solely on three parameters: number of carbon atoms, number of heteroatoms in the molecule under consideration, and a regression constant. A MLR model was trained on an extensive but undisclosed industrial proprietary collection of 95,809 molecules with experimental log *P* values.

MLOGP. MLOGP is a pioneer topological method developed by Moriguchi[25,26] relying on a linear relationship involving 13 simple parameters describing the topology of the molecular structure and a regression constant. The model was trained on 1230 molecules with measured *n*-octanol/water partition coefficients. MLOGP was implemented in various commercial packages, e.g., ADMET Predictor (Simulations Plus, Inc., http://www.simulations-plus.com) and Dragon[66] with striking discrepancies regarding performance in benchmark studies.[16,17] The reason seems to be the problematic interpretation and coding of

dx.doi.org/10.1021/ci500467k | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

some ambiguous topological descriptors. In the present study, an in-house implementation was employed based on the description given by Lipinski in his effort to refactorize the rule-of-five using MLOGP instead of CLOGP.[10] The descriptors were translated to SMARTS and the recognition engine used OpenBabel together with homemade Python scripts.

SILICOS-IT. The log *P* calculator proposed by Silicos-it is a hybrid method relying on 27 fragments and 7 topological descriptors as described in the online documentation (http://silicos-it.be.s3-website-eu-west-1.amazonaws.com/software/filter-it/1.0.2/filter-it.html, accessed October 2014). It was trained on 23,455 molecules with experimental *n*-octanol/water partition values from the Syracuse PHYSPROP database (http://www.srcinc.com/what-we-do/environmental/scientific-databases.html). The training set covers molecular weights between 200 and 600 and has a log $P_{o/w}$ range between −2.6 and 8.1. For our study, the predictions were computed using the FILTER-IT program (version 1.0.2, 2013, www.silicos-it.be).

SLOGP. SLOGP is an atomic method based on the fragmental system by Wildman and Crippen including 142 fragments and trained on 9920 molecular structures.[24] The MOE commercial modeling package (Chemical Computing Group, http://www.chemcomp.com) has an implementation of SLOGP. For the present study, we employed an in-house implementation based on SMARTS recognition pattern using OpenBabel and tailor-made Perl scripts.

XLOGP3. The version 3 of the XLOGP atomic model is based on a system of 87 fragments and two corrective factors. The main methodological novelty is the use of a library of reference compounds with known experimental partition coefficient values.[23] If the input structure is similar to a reference compound, only the fragments differentiating them are treated and the corresponding log *P* contributions added to the reference structure log *P* value. Otherwise, a purely atomic prediction is performed. This library of reference compounds is also the training set and involves 8199 structures with measured partition coefficients. For our study, the computation was performed through the command-line Linux executable (version 3.2.2). Inputs were in SDF format. The conversion from SMILES was achieved by Openbabel,[54] all hydrogens were added and neutral forms preferred. For the sake of avoiding biases, both the default value (XLOGP3) and the value obtained by the purely atomic methodology without knowledge-based reference starting point were computed (XLOGP3-aa, for all-atom).

Statistical Evaluation of Computed Values. The accuracy of the log $P_{o/w}$ values computed by the different models and methods was assessed using two standard error metrics. The root mean squared error (RMSE) was calculated to quantify the average error on all computed values compared to the experimental ones. Because the individual errors are squared before averaged, the RMSE is more impacted by large errors than by small errors. The mean absolute error (MAE) was calculated for every prediction, too. It is the average over the data set of the differences between individual computed value and the corresponding experimental measure. In contrast with RMSE, MAE is a linear function, which means that all individual differences are weighted equally in the average. Both RMSE and MAE range from zero to infinity, with zero being a perfect accuracy for all calculations.

These error metrics as well as the correlation coefficient (r) and the variation of regression coefficients (α_1, β_1, C_1, or α_2, β_2, C_2 for eqs 4 and 5, respectively) were used to evaluate the intrinsic statistical significance of the GB/SA models. Their robustness tested by *Y*-scrambling was also evaluated this way. The parameter k is added to assess the predictive power of various models and methods on external test sets. The parameter k is the slope of the linear regression when the line is forced going through the origin. Finally, the cross-validated correlation coefficient q^2_{CV} together with $RMSE_{CV}$ and MAE_{CV} are measures for internal validation (see the Internal Validation and Data Randomization section).

RESULTS AND DISCUSSION

Great care was taken for building a relevant large training set of molecular structures linked with trustable measurements of *n*-octanol/water partition. The merger of four public sources followed by data cleansing led to the collection of 17,511 data points. The range of experimental log $P_{o/w}$ values of the entire training set is broad, from −8.00 to 11.29. As depicted in Figure 1

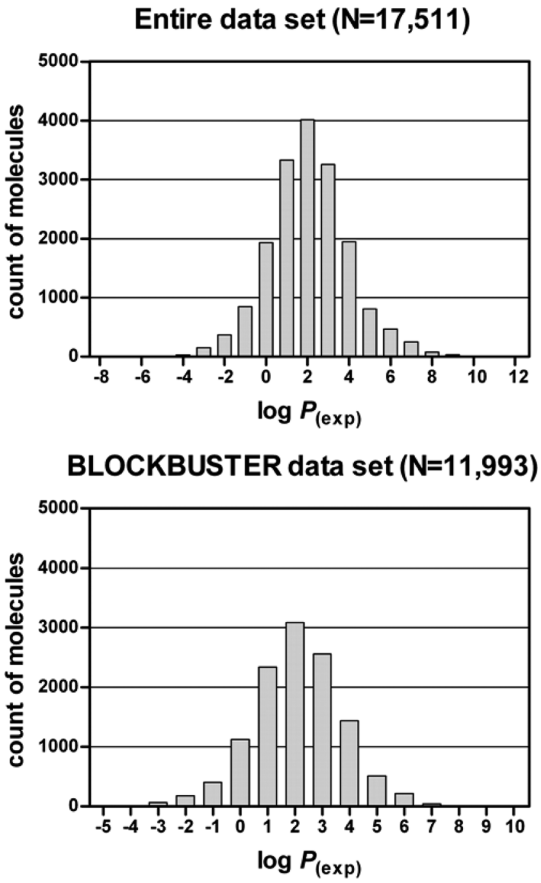


Figure 1. Distribution of experimental *n*-octanol/water partition coefficient values (log $P_{(exp)}$) for the entire 17,511-molecule data set (upper panel, mean 2.06, standard deviation 1.92) and for the 11,993-molecule BLOCKBUSTER set (lower panel, mean 2.09, standard deviation 1.64).

(upper panel), the distribution appears nearly normal with a mean at 2.06 log *P* unit and a standard deviation of 1.92. Besides, the molecular weight ranges between 18.0 and 1550.2 with a mean at 262.4 and a standard deviation of 116.4.

The computation of GB and SASA values for this large number of chemical structures was affordable thanks to the automatic calculation of CHARMM topologies and parameters with SwissParam,[63] efficient GBMV2 implementation,[48,49] and

dx.doi.org/10.1021/ci500467k | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

Table 1. Log $P_{o/w}$ Multilinear Regression Models Built on the Entire Training Set (17,511 molecules) Considering $\Delta\Delta G_{solv,elec}$ and ΔSASA as Electrostatic and Nonpolar Parameters of the Free Energy of Solvation, Respectively[a]

		regression coefficients[b]					
model #	number of conformers	α_1 (mol/kcal)	β_1 (Å^{-2})	C_1	r[c]	RMSE[d]	MAE[e]
1	1	1.311	−0.0779	−3.671	0.72	1.33	1.03
2	20	1.290	−0.0778	−3.594	0.74	1.30	1.00
3	200	1.272	−0.0792	−3.720	0.74	1.29	0.99

[a]1, 20, or 200 conformations per molecule were generated. [b]Multilinear regression coefficients of eq 4. [c]Correlation coefficient. [d]Root mean square error. [e]Mean average error.

Table 2. Log $P_{o/w}$ Multilinear Regression Models Built on the Entire Training Set (17,511 molecules) Considering $\Delta\Delta G_{solv,elec}$ and SASA^w as Electrostatic and Nonpolar Parameters of the Free Energy of Solvation, Respectively[a]

		regression coefficients[b]					
model #	number of conformers	α_2 (mol/kcal)	β_2 (Å^{-2})	C_2	r[c]	RMSE[d]	MAE[e]
4	1	1.367	0.0085	0.187	0.72	1.33	1.03
5	20	1.336	0.0086	0.221	0.74	1.30	1.00
6	200	1.312	0.0088	0.140	0.74	1.29	0.99

[a]1, 20, or 200 conformations per molecule were generated. [b]Multilinear regression coefficients of eq 5. [c]Correlation coefficient. [d]Root mean square error. [e]Mean average error.

analytic determination of SASA in CHARMM.[50] At the time of writing this manuscript, the entire computational procedure described in the Methods section, from SMILES input to log $P_{o/w}$ output, took 1 to 2 s real-time on a single core of Intel Core2 Q6600 (2.40Ghz) for a single conformation, depending on the size of the molecule. The computation is not optimized yet, and a future production version is expected to significantly reduce the computational time in order to meet large-scale medicinal chemistry project requirements.

Relationship between Computed Implicit Solvation and *n*-Octanol/Water Partition. Variants in the methodology at different levels of descriptive approximations were applied to evaluate the statistical and physical relevance of our GB/SA-based MLR models that link solvation free energy and experimental *n*-octanol/water partition. Table 1 shows the regression coefficients, correlation coefficient, and both error metrics (RMSE and MAE) for the models built on the entire training set (17,511 molecules), generating 1, 20, 200 conformations per molecule and accounting for the difference of solvent accessible surface areas in both solvents (ΔSASA, eq 4).

Using a single conformation to determine geometry-dependent parameters may be seen as a too crude approximation *a priori*. However, model 1, as elementary as it can be, validates the relationship between experimental log $P_{o/w}$ and GB/SA parameters computed in octanolic and aqueous implicit media, with a linear correlation coefficient of 0.72. The MLR coefficients illustrate the physical significance of the proposed approach. Indeed the $\Delta\Delta G_{solv,elec}$ parameter is weighted by a coefficient close to the theoretical value of $1/RT$ = 1.66 mol/kcal (α_1 = 1.311; refer to eq 4), which supports the relevance of the calculated differential solvation energy. The ΔSASA parameter coefficient is $\beta_1 = -0.0779$ Å^{-2} (refer to eq 4). According to the GB/SA theory of solvation energy, $\beta 1$ should be proportional to the difference of surface tensions in water and *n*-octanol. However, the coefficients of the GB/SA parameters obtained by linear regression on a set of experimental observations might not be strictly equal to the theoretical values. Clearly, the GB/SA approach implies an approximate description of the solute/bisolvent system. First, the *n*-octanol molecule is considered as a sphere, which is far from realistic. This impacts both the electrostatic and the nonpolar parts of the equation. Furthermore, the genuine octanolic phase is wetted by 4.13% of water,[67,68] which suggests an even more complex micellar-like reticulated nature of the water-saturated octanolic medium.[69] Second, the interface between both solvents is ignored by the model, which considers differential solubility instead of transfer between both phases. We believe that these approximations influence both the α_1 and β_1 coefficients of $\Delta\Delta G_{solv,elec}$ and ΔSASA, respectively. The approximations that are independent from the GB/SA parameters are accounted for by the regression constant ($C_1 = -3.671$; refer to eq 4).

The difficulty to model the octanolic phase as an implicit solvent led us to build other MLR models on the entire training set (17,511 molecules) by approaching the nonpolar contribution to the free energy with the water accessible surface area solely (SASA^w, eq 5). Regression coefficients, as well as the correlation coefficient and both error metrics (RMSE and MAE) for those models are presented in Table 2. For the single conformation model using SASAw (model 4), the correlation and the error between calculated and experimental values are strictly equal to the corresponding model using ΔSASA (model 1). Therefore, the use of SASA^w instead of ΔSASA has no impact the statistical quality of the model, demonstrating that SASA^o does not contribute significantly to model 1, model 2, and model 3. Again for model 4, the regression coefficients are arguing for a physically meaningful linear relationship between log $P_{o/w}$ and solvation free energies computed in both implicit solvent phases. Indeed, the weighting of $\Delta\Delta G_{solv,elec}$ is close to $1/RT$ (α_2 = 1.3669 mol/kcal; refer to eq 5). The SASA^w weighting is β_2 = 0.0085 Å^{-2} (refer to eq 5) and in the same order of magnitude than σ/RT, where σ is the nonpolar surface tension for water, classically defined[35,70,71] as 0.0072 kcal/mol Å^2. Moreover, the regression constant is further supporting our working hypothesis relying on physics theory in depicting a *Y*-intercept close to origin (C_2 = 0.1867 ; refer to eq 5). Overall, this indicates that considering SASA^w for the nonpolar contributions of energies in our GB/SA models is a valid approximation. In fact, SASA^o and SASA^w are strongly intercorrelated with a linear coefficient r = 0.99. Thus, SASA^o does not bring relevant additional information to the description of nonpolar effects and can be withdrawn from the linear model to calculate log $P_{o/w}$. As a result in model 4, the

dx.doi.org/10.1021/ci500467k | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

Table 3. Log $P_{o/w}$ Multilinear Regression Models Built on Reduced Training Subsets Using $\Delta\Delta G_{solv,elec}$ and SASA^w as Electrostatic and Nonpolar Parameters of the Free Energy of Solvation, Respectively[a]

		regression coefficients[b]					
model #	subset (number of molecules)	α_2 (mol/kcal)	β_2 (Å^{-2})	C_2	r[c]	RMSE[d]	MAE[e]
7	RB10 (N = 16,572)	1.324	0.0090	−0.099	0.72	1.29	1.00
8	RB5 (N = 12,647)	1.387	0.0109	−0.748	0.73	1.25	0.97
9	BLOCKBUSTER (N = 11,993)	1.161	0.0084	−0.068	0.72	1.14	0.89
10	DRUG (N = 3,754)	0.915	0.0065	0.261	0.60	1.00	0.79

[a]One conformation per molecule was generated. [b]Multilinear regression coefficients of eq 5. [c]Correlation coefficient. [d]Root mean square error. [e]Mean average error.

complex nature of the water-saturated octanolic phase mentioned earlier and its impact on partition are accounted for by the coefficients of the regression, which make physical sense.

Accounting for Molecular Flexibility. In order to investigate molecular flexibility, we generated 20 or 200 conformations of each molecule before computing the GB and SA values for each conformer in both aqueous and octanol-like media. MLR models were built using Boltzmann-averaged values. The impact on the regressions is shown in Tables 1 and 2 for models either using ΔSASA (model 2 and model 3) or SASA^w (model 5 and model 6) as the nonpolar term. The statistical quality of the models improved with the number of conformations involved in the computation of the solvation free energies, as indicated by slightly stronger linear correlations (r) and smaller errors. Moreover, both RMSE (individual errors squared) and MAE (individual errors weighted equally) are reduced by the same order of magnitude. This indicates that generating multiple conformers did not fix important deviations in the geometries leading to a few large errors, but rather polished some subtle tridimensional effects influencing the GB and/or SASA parameters by taking an average. Moreover, generating multiple conformers did not significantly affect the regressions coefficients and constants of eq 4 (α_1, β_1, and C_1 in Table 1) or eq 5 (α_2, β_2, and C_2 in Table 2). Averaging out GB/SA parameters from 200 conformations generated with OMEGA led to the best models with an average error of 0.99 log unit and a correlation coefficient of 0.74 (either accounting for ΔSASA in model 3 or only for SASA^w in model 6). To gauge the statistical relevance of model 3 and model 6, it is worthy to note that the GB/SA parameter coefficients were generated from about seven million tridimensional molecular geometries.

Another way to consider the impact of molecular flexibility was to train GB/SA MLR models on two reduced subsets introducing thresholds of maximum rotatable bounds (10 or 5 for subset RB10 and RB5, respectively). Table 3 shows the regression data and error metrics for model 7 (RB10) and model 8 (RB5) generated on one single conformation per molecule and using SASA^w as the nonpolar term (refer to eq 5). Comparing model 7 (trained on 16,572 molecules) with model 4 (trained on 17,511 molecules) shows that filtering out flexible molecules reduced the error on the calculated log $P_{o/w}$ by 0.04 and 0.03 log unit for RMSE and MAE, respectively. In contrast, the correlation coefficient did not increase. The trend of smaller errors for more rigid molecules is confirmed by lowering the maximum number of rotatable bonds to five which filtered out 4864 molecules. As a consequence, model 8 (RB5, trained on 12,647 molecules) demonstrated smaller RMSE (by 0.04) and MAE (by 0.03) compared to model 7 (RB10, trained on 16,572 molecules). Besides, the correlation did not notably improve (r = 0.72 for model 7 and r = 0.73 for model 8). Similarly, the MLR coefficients and constants were more impacted in model 8, trained on the more rigid RB5 data set, compared to model 7, built on RB10 (α_2, β_2, and C_2 in Table 3). This can be attributed to the number of data points withdrawn from the data set; less than 5.5% of observations were removed for RB10 models compared to more than 28% for RB5 models. Remarkably, the physical description of n-octanol/water partition given by eq 5 remained identical in the case of reduced data sets of more rigid molecules. The exact same conclusion can be drawn for eq 4 from model S1 and model S2 trained on the RB10 and RB5 subsets, respectively, but considering ΔSASA (Table S1, Supporting Information). Altogether, these results confirm that generating the GB/SA values on a single conformation cannot account for all subtle effects impacting solubility and partition. The influence of molecular flexibility is bigger for molecules with many rotatable bounds. The more flexible the molecule is, the stronger the potential bias is. Nonetheless, given the statistical quality of the MLR models, the physical relevance of our GB/SA approach for log $P_{o/w}$ is undoubtedly preserved by the single conformation approximation.

MD simulations were carried out in an attempt to further study the link between molecular flexibility and GB/SA regression models for the n-octanol/water partition. Two hundred molecules covering the entire range of log $P_{o/w}$ of the training set were submitted to MD simulations at ambient temperature. GB/SA values were generated for conformations along both implicit n-octanol and water trajectories. MLR models were built on the fly, training either eq 4 or eq 5 on the 200-molecule set. After 7.5 ns, RMSE was equal to 1.45 and 1.39 log P unit for eqs 4 and eq 5, respectively. For the sake of comparison, the MLR models were trained for 1, 20, and 200 conformations using OMEGA as geometry generator for the same 200 molecules. The RMSE of the single conformation models were significantly lower (1.36 and 1.34 for eq 4 and eq 5, respectively) than those of the 7.5 ns MD. It follows that time-consuming MD simulations at room temperature cannot straightforwardly give a more relevant physical description of the dual implicit solvent system. The generation of multiple conformers in vacuum then optimized in the implicit solvent media to calculate the Boltzmann-averaged GB/SA values appears more suited to improve model relevance by accounting for molecular flexibility. Indeed, for 20 conformations generated with OMEGA on the 200-molecule set, RMSE were equal to 1.28 and 1.30 when using eq 4 and eq 5, respectively. For 200 conformations, RMSE values were equal to 1.24 and 1.27 when using eq 4 and eq 5, respectively. This increased accuracy of the GB/SA MLR models as a function of the number of geometries optimized in implicit solvents illustrates again the physical relevance of our approach to describe partition between n-octanol and water.

Focus on Druglikeness. With the aim of meeting medicinal chemistry and drug discovery needs, we attempted to test whether the one-conformation approximation can be valid to

dx.doi.org/10.1021/ci500467k | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

estimate log $P_{o/w}$ of druglike molecules with adequate accuracy. Two additional reduced subsets were built using two different druglikeness filters, DRUG as described in ref 56 and BLOCKBUSTER as defined in OpenEye's FILTER online documentation (http://www.eyesopen.com/docs/filter/current/html/filter_files.html?highlight=blockbuster). The GB and SASA^w values were generated for one single conformation per molecule, then eq 5 was trained on the BLOCKBUSTER and DRUG subsets to generate model 9 and model 10, respectively. Equation 4 was also trained on the subsets to generate model S3 and model S4. As it appears in Table 3 and Table S1 of the Supporting Information, similarly to RB10 and RB5 models, BLOCKBUSTER and DRUG models (model 9, model 10, model S3, and model S4) showed better accuracy with smaller error denoted by lower RMSE and MAE. However, comparison of error metrics should be taken with caution because the statistical relevance of such models calls for closer analysis. Considering the DRUG subset, the number of data points was considerably reduced as 78.6% of the molecular structures failed to satisfy the drastic filter yielding a 3754-molecule set. Also important, the range of measured log $P_{o/w}$ values was narrowed by the druglikeness criteria (from −3.93 to 6.46). This limited dispersion of observations had a direct influence on the error metrics defined by averages. Consequently, RMSD and MAE were mechanically smaller, even though the statistical relevance of the MLR model is poorer, as indicated by the significantly lower correlation coefficient of both models trained on the severely less diverse DRUG data set (r = 0.60 for both model 10 built with eq 5 and model S4 built with eq 4). Moreover the physical relevance of these models is further disputed by the linear regression coefficients. Indeed, α_2, β_2, and C_2 for model 10 as well as α_1, β_1, and C_1 for model S4 are notably dissimilar compared to related models built with eq 5 and eq 4, respectively. This highlights the importance of a large and diverse molecular set to generate meaningful and relevant log $P_{o/w}$ linear models strongly correlated with the GB/SA values derived from implicit solvation theory. In contrast to the DRUG models, those trained on the larger and more diverse BLOCKBUSTER subset are of clear interest. By applying a softer filter, 31.5% of the molecular structures were filtered out yielding a large and diverse yet druglike training set of 11,993 molecules with an appreciably broad experimental log $P_{o/w}$ range between −4.12 and 9.96 following an approximately normal distribution with a mean of 2.09 and a standard deviation of 1.64 (refer to Figure 1, lower panel). Comparison of MLR models built with eq 5 indicates an increased accuracy of model 9 trained on the BLOCKBUSTER subset with respect to model 4 trained on the unfiltered entire data set (17,511 molecules). RMSE and MAE for log P computed by model 9 compared to experimental values are as low as 1.14 and 0.89 log unit, respectively. The correlation is not stronger (r = 0.72), while the linear coefficients and constant of eq 5 are notably well compliant with the GB/SA theory. Indeed $\Delta\Delta G_{solv,elec}$ and SASA^w weightings are in the order of $1/RT$ and σ/RT, respectively, as well the Y-intercept is close to origin. All 17,511 data points in Figure 2 represent log $P_{o/w}$ as computed by the GB/SA model generated on the entire training set (model 4) versus experimental log $P_{o/w}$ measurements. Red crosses depict the 5,518 molecular structures that were rejected by the BLOCKBUSTER filter. The RMSE of model 4 is 1.33 log P unit. A total of 73 molecules have computed log P values exceeding 3-fold RMSE error, 34 of which are overestimated (Chart S1 and Table S2, Supporting Information) and 39 are underestimated (Chart S2 and Table S3, Supporting Information).

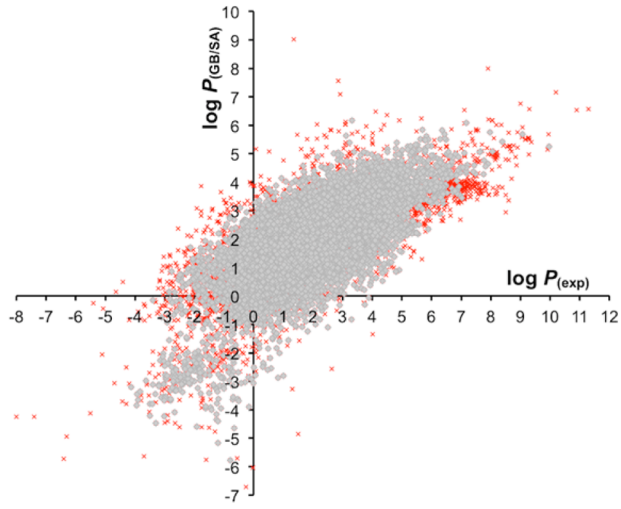


Figure 2. Log $P_{o/w}$ calculated by model 4 (log $P_{(GB/SA)}$) versus experimental values (log $P_{(exp)}$) for the 17,511 molecules of the entire training set. Data for the 11,993 molecules that passed the BLOCKBUSTER druglikeness filter are shown as gray dots, while data for the remaining 5,518 molecules are shown as red crosses.

Most of the outlying computed values are eliminated by the BLOCKBUSTER filter. This should not suggest that our predictive method would be less performing for nondruglike molecules. On the contrary, as shown on the population histogram in Figure 3, numerous molecules failing to satisfy

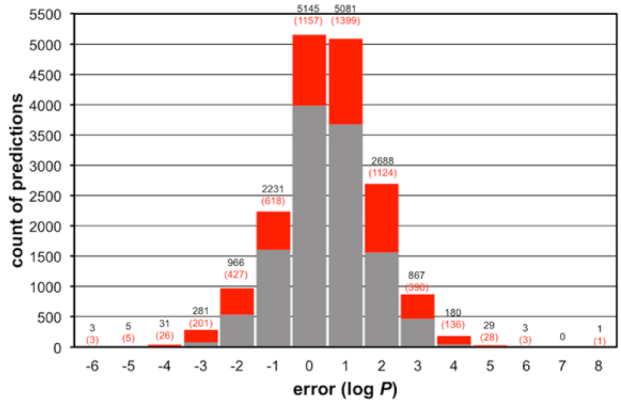


Figure 3. Distribution of errors on log $P_{o/w}$ computed by model 4. The gray bars represent the molecules satisfying the BLOCKBUSTER druglike criteria, and the red bars are for the filtered-out molecules. The numbers over bars are the population count in the 1 − log P unit bin: in black, the total count, and in red in brackets, the filtered-out molecules.

the BLOCKBUSTER criteria were actually accurately calculated. For instance, log P of 2556 filtered-out molecules (so more than 46% of the nondruglike molecules according to the BLOCKBUSTER criteria) was computed by model 4 with an error lower than one log unit. The advantage of using the BLOCKBUSTER filter stems from the cleanup of the training set from structures with properties preeminently dissimilar from druglike molecules. As a result, 11,993 data points remained and are represented by gray dots in Figure 2. As described above, retraining eq 5 on this BLOCKBUSTER subset led to model 9, whose statistical significance is clearly indicated on the calculated-versus-observed graph in Figure 4 showing the vast majority of points close to identity (dashed line). The distribution of errors is nearly normal

dx.doi.org/10.1021/ci500467k | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

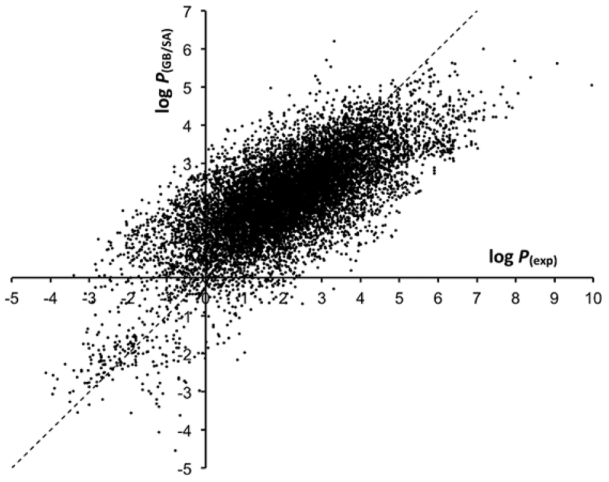


Figure 4. Log $P_{o/w}$ calculated by model 9 (log $P_{(GB/SA)}$) vs experimental log $P_{o/w}$ (log $P_{(exp)}$) for the 11,993 molecules of the BLOCKBUSTER subset. The identity line is dashed.

as depicted by the frequency histogram in Figure 5. The standard deviation (i.e., RMSE) is 1.14 log P unit. Less than 36% of the

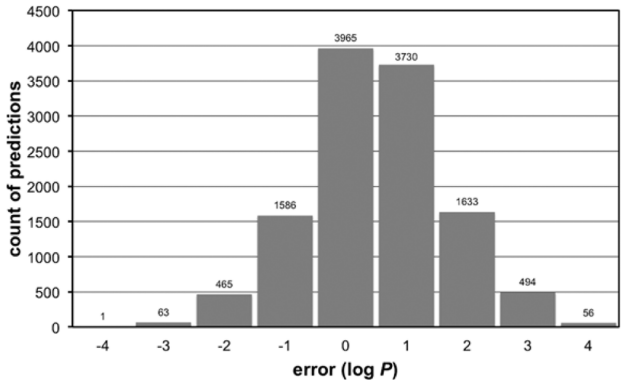


Figure 5. Distribution of errors on log $P_{o/w}$ computed by model 9. The number over bars is the population count in the 1 − log P unit bin.

calculated log P values are above one log unit absolute error. Only 26 molecules have calculated values exceeding 3-fold RMSE error, 13 of which are overestimated (structures and values in Chart S3, Supporting Information). These include mainly short peptide-like structures whose deviations could be related to a inappropriate tridimensional geometry. Three small aromatic sulfonic acids are overestimated as well, which might exemplify how our physics-based methodology can be affected by empirical force field and/or GB incorrect parametrization for a specific moiety. However, taken as a whole, the 13 experimental log $P_{o/w}$ values being all very negative lead us to think retrospectively that measurement error cannot be excluded. The most dramatic example is compound EPI10850 with an apparently overestimated values of Δ(log P) = 3.88 returned by model 9. The computed log P for this molecule is also much higher than the experimental values when submitted to empirical methods (by 3.57, 4.36, 5.01, 4.38, 5.98, and 5.16 for KOWWIN, MANNHOLD, MLOGP, SILICOS-IT, SLOGP, and XLOGP3, respectively). Because it is unlikely that all methods share the same bias, one can assume that, at least for this molecule, the error is due to experimental measurement. On the other hand, model 9 underestimates the lipophilicity of 13 compounds (structures and values shown in Chart S4, Supporting Information). Whereas finding common structural features that could explain the error on calculated log P remains difficult, some of these structures involve a long aliphatic chain. This latter can adopt several low energy conformations when solvated, which could influence the SASA parameters described from a unique tridimensional geometry by model 9. Accounting for molecular flexibility should improve the accuracy of the GB/SA methodology in that case. For instance, for the most flexible compound EPI06615, model 9 (single conformer) underestimated the partition coefficient by Δ(log P) = −3.55, while model 3 (200 conformers) underestimated it by only Δ(log P) = −2.65. Overall, the relatively small number of outlying computed values and the theoretically strong hypotheses able to explain the apparent deviations do not challenge the statistical or physical relevance of our physics-based approach to describe *n*-octanol/water partition.

The comparison of model S3 (Table S1, Supporting Information, eq 4 on the BLOCKBUSTER subset) with model 1 (eq 4 on the entire unfiltered training set) also shows lower errors, not significantly stronger correlation and MLR coefficients in better accordance with GB/SA theory for the model trained on the BLOCKBUSTER molecular set. Comparing the results on DRUG and BLOCKBUSTER sets brings to light the negative impact of applying too drastic filters for building data set that are not suited to train MLR models for *n*-octanol/water partition estimation. However, by employing sound softer criteria, it was possible to build a molecular data set large enough and well balanced between diversity and druglikeness. The resulting MLR models showed good agreement with implicit solvation theory and clear statistical relevance. This makes model 9, trained on the BLOCKBUSTER data set and using SASA^{w} for the nonpolar part of the free energy of solvation, a promising log $P_{o/w}$ predictive method for drug design purposes, whose internal validity and robustness have to be assessed by cross-validation and data randomization. Finally, predictive power needs to be evaluated on external test sets.

Internal Validation and Data Randomization. Equation 5 involves only three parameters ($\Delta\Delta G_{solv,elec}$, SASA^{w}, and a constant) to estimate log $P_{o/w}$ and was trained on molecular sets of tens of thousands of experimental observations. This limits dramatically the risk of overfitting. To further probe the internal

Table 4. Five-Fold Cross-Validation for Selected Models[a]

model #	number of molecules (training/validation)	number of conformers	q^2_{CV}[b]	$RMSE_{CV}$[c]	MAE_{CV}[d]
4	14,009/3502	1	0.52	1.33	1.03
5	14,009/3502	20	0.54	1.30	1.00
6	14,009/3502	200	0.55	1.29	0.99
9	9595/2398	1	0.52	1.14	0.89

[a]Data sets were split five folds in training/validation for each model. [b]Cross-validated correlation coefficient. [c]Root mean square error averaged on the five validation sets. [d]Mean average error averaged on the five validation set.

dx.doi.org/10.1021/ci500467k | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

Table 5. Comparative Predictive Capacity on the *Martel* External Test Set Consisting of 706 Diverse Druglike Molecules

method						
name	class	number of parameters[a]	r^b	k^c	RMSE[d]	MAE[e]
model 4	physics-based	3	0.59	0.71	1.56	1.24
model 9	physics-based	3	0.64	0.72	1.40	1.18
KOWWIN	fragmental	400	0.68	0.81	1.40	1.08
MANNHOLD	topological	3	0.64	0.66	1.61	1.39
MLOGP	topological	14	0.45	0.39	2.80	2.51
SILICOS-IT	fragmental/topological	34	0.59	0.82	1.45	1.10
SLOGP	atomic	142	0.63	0.83	1.40	1.07
XLOGP3	atomic/knowledge-based	89	0.77	0.82	1.17	0.92

[a]For physics-based and topological methods: number of parameters in the equation. For fragmental and atomic methods: number of fragments and correction factors. [b]Correlation coefficient. [c]Slope of the correlation line when forced going through the origin. [d]Root mean square error. [e]Mean average error.

validity and the robustness of the most interesting models, we applied five-fold cross-validation and Y-scrambling procedures.

Cross-validation was achieved for model 4, model 5, model 6 (all trained on the 17,511-molecule set) as well as for model 9 (trained on the 11,993-molecule BLOCKBUSTER set). All entries were sorted by experimental log $P_{o/w}$ values, and one in every five molecules—taken in this order—was assigned to the validation set. This procedure was repeated five times starting from entries 1, 2, 3, 4, or 5 of the sorted list. So five different pairs of training/validation sets were established covering the broadest possible value range. For model 4, model 5, and model 6, the number of molecules for training/validation is 14,009/3,502, and for model 9 is 9,595/2,398. For each model, regression coefficients of eq 5 (α_2, β_2, and C_2) were obtained from each training set and were used to calculate log $P_{o/w}$ values of the corresponding validation set. The internal validity of the four models is given by cross-validation metrics in Table 4. All four models show a cross-validated correlation coefficient q^2_{CV} higher than 0.5, which may be considered as a first indication of robustness.[72] Not surprisingly, the most robust is model 6, generated on 200 conformers, with q^2_{CV} = 0.55. More importantly, the error metrics averaged on the predictions of the five validation sets ($RMSE_{CV}$ and MAE_{CV} in Table 4) are identical to the error metrics on the calculation of the whole training sets of parent models (RMSE and MAE in Tables 2 and 3). This constancy of accuracy in the computed values when training or cross-validating the models illustrates the internal quality and robustness of our log $P_{o/w}$ estimation based on GB/SA calculations.

Model 4, model 5, model 6, and model 9 were subjected to further validation by data randomization in order to ascertain that the correlations are not due to serendipity. By randomly swapping the experimental log $P_{o/w}$ values between all molecules while keeping their GB/SA parameters, we checked whether it was possible to fit eq 5 with decent MLR statistics. This procedure was repeated 25,000 times for each model. No statistically significant linear correlation could be found for any randomized set. Indeed, correlation coefficients ranged from 0.00 to 0.03 for model 4, model 5, and model 6 built on the entire data set and from 0.00 to 0.04 for model 9 built on the reduced BLOCKBUSTER training set. The RMSE was 1.92 for model 4, model 5, and model 6 and 1.64 for model 9. This strongly indicates that the occurrence of chance correlations was not present. This is a further solid evidence of the robustness of the presented models linking measured partition and computed GB/SA solvation free energies in *n*-octanol and water.

External Predictive Power Compared to Empirical Predictors. Two different diverse and truly external test sets, the so-called *Martel* and *17-drug* test sets, were built to guarantee a fair comparison between the predictive capacity of six widely used empirical log *P* predictors (KOWWIN, MANNHOLD, MLOGP, SILICOS-IT, SLOGP, and XLOGP3, described in the Methods section) and our most interesting physics-based models for drug design purpose: model 4 and model 9. Those models use one single conformation per molecule, were trained against the entire or the BLOCKBUSTER training sets, respectively, and were validated internally by cross-validation and data randomization.

A proper assessment of the predictive power of models requires an outline of the chemical space covered by the range of every descriptor calculated for all entry in the training set.[72,73] This allows determining whether a test molecule of an external set is found inside or outside the model's applicability domain. In the former case, the prediction comes from an interpolative calculation and is considered more reliable than a prediction determined by extrapolation.[74,75] The applicability domain of model 4 can be defined as a box bounded by extreme values of $SASA^w$ (126.14 to 1960.60 Å^2) and $\Delta\Delta G_{solv,elec}$ (−13.47 to 1.37 kcal/mol). The filtering of the BLOCKBUSTER training set led to a reduced applicability domain for Model 9; $SASA^w$ from 282.98 to 1143.76 Å^2 and $\Delta\Delta G_{solv,elec}$ from −7.54 to 0.36 kcal/mol. One advantage of models relying on two descriptors only is the simple visualization of their applicability domain. In the Supporting Information are the two-dimensional plots of $\Delta\Delta G_{solv,elec}$ versus $SASA^w$ computed for structures contained in the entire 17,511-molecule training set of model 4 (Figure S1, black dots left panel) and in the 11,993-molecule BLOCKBUSTER training set of model 9 (Figure S1, black dots right panel). Apart from confirming the less extended applicability domain for model 9 compared to model 4, these graphs allow defining more precisely the interpolation regions and identifying less populated zones within.

The *Martel* test set contains 706 druglike compounds with high chemical diversity but a quite narrow, rather hydrophobic, experimental log $P_{o/w}$ range (from 0.30 to 6.96). Notably, all measurements were performed in a unique laboratory using a standardized liquid chromatography method.[57] None of the 706 chemical structures belongs to our and most probably to any training set of the six empirical $P_{o/w}$ predictors. Moreover all molecules of the *Martel* test set are inside the applicability domain of both model 4 and model 9 as shown in Figure S1 of the Supporting Information. Therefore, all predictions were interpolative calculations for our models and considered reliable.

dx.doi.org/10.1021/ci500467k | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

Table 6. Comparative External Predictive Capacity on the *17-drug* External Test Set Consisting of Drugs

method						
name	class	number of parameters[a]	r^b	k^c	RMSE[c]	MAE[d]
model 4	physics-based	3	0.94	1.01	0.55	0.41
model 9	physics-based	3	0.94	1.01	0.52	0.38
KOWWIN	fragmental	400	0.86	1.12	1.18	0.89
MANNHOLD	topological	3	0.94	0.91	0.86	0.66
MLOGP	topological	14	0.76	0.79	1.09	0.77
SILICOS-IT	fragmental/topological	34	0.72	1.02	1.30	1.05
SLOGP	atomic	142	0.75	0.98	1.42	0.94
XLOGP3	atomic/knowledge-based	89	0.81	1.15	1.42	0.89

[a]For physics-based and topological methods: number of parameters in the equation. For fragmental and atomic methods: number of fragments and correction factors. [b]Correlation coefficient. [c]Slope of the correlation line when forced going through the origin. [d]Root mean square error. [e]Mean average error.

However, as shown in Table 5, all methods have a modest predictive power on the *Martel* test set. Indeed, correlations between predicted and experimental log $P_{o/w}$ are poor (r ranging from 0.45 for MLOGP to 0.77 for XLOGP3). Moreover, when the correlation lines are forced to go through the origin, their slopes (k ranging from 0.39 for MLOGP to 0.83 for SLOGP) do not satisfy the criteria $0.85 \leq k \leq 1.15$ given by Golbraikh and Tropsha.[76] The inaccuracy of prediction is also indicated by high error metrics. Model 4 is performing slightly worse than model 9. This latter, showing RMSE and MAE of 1.40 and 1.18 log P unit, respectively, compares favorably with the six empirical methods. In detail, model 9 predicted log $P_{o/w}$ with a lower error than both topological predictors MANNHOLD and MLOGP (the latter being in this case the less predictive of all tested methods with an average error exceeding 2.5 log unit). The predictive power of model 9 is similar to that of KOWWIN, SLOGP, and SILICOS-IT, which are three well-trusted and renowned tools. Only XLOGP3 is clearly performing better in this comparative test with an average error lower than one log unit. This is irrespective to the knowledge-based approach of XLOGP3 as XLOGP3-aa (all-atom, results not shown) returned identical predicted values. Taken together, these results make us confident on the fairness and relevance of the *Martel* test set as a benchmarking tool in the sense that it involves molecular structures unrelated to typical data sets used to train log P predictors. Therefore, the prediction is challenging for all tested methods. Noticeably, our GB/SA models performed similarly to the six typical empirical methods tested.

The *17-drug* test set is detailed in Chart 1. Despite its small size, this molecular set includes chemically diverse structures with a relatively broad lipophilicity ranging from log $P_{o/w}$ −1.10 to 5.90. Its significance stems from the facts that molecules are (i) related to clinical drugs and (ii) truly external to any training set of models and methods compared here. The overall predictive data are given in Table 6. In contrast with the *Martel* test set, all methods but MLOGP have a satisfactory external predictive power on the *17-drug* test set. Indeed, predicted and experimental log P values are highly linearly correlated (r ranging from 0.72 for SILICOS-IT to 0.94 for model 4, model 9, and MANNHOLD). Moreover, all models except MLOGP (k = 0.79) satisfy the criteria of $0.85 \leq k \leq 1.15$,[76] with the closest to 1 being model 4 and model 9 (k = 1.01 for both GB/SA models). Regarding error metrics, GB/SA models outperformed all six empirical methods compared. The average errors are as low as 0.41 and 0.38 log unit, whereas the RMSE are 0.55 and 0.52 log unit for model 4 and model 9, respectively. Among empirical methods, only MANNHOLD has a RMSE lower than 1 log unit, and remarkably, XLOGP3, which showed the best predictive capacity for the *Martel* test set, has the highest RMSE with 1.42 log unit (same as SLOGP). This is the symptom of a few large errors. Beside error metrics, it is interesting to note that the strongest linear correlation is for three methods that links *n*-ocatnol/water partition with only three parameters (refer to Table 6). However, as shown in Figure 6, the correlation line of MANNHOLD model intercepts the *Y*-axis at 1.64 log unit, whereas the physics-based model 4 and model 9 show *Y*-intercepts at 0.00 and 0.18, respectively. This deviation in MANNHOLD predictions is also indicated by k dropping to 0.91. At least for the *17-drug* test set, parameters from GB/SA theory are more descriptive of the physics of *n*-octanol/water partition than the simpler but crude counts of carbons and heteroatoms. Furthermore, referring to Figure S1 of the Supporting Information, all molecules of this test set are inside the applicability domain of model 4, and only Spinosad is outside of model 9 interpolation region (pointed in the right graph). Hence the log $P_{o/w}$ value of Spinosad predicted by model 9 was obtained by extrapolation and could be less reliable. Actually, the lipophilicity of this large molecule is clearly overestimated by our GB/SA-based models (see Chart 1). All other predictions are trustworthy because they are computed by interpolative calculations. This analysis shows that GB/SA models outperformed all six topological and fragmental methods for log $P_{o/w}$ prediction of the molecules included in the *17-drug* external test set.

CONCLUSION

The multilinear regression models presented in this article demonstrated the capacity to linearly correlate experimental log $P_{o/w}$ with solvation free energy computed in implicit media by the GB/SA approach. In particular, the statistical quality of model 9, trained on a chemically diverse druglike data set of 11,993 molecular structures (r = 0.72, MAE = 0.89, and RMSE = 1.14 versus experimental log P) and its internal validity (q^2_{CV} = 0.52, MAE_{CV} = 0.89, and $RMSE_{CV}$ = 1.14 for five-fold cross-validation) were satisfactory. Moreover, the predictive capacity of model 9 was good in comparison with six reference empirical predictors on two different external test sets: the diverse *Martel* test set (N = 706, r = 0.64, MAE = 1.18, and RMSE = 1.40) and the "real"*17-drug* test set (N = 17, r = 0.94, MAE = 0.38, and RMSE = 0.52).

It has to be emphasized that our models only rely on two physical descriptors ($\Delta\Delta G_{solv,elec}$ and $SASA^w$) and a regression constant. This makes our methodology less prone to overfitting compared to purely empirical methods involving more

dx.doi.org/10.1021/ci500467k | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

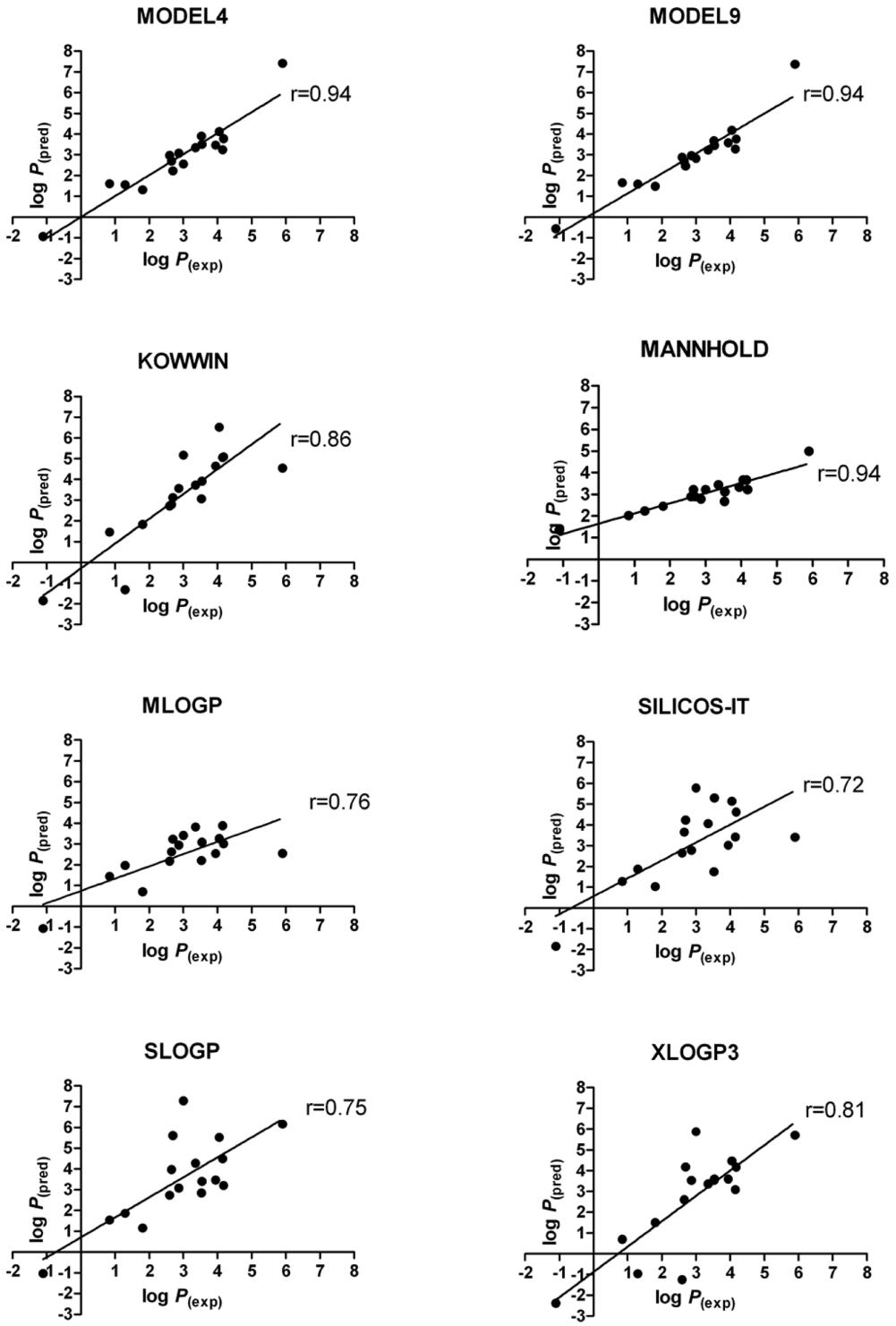


Figure 6. Predicted log $P_{o/w}$ (log $P_{(pred)}$) by two GB/SA models (model 4 and model 9) and six empirical methods (KOWWIN, MANNHOLD, MLOGP, SILICOS-IT, SLOGP, and XLOGP3) versus experimental values (log $P_{(exp)}$) on the *17-drug* external test set. The lines represent the linear correlations, whose coefficient (r) is noted.

parameters. We decided not to implement descriptors other than those describing the solvation free energy in order to keep the approach purely physics-based. The aim was to favor robustness by increasing the reliability of predictions outside applicability domain and to make the models as global as possible, although focusing the predictive power assessment on druglike external sets. Furthermore, a three-parameter linear model is simple enough for straightforward interpretation in chemical terms. The objective here is to help understanding the subtle structural and electronic factors governing the partition of a specific compound (such as regioisomerism, sterical hindrance, or intramolecular interactions) and therefore to support efficiently the design of molecules modified to target the desired lipophilicity. Because our description of partition has a strong rationale based on GB/SA theory, the lipophilic contributions are defined at the atomic level. In contrast with empirical atomic methods (e.g., ALOGP,

dx.doi.org/10.1021/ci500467k | *J. Chem. Inf. Model.* XXXX, XXX, XXX−XXX

SLOGP, or XLOGP) and because the degree of buriedness of each atom is taken into account, the overall shape of the molecule influences the contributive values. It follows that each individual atomic contribution to log $P_{o/w}$ is not fixed in a fragmental system but calculated for each input molecule because distinctively dependent on the molecular environment. In Figures 7 and 8 are

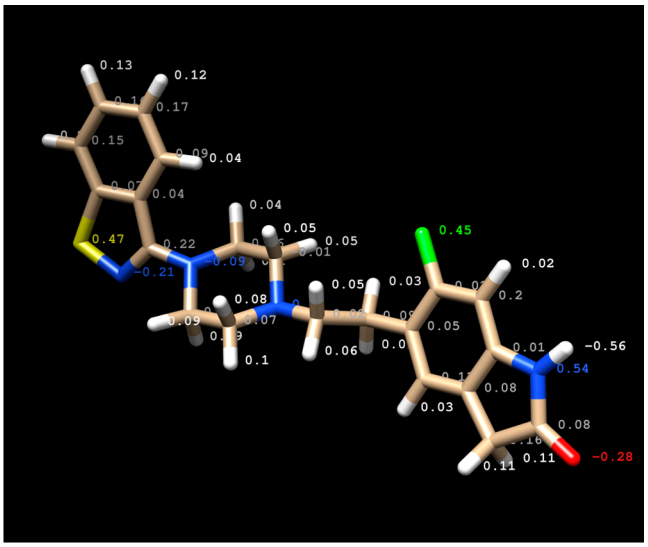


Figure 7. Atomic contributions to *n*-octanol/water partition as computed by the GB/SA approach and plotted on the Ziprasidone conformer generated by OMEGA. Label colors relate to elements. The sum of atomic contributions corresponds to the log $P_{o/w}$ calculated by model 9, log $P_{(GB/SA)}$ = 3.94.

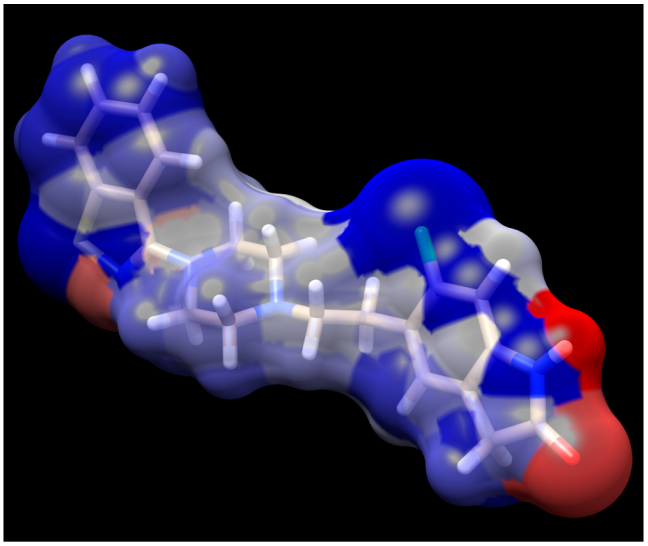


Figure 8. Atomic contributions as defined by model 9 mapped on the semitransparent SASA around the Ziprasidone conformer generated by OMEGA. Color-scale is from red for the most polar to blue for the most hydrophobic regions; white depicts close to zero regions.

examples of tridimensional representations of the lipophilic contributions computed according to model 9 for the antipsychotic drug Ziprasidone, which belongs to our training set (DB00246). The conformer generated by OMEGA was loaded in the UCSF Chimera molecular graphics environment[77] (version 1.9, 2014, http://www.cgl.ucsf.edu/chimera/). Each atom can be labeled by its numerical contribution to the *n*-octanol/water partition coefficient as in Figure 7. Summing all contributions returns the calculated log $P_{(pred)}$ = 3.94, which in the case of Ziprasidone compares well with the experimentally measured log $P_{(exp)}$ = 3.80. It is worth remembering that this calculation is based on GB and SA values only and does not require any empirical corrective term. Figure 8 illustrates another level of description by projecting lipophilic contributions on the SASA as computed by Chimera. The lipophilicity map highlights a few clearly polar regions in red (mainly, the lactam moiety, the nitrogen of the benzothiazole ring, and to a much lesser extent the piperazine) and largely hydrophobic regions in blue (mainly, the chloro aromatic substitutent, the sulfur of the benzothiazole ring, and the aromatic cycles), which explain the predominantly hydrophobic nature of the molecule with a global log $P_{o/w}$ around 4. This very detailed description of lipophilicity shows a realistic picture of the physicochemical nature of the molecule under study in the sense that it explains how the molecule is perceived by its environment. Altogether, these analyses indicate to which extent our approach can be informative for computed-assisted drug design, including optimization steps involving tridimensional graphics visualization.

The methodology presented in this article, which links solvation free energy and *n*-octanol/water partition coefficient, will be made available to the scientific community through a prediction tool called iLOGP, for implicit log *P*. However, the methodology is obviously perfectible. The room for improvement lies first in the fact that although bearing the great advantage of being large, diverse, and cleansed, the structural description of the 17,511-molecule data set lacks stereochemical information. Indeed, the sources of chemical information were too heterogeneous and inaccurate to allow defining the stereochemistry for all entries. As a consequence, the tridimensional geometry was randomized with respect to chiral centers and double-bond configurations. Because GB/SA calculations are performed on the tridimensional structure of the compound, it cannot be excluded that this uncertainty impacts the calculation of GB/SA parameters depending on the chemical nature of the molecule under consideration. It is therefore likely that some noise is added to the statistical model. The influence of stereochemistry shall be evaluated at the time of building the actual predictive log $P_{o/w}$ model by retraining the MLR on smaller molecular sets with defined stereochemical information. Second, we have shown that describing the molecular structure by multiple geometries slightly improved the statistical relevance and internal quality of the GB/SA MLR models for *n*-octanol/water partition. However, in the drug design context where numerous compounds are treated, this improvement cannot counterbalance the loss in computational efficiency—even though various levels of approximation could be chosen as a function of the number of molecules to be handled. Nevertheless, this indicates the important influence of the tridimensional geometry for such predictive models. Whereas here we have focused on a unique conformation generator, OMEGA, mainly because of its speed, other tridimensional engines do exist and could be evaluated at the time of making our models evolve toward a production cheminformatic tool. Third, unavoidable approximations in the physical models impact the calculation of the solvation free energies in implicit solvents. Noticeably, the MLR equation parameters rely on the Born radii definition and force field parametrization. However, any improvement in new implementations of GBMV2, CHARMM, or SwissParam can straightforwardly be included in our methodological workflow. Finally, the computation of log *P* from a single SMILES currently

takes 1 to 2 s depending on the size of the molecule. This is fast enough for most of the drug design activities. Although this speed will never be at the level of topological or fragmental methods, which avoid the demanding step of generating tridimensional geometries, code and implementation improvements are expected to reduce the time of computation substantially.

The fact remains that the size, diversity, and quality of the employed data together with the straightforward methodology involving molecular mechanics, GB, and SASA are adequate to produce simple yet physically meaningful models able to strongly correlate the computation of implicit solvation energies with experimental *n*-octanol/water partition measurements. The promising robustness and predictive power demonstrated by the models encourage the future addition of the iLOGP predictor to the portfolio of drug design and molecular modeling tools developed by the SIB Swiss Institute of Bioinformatics, which currently includes SwissBioisostere,[78] SwissDock,[79] SwissParam,[63] SwissSidechain,[80] and SwissTargetPrediction.[81] The final implementation of iLOGP will benefit from other empirical methods through a consensus approach. Indeed, it has been demonstrated that consensus values of log $P_{o/w}$ averaging over multiple computational methods can be accurate at predicting experimental *n*-octanol/water partition coefficients.[16] The computational methods involved in the consensus should be as diverse as possible in the description of the chemical structure. In that respect, our original physics-based approach is a good candidate to complement favorably typical fragmental and topological methods.

■ ASSOCIATED CONTENT

Supporting Information

Data as mentioned in the text. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Authors

*E-mail: olivier.michielin@isb-sib.ch (O.M.).
*E-mail: vincent.zoete@isb-sib.ch (V.Z.).

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors are deeply thankful to the SIB Swiss Institute of Bioinformatics and to its center for high-performance computing (Vital-IT, http://www.vital-it.ch) for providing computational resources. Gratitude is also expressed to Ute Röhrig for the thorough review of the manuscript, as well as to other Molecular Modeling Group members: Justyna Iwaszkiewicz, Maria Johansson, David Gfeller, Aurélien Grosdidier, Prasad Chaskar, and Matthias Wirth for their helpful advises. Special thanks go to Sophie Martel at Debiopharm for sharing a bit of her great expertise in experimental physicochemistry. We are grateful to OpenEye Scientific Software, Inc. (http://www.eyesopen.com) and ChemAxon, Ltd. (http://www.chemaxon.com) for the respective academic license agreements. Molecular graphics and tridimensional pictures were produced by the UCSF Chimera package (http://www.cgl.ucsf.edu/chimera/).

■ REFERENCES

(1) Plika, V.; Testa, B.; van de Waterbeemd, H. Lipophilicity: The Empirical Tool and the Fundamental Objective. An Introduction. In *Lipophilicity in Drug Action and Toxicology*; Methods and Principles in Medicinal Chemistry; Wiley-VCH Verlag GmbH: Weinheim, Germany, 1996; pp 1−6.

(2) Liu, X.; Testa, B.; Fahr, A. Lipophilicity and its relationship with passive drug permeation. *Pharm. Res.* **2010**, *28*, 962−977.

(3) Garrido, N. M.; Queimada, A. J.; Jorge, M.; Macedo, E. A.; Economou, I. G. 1-Octanol/water partition coefficients of N-alkanes from molecular simulations of absolute solvation free energies. *J. Chem. Theory Comput.* **2009**, *5*, 2436−2446.

(4) Arnott, J. A.; Planey, S. L. The influence of lipophilicity in drug discovery and design. *Expert Opin. Drug Discovery* **2012**, *7*, 863−875.

(5) Yazdanian, M. Overview of determination of biopharmaceutical properties for development candidate selection. *Curr. Protoc. Pharmacol.* **2013**, *9.17*, 1−8.

(6) van de Waterbeemd, H.; Gifford, E. ADMET in silico modelling: Towards prediction paradise? *Nat. Rev. Drug Discovery* **2003**, *2*, 192−204.

(7) Dearden, J. C. In silicoprediction of ADMET properties: How far have we come? *Expert Opin. Drug Metab. Toxicol.* **2007**, *3*, 635−639.

(8) Kenny, J. R. Predictive DMPK: In silico ADME predictions in drug discovery. *Mol. Pharmaceutics* **2013**, *10*, 1151−1152.

(9) Smith, D. A. Evolution of ADME science: Where else can modeling and simulation contribute? *Mol. Pharmaceutics* **2013**, *10*, 1162−1170.

(10) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Delivery Rev.* **2001**, *46*, 3−26.

(11) Nurisso, A.; Bravo, J.; Carrupt, P.-A.; Daina, A. Molecular docking using the molecular lipophilicity potential as hydrophobic descriptor: Impact on GOLD docking performance. *J. Chem. Inf. Model.* **2012**, *52*, 1319−1327.

(12) Shoichet, B. K.; Leach, A. R.; Kuntz, I. D. Ligand solvation in molecular docking. *Proteins* **1999**, *34*, 4−16.

(13) Gleeson, M. P.; Hersey, A.; Montanari, D.; Overington, J. Probing the links between in vitro potency, ADMET and physicochemical parameters. *Nat. Rev. Drug Discovery* **2011**, *10*, 197−208.

(14) Meanwell, N. A. Improving drug candidates by design: A focus on physicochemical properties as a means of improving compound disposition and safety. *Chem. Res. Toxicol.* **2011**, *24*, 1420−1456.

(15) Kolář, M.; Fanfrlík, J.; Lepšík, M.; Forti, F.; Luque, F. J.; Hobza, P. Assessing the accuracy and performance of implicit solvent models for drug molecules: Conformational ensemble approaches. *J. Phys. Chem. B* **2013**, *117*, 5950−5962.

(16) Mannhold, R.; Poda, G. I.; Ostermann, C. Calculation of molecular lipophilicity: State-of-the-Art and comparison of log P methods on more than 96,000 compounds. *J. Pharm. Sci.* **2009**, *98*, 861−893.

(17) Tetko, I. V.; Poda, G. I.; Ostermann, C.; Mannhold, R. Large-scale evaluation of log P predictors: Local corrections may compensate insufficient accuracy and need of experimentally testing every other compound. *Chem. Biodiversity* **2009**, *6*, 1837−1844.

(18) Klopman, G.; Li, J.-Y.; Wang, S.; Dimayuga, M. Computer automated log P calculations based on an extended group contribution approach. *J. Chem. Inf. Model.* **1994**, *34*, 752−781.

(19) Meylan, W. M.; Howard, P. H. Estimating log P with atom/fragments and water solubility with log P. *Perspect. Drug Discovery Des.* **2000**, *19*, 67−84.

(20) Ghose, A. K.; Crippen, G. M. Atomic physicochemical parameters for three-dimensional structure-directed quantitative structure-activity relationships. I. Partition coefficients as a measure of hydrophobicity. *J. Comput. Chem.* **1986**, *7*, 565−577.

(21) Ghose, A. K.; Viswanadhan, V. N.; Wendoloski, J. J. Prediction of hydrophobic (lipophilic) properties of small organic molecules using fragmental methods: An analysis of ALOGP and CLOGP methods. *J. Phys. Chem. A* **1998**, *102*, 3762−3772.

(22) Wang, R.; Fu, Y.; Lai, L. A new atom-additive method for calculating partition coefficients. *J. Chem. Inf. Model.* **1997**, *37*, 615−621.

(23) Cheng, T.; Zhao, Y.; Li, X.; Lin, F.; Xu, Y.; Zhang, X.; Li, Y.; Wang, R.; Lai, L. Computation of octanol−water partition coefficients by

guiding an additive model with knowledge. *J. Chem. Inf. Model.* **2007**, *47*, 2140−2148.
(24) Wildman, S. A.; Crippen, G. M. Prediction of physicochemical parameters by atomic contributions. *J. Chem. Inf. Model.* **1999**, *39*, 868−873.
(25) Moriguchi, I.; Shuichi, H.; Nakagome, I.; Hirano, H. Comparison of reliability of log P values for drugs calculated by several methods. *Chem. Pharm. Bull.* **1994**, *42*, 976−978.
(26) Moriguchi, I.; Shuichi, H.; Liu, Q.; Nakagome, I.; Matsushita, Y. Simple method of calculating octanol/water partition coefficient. *Chem. Pharm. Bull.* **1992**, *40*, 127−130.
(27) Reynolds, C. H. Estimating lipophilicity using the GB/SA continuum solvation model: A direct method for computing partition coefficients. *J. Chem. Inf. Model.* **1995**, *35*, 738−742.
(28) Jorgensen, W. L. Free energy calculations: A breakthrough for modeling organic chemistry in solution. *Acc. Chem. Res.* **1989**, *22*, 184−189.
(29) Kollman, P. Free energy calculations: Applications to chemical and biochemical phenomena. *Chem. Rev.* **1993**, *93*, 2395−2417.
(30) DeBolt, S. E.; Kollman, P. A. Investigation of structure, dynamics, and solvation in 1-octanol and its water-saturated solution: Molecular dynamics and free-energy perturbation studies. *J. Am. Chem. Soc.* **1995**, *117*, 5316−5340.
(31) Essex, J. W.; Reynolds, C. A.; Richards, W. G. Theoretical determination of partition coefficients. *J. Am. Chem. Soc.* **1992**, *114*, 3634−3639.
(32) Duffy, E. M.; Jorgensen, W. L. Prediction of properties from simulations: Free energies of solvation in hexadecane, octanol, and water. *J. Am. Chem. Soc.* **2000**, *122*, 2878−2888.
(33) Gilson, M. K.; Honig, B. Calculation of the total electrostatic energy of a macromolecular system: Solvation energies, binding energies, and conformational analysis. *Proteins* **1988**, *4*, 7−18.
(34) Born, M. Volumen und hydratationswärme der ionen. *Z. Phys.* **1920**, *1*, 45−48.
(35) Still, W. C.; Tempczyk, A.; Hawley, R. C.; Hendrickson, T. Semianalytical treatment of solvation for molecular mechanics and dynamics. *J. Am. Chem. Soc.* **1990**, *112*, 6127−6129.
(36) Qiu, D.; Shenkin, P. S.; Hollinger, F. P.; Still, W. C. The GB/SA continuum model for solvation. A fast analytical method for the calculation of approximate Born radii. *J. Phys. Chem. A* **1997**, *101*, 3005−3014.
(37) Jorgensen, W. L.; Tirado-Rives, J. The OPLS [optimized potentials for liquid simulations] potential functions for proteins, energy minimizations for crystals of cyclic peptides and crambin. *J. Am. Chem. Soc.* **1988**, *110*, 1657−1666.
(38) Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J. Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. *J. Am. Chem. Soc.* **1996**, *118*, 11225−11236.
(39) Best, S. A.; Merz, K. M., Jr; Reynolds, C. H. GB/SA-based continuum solvation model for octanol. *J. Phys. Chem. B* **1997**, *101*, 10479−10487.
(40) Halgren, T. A. Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94. *J. Comput. Chem.* **1998**, *17*, 490−519.
(41) Halgren, T. A. Merck molecular force field. II. MMFF94 van der Waals and electrostatic parameters for intermolecular interactions. *J. Comput. Chem.* **1996**, *17*, 520−553.
(42) Halgren, T. A. Merck molecular force field. V. Extension of MMFF94 using experimental data, additional computational data, and empirical rules. *J. Comput. Chem.* **1996**, *17*, 616−641.
(43) Halgren, T. A.; Nachbar, R. B. Merck molecular force field. IV. Conformational energies and geometries for MMFF94. *J. Comput. Chem.* **1998**, *17*, 587−615.
(44) Halgren, T. A. Merck molecular force field. III. Molecular geometries and vibrational frequencies for MMFF94. *J. Comput. Chem.* **1996**, *17*, 553−586.
(45) Best, S. A.; Merz, K. M.; Reynolds, C. H. Free energy perturbation study of octanol/water partition coefficients: Comparison with continuum GB/SA calculations. *J. Phys. Chem. B* **1999**, *103*, 714−726.
(46) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *J. Am. Chem. Soc.* **1995**, *117*, 5179−5197.
(47) Totrov, M. Accurate and efficient generalized Born model based on solvent accessibility: Derivation and application for logP octanol/water prediction and flexible peptide docking. *J. Comput. Chem.* **2004**, *25*, 609−619.
(48) Lee, M. S.; Salsbury, F. R.; Brooks, C. L. Novel generalized Born methods. *J. Chem. Phys.* **2002**, *116*, 10606−10614.
(49) Lee, M. S.; Feig, M.; Salsbury, F. R.; Brooks, C. L. New analytic approximation to the standard molecular volume definition and its application to generalized Born calculations. *J. Comput. Chem.* **2003**, *24*, 1348−1356.
(50) Brooks, B. R.; Brooks, C. L.; MacKerell, A. D.; Nilsson, L.; Petrella, R. J.; Roux, B.; Won, Y.; Archontis, G.; Bartels, C.; Boresch, S.; Caflisch, A.; Caves, L.; Cui, Q.; Dinner, A. R.; Feig, M.; Fischer, S.; Gao, J.; Hodoscek, M.; Im, W.; Kuczera, K.; Lazaridis, T.; Ma, J.; Ovchinnikov, V.; Paci, E.; Pastor, R. W.; Post, C. B.; Pu, J. Z.; Schaefer, M.; Tidor, B.; Venable, R. M.; Woodcock, H. L.; Wu, X.; Yang, W.; York, D. M.; Karplus, M. CHARMM: The biomolecular simulation program. *J. Comput. Chem.* **2009**, *30*, 1545−1614.
(51) Meylan, W. M.; Howard, P. H.; Boethling, R. S. Improved method for estimating water solubility from octanol/water partition coefficient. *Environ. Toxicol. Chem.* **1996**, *15*, 100−106.
(52) Knox, C.; Law, V.; Jewison, T.; Liu, P.; Ly, S.; Frolkis, A.; Pon, A.; Banco, K.; Mak, C.; Neveu, V.; Djoumbou, Y.; Eisner, R.; Guo, A. C.; Wishart, D. S. DrugBank 3.0: A comprehensive resource for "omics" research on drugs. *Nucleic Acids Res.* **2010**, *39*, D1035−D1041.
(53) Ihlenfeldt, W. D.; Voigt, J. H.; Bienfait, B.; Oellien, F.; Nicklaus, M. C. Enhanced CACTVS browser of the Open NCI database. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 46−57.
(54) O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. OpenBabel: An open chemical toolbox. *J. Cheminform.* **2011**, *3*, 33.
(55) Wang, Y.; Xiao, J.; Suzek, T. O.; Zhang, J.; Wang, J.; Zhou, Z.; Han, L.; Karapetyan, K.; Dracheva, S.; Shoemaker, B. A.; Bolton, E.; Gindulyte, A.; Bryant, S. H. PubChem's bioassay database. *Nucleic Acids Res.* **2011**, *40*, D400−D412.
(56) Oprea, T. I. Property distribution of drug-related chemical databases. *J. Comput. Aided Mol. Des* **2000**, *14*, 251−264.
(57) Martel, S.; Gillerat, F.; Carosati, E.; Maiarelli, D.; Tetko, I. V.; Mannhold, R.; Carrupt, P.-A. Large, chemically diverse dataset of logP measurements for benchmarking studies. *Eur. J. Pharm. Sci.* **2013**, *48*, 21−29.
(58) Irwin, J. J.; Shoichet, B. K. ZINC − A free database of commercially available compounds for virtual screening. *J. Chem. Inf. Model.* **2005**, *45*, 177−182.
(59) Law, V.; Knox, C.; Djoumbou, Y.; Jewison, T.; Guo, A. C.; Liu, Y.; Maciejewski, A.; Arndt, D.; Wilson, M.; Neveu, V.; Tang, A.; Gabriel, G.; Ly, C.; Adamjee, S.; Dame, Z. T.; Han, B.; Zhou, Y.; Wishart, D. S. DrugBank 4.0: Shedding new light on drug metabolism. *Nucleic Acids Res.* **2013**, *42*, D1091−D1097.
(60) Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Model.* **1988**, *28*, 31−36.
(61) Hawkins, P. C. D.; Nicholls, A. Conformer generation with OMEGA: Learning from the data set and the analysis of failures. *J. Chem. Inf. Model.* **2012**, *52*, 2919−2936.
(62) Hawkins, P. C. D.; Skillman, A. G.; Warren, G. L.; Ellingson, B. A.; Stahl, M. T. Conformer generation with OMEGA: Algorithm and validation using high quality structures from the protein databank and cambridge structural database. *J. Chem. Inf. Model.* **2010**, *50*, 572−584.
(63) Zoete, V.; Cuendet, M. A.; Grosdidier, A.; Michielin, O. SwissParam: A fast force field generation tool for small organic molecules. *J. Comput. Chem.* **2011**, *32*, 2359−2368.

(64) Knight, J. L.; Brooks, C. L., III. Surveying implicit solvent models for estimating small molecule absolute hydration free energies. *J. Comput. Chem.* **2011**, *32*, 2909−2923.
(65) Ryckaert, J.-P.; Ciccotti, G.; Berendsen, H. J. C. Numerical integration of the Cartesian equations of motion of a system with constraints: Molecular dynamics of N-alkanes. *J. Comput. Phys.* **1977**, *23*, 327−341.
(66) Mauri, A.; Consonni, V.; Pavan, M.; Todeschini, R. DRAGON software: An easy approach to molecular descriptor calculations. *MATCH* **2006**, *56*, 237−248.
(67) Leahy, D. E. Intrinsic molecular volume as a measure of the cavity term in linear solvation energy relationships: Octanol−water partition coefficients and aqueous solubilities. *J. Pharm. Sci.* **1986**, *75*, 629−636.
(68) el Tayar, N.; Tsai, R. S.; Testa, B.; Carrupt, P. A.; Leo, A. Partitioning of solutes in different solvent systems: The contribution of hydrogen-bonding capacity and polarity. *J. Pharm. Sci.* **1991**, *80*, 590−598.
(69) Franks, N. P.; Abraham, M. H.; Lieb, W. R. Molecular organization of liquid−octanol: An X-ray diffraction analysis. *J. Pharm. Sci.* **1993**, *82*, 466−470.
(70) Hasel, W.; Hendrickson, T. F.; Still, W. C. A rapid approximation to the solvent accessible surface areas of atoms. *Tetrahedron Comput. Methodol.* **1988**, *1*, 103−116.
(71) Gohlke, H.; Kiel, C.; Case, D. A. Insights into protein−protein binding by binding free energy calculation and free energy decomposition for the Ras−Raf and Ras−RalGDS complexes. *J. Mol. Biol.* **2003**, *330*, 891−913.
(72) Tropsha, A.; Gramatica, P.; Gombar, V. K. The importance of being earnest: Validation is the absolute essential for successful application and interpretation of QSPR models. *QSAR Comb. Sci.* **2003**, *22*, 69−77.
(73) Gramatica, P. Principles of QSAR models validation: Internal and external. *QSAR Comb. Sci.* **2007**, *26*, 694−701.
(74) Tetko, I. V.; Sushko, I.; Pandey, A. K.; Zhu, H.; Tropsha, A.; Papa, E.; Oberg, T.; Todeschini, R.; Fourches, D.; Varnek, A. Critical assessment of QSAR models of environmental toxicity against *Tetrahymena pyriformis*: Focusing on applicability domain and overfitting by variable selection. *J. Chem. Inf. Model.* **2008**, *48*, 1733−1746.
(75) Sahigara, F.; Mansouri, K.; Ballabio, D.; Mauri, A.; Consonni, V.; Todeschini, R. Comparison of different approaches to define the applicability domain of QSAR models. *Molecules* **2012**, *17*, 4791−4810.
(76) Golbraikh, A.; Tropsha, A. Beware of Q2! *J. Mol. Graph. Model.* **2002**, *20*, 269−276.
(77) Pettersen, E. F.; Goddard, T. D.; Huang, C. C.; Couch, G. S.; Greenblatt, D. M.; Meng, E. C.; Ferrin, T. E. UCSF chimera − A visualization system for exploratory research and analysis. *J. Comput. Chem.* **2004**, *25*, 1605−1612.
(78) Wirth, M.; Zoete, V.; Michielin, O.; Sauer, W. H. B. SwissBioisostere: A database of molecular replacements for ligand design. *Nucleic Acids Res.* **2013**, *41*, D1137−D1143.
(79) Grosdidier, A.; Zoete, V.; Michielin, O. SwissDock, a protein−small molecule docking Web service based on EADock DSS. *Nucleic Acids Res.* **2011**, *39*, W270−W277.
(80) Gfeller, D.; Michielin, O.; Zoete, V. SwissSidechain: A molecular and structural database of non-natural sidechains. *Nucleic Acids Res.* **2013**, *41*, D327−D332.
(81) Gfeller, D.; Grosdidier, A.; Wirth, M.; Daina, A.; Michielin, O.; Zoete, V. SwissTargetPrediction: A Web server for target prediction of bioactive small molecules. *Nucleic Acids Res.* **2014**, *42*, W32−W38.