“Big Data” Fast Chemoinformatics Model to Predict Generalized Born Radius and Solvent Accessibility as a Function of Geometry

Dragos Horvath, Gilles Marcou, Alexandre Varnek

"Big Data" Fast Chemoinformatics Model to Predict Generalized Born Radius and Solvent Accessibility as a Function of Geometry

Journal:	*Journal of Chemical Information and Modeling*
Manuscript ID	ci-2019-01172s.R2
Manuscript Type:	Article
Date Submitted by the Author:	n/a
Complete List of Authors:	Horvath, Dragos; Universite de Strasbourg, Laboratoire de Chemoinformatique-UMR7140 Marcou, Gilles; Universite de Strasbourg, Faculty of Chemistry Varnek, Alexandre; Universite de Strasbourg,

"Big Data" Fast Chemoinformatics Model to Predict Generalized Born Radius and Solvent Accessibility as a Function of Geometry

Dragos Horvath [a*], Gilles Marcou [a], Alexandre Varnek [a]

[a] Laboratory of Chemoinformatics, UMR 7140 University of Strasbourg/CNRS, 4 rue Blaise Pascal, 67000 Strasbourg, France

* Corresponding Author (dhorvath@unistra.fr)

Abstract.

The Generalized Born (GB) solvent model is offering the best accuracy/computing effort ratio and yet requires drastic simplifications to estimate of the Effective Born Radii (EBR), in bypassing a too expensive volume integration step. EBR are a measure of the degree of burial of an atom, and not very sensitive to small changes of geometry: in Molecular Dynamics (MD), the costly EBR update procedure is not mandatory at every step. This work however aims at implementing a GB model into the S4MPLE evolutionary algorithm, with mandatory EBR updates at each step triggering arbitrarily large geometric changes. Therefore, a Quantitative Structure-Property Relationship (QSPR) has been developed in order to express the EBRs as a linear function of both topological neighborhood and the geometric occupancy of the space around atoms. A training set of 810 molecular systems, starting from fragment-like, to drug-like compounds, proteins, host-guest systems and ligand-protein complexes has been compiled. For each species, S4MPLE generated several hundreds of random conformers. For each atom in each geometry of each species, its “standard” EBR was calculated by numeric integration and associated to topological and geometric descriptors of the atom neighborhood. This training set (EBR, atom descriptors) involving >5M entries was subjected to a boot-strapping multilinear regression process with descriptor selection. In parallel, the strategy was repurposed to also learn atomic solvent-accessible areas (SA), based on the same descriptors. Resulting linear equations were challenged to predict EBR and SA values, respectively, for a similarly compiled external set of >2,000 new molecular systems. Solvation energies calculated with estimated EBR and SA match "standard" energies within the typical error of a force-field based approach (a few kcal/mol). Given the extreme diversity of molecular systems covered by the model, this simple EBR/SA estimator covers a vast applicability domain.

Keywords: Generalized Born model, continuum solvent models, multilinear regression, fast approximate estimation of Born radii

Abbreviations: GB – generalized Born, EBR – Effective Born Radius (of an atom), F – Fraction of Accessible Surface Area (of an atom), SA – Surface Area

1. Introduction.

After the initial introduction of the Generalized Born[1] (GB) solvent model[2, 3], its relative simplicity and good performance prompted the community[4, 5] to envisage various workarounds around its main bottleneck: the need to estimate "Effective" Born Radii (EBR) of atoms. These key parameters are related to the degree of burial of atoms inside the low-dielectric solute. For a monoatomic ion, the EBR equals its ionic radius as postulated in Born's ion solvation theory[6]. For a charged atom in a molecule, its EBR would be the radius value required to have Born's above-cited formula return the actual solvation energy of the low-dielectric "blob" representing the actual molecule, with partial charges of all other atoms set to zero[7]. The presence of other atoms surrounding the atom of interest in this "Gedankenexperiment" has a two-fold impact. First, the dielectric interface is pushed away from the considered atom, as the low-dielectric neighbors displace the solvent. Thus, for the case of a charged atom, completely screened from the solvent at the center of a nearly spherical globular protein, its EBR would roughly match the radius of the molecular sphere (tens of Å). However, if the spherical symmetry of the system is broken, the polarization of the dielectric interface can no longer be expressed by analytical functions of the molecular geometry, but requires complex volume or boundary-based integration of the Poisson-Boltzmann equation (PB)[8-11]. The standard approach to estimate EBR values by volume integration[12, 13]:

$$EBR_i^{-1} = R_i^{-1} - \frac{1}{4\pi}\iiint_{R_i}^{\infty} \beta(\vec{r})\frac{d^3\vec{r}}{r^4} \tag{1}$$

is performed in spherical coordinates centered on atom *i,* over all points outside the actual atomic sphere of radius R_i, where the burial status β is a toggle function equaling one in all the space points within the solute (inside the molecular or van der Waals surface, respectively), and zero within the solvent-occupied space. Albeit still too time-consuming to be used at every step of molecular simulation, this approach is already a stark simplification, ignoring any distortion of the dielectric displacement vector field caused by an arbitrary charge distribution (lacking spherical symmetry). Precise[7] EBR estimation *via* PB calculations would significantly enhance the accuracy of the GB model. Interestingly, the authors highlighted that usage of these accurate EBR in a 6 ns molecular dynamics (MD) simulation, all while assumed to be constant and equal to the ones derived on hand of the departure geometry, still outperforms the "classical" GB model. EBR values tend to be rather weakly affected by small geometric fluctuations. Thus, for application[14] in MD simulations, EBR update needs not to be performed very often – certainly not at every step. Unfortunately, non-physical conformational sampling tools like the evolutionary process within our in-house software S4MPLE[15-17] perform arbitrarily remote jumps in conformational space, rendering EBR recalculation mandatory after each such step. In this herein envisaged scenario, the need for rapid EBR assessment is much more stringent.

A fast evaluation of EBRs must avoid the explicit volume integration prone by equation (1). This is easier to achieve if the low-dielectric interior of the solute is defined by the van der Waals rather than the molecular surface. If so, approximate analytical solutions of the integral can be proposed, and empirically parameterized in order to compensate for systematic errors[5, 18-24]. Herein, EBR values are rendered as (complex) empirical functions of interatomic distances, starting from the analytical expression of the integral (1) for the simple case when atom i is in presence of a single and non-overlapping atom j ($r_{ij}>R_i+R_j$). Then, empirical scaling factors are introduced as corrections for the actual overlap of atomic spheres. GB terms based on these pairwise approximations are analytically differentiable[25]. Such pairwise decomposition schemes were even tailor-made for proteins only, exploiting the specificity of back bone versus side chain atoms[26]. Alternatively, since the dielectric boundary is intrinsically ill-defined, an appealing alternative is[27, 28] to express $\beta(\vec{r})$ as a sum of Gaussians centered on atoms, herewith enabling an analytical approach to the integrals. GB models were also proposed as docking-specific, active site grid-based reformulations[29, 30], and recently adapted for GPU[31, 32] or hybrid CPU/GPU computing[33, 34].

However, with the van der Waals surface as dielectric boundary, the reentrant volumes close to the intersection of atomic spheres erroneously count as solvated, whereas they should in principle be also assigned β=1. Empirical correction schemes are specifically targeting[12, 35-37] this problem – but, in doing so, they often specialize on a given compound class (macromolecules, drug-like compounds, *etc*). Empirically parameterized solutions nevertheless rely on quite complex expressions and are typically system size-dependent: accurate solutions for small molecules[38] trigger large systematic errors for macromolecules – reciprocally, protein[39]- and protein-ligand-tailored approaches[40] may not be accurate when applied to other host-guest systems, for example[41].

Also, an accurate estimation of EBR values is in principle not compatible with the use of distance cutoffs, which are the most widely used way to speed up molecular mechanics calculations. Stopping the integration in equation (1) at some finite cutoff *c*, or alternatively truncating pairwise analytical terms to atoms within this cutoff distance would automatically result in an estimated $EBR \leq c$, irrespectively how deeply the atom is actually buried. In terms of absolute estimations of EBR values, this may result in errors of tens of Ǻ. Nevertheless, this might be acceptable in as far as the "self-contribution" of atom *i* to the GB solvation energy (of the order of Q_i^2/EBR_i, with Q_i being the partial charge) is already small enough at $EBR=c$ and will not fluctuate significantly during the simulation of interest.

As noted, the "standard" GB model based on equation (1) is already a stark simplification. More fundamentally, the hypothesis of a structureless dielectric solvent is *per se* flawed whenever any specific, strong solute-water interactions come into play. Even the mathematically exact solutions are tributary to some hand-waving definition of the dielectric interface: the atomic radii used to define it (including or not some offset value) are best viewed as fitable parameters of the model. Thus, no model can be better than the weakest of its hypotheses.

Therefore, our approach consists in voluntarily giving up the design of simplified expressions of the integral in favor of a machine-learning approach aimed at approximating EBR values as functions of simple topological and geometric indices. We aimed at a global model, applicable for both macromolecules and small ligands (even fragments, given the interest in Fragment-based drug design[42]). Designed to enter the general-purpose S4MPLE program, it should be applicable for both conformational sampling (drug-like compounds, small peptides, flexible protein loops) and single/multi-ligand docking (into proteins or other hosts).

S4MPLE (Sampler For Multiple Protein-Ligand Entities) is a molecular modeling program based on a Lamarckian genetic algorithm. Its inbuilt genetic operators (mutations, cross-overs) include various heuristic modifications of structures in the current populations – shake-up by high-temperature MD, fragment recombination, random torsion angle change, random hydrophobic contact or hydrogen bond formation (implying rototranslation of loose fragments, which are implicitly considered as ligands. The "site" entity must include at least one fixed atom). It can be employed for a wide variety of simulation types: conformational sampling of ligands or small peptides, and docking of both fragment-sized and drug-sized compounds. There is no limit with respect to the number of considered entities – simultaneous docking of multiple ligands is supported. The energy function uses AMBER[43] and GAFF[44] to respectively simulate peptide and small organic moieties of the considered system. Here, all simulations are performed with the "Fit FF" energy scheme described, calibrated and validated previously[17].

Machine-learning chemical context-specific (but geometry-invariant) EBR values associated to SMARTS-based definitions of atom types in various chemical environments[45] resulted in successful prediction vacuum-to-water transfer energies of small rigid molecules. Such an approach cannot be used for docking – when ligand atoms witness dramatic EBR increases while buried into the active site – and even less for macromolecular simulations. Yet, it has the merit to point out that part of the degree of burial of each atom, stemming from its direct topological neighbors, is basically geometry-independent (no significant impact by bond length or valence angle fluctuations).

The "morphometric approach" is another empirical strategy to calculate hydration free energies expressed as linear combinations of specific geometric terms. Recently[46], a protein-specific empirical linear combination of geometric indices and "standard" GB term was fitted to reproduce 3D-RISM hydration energy values. Geometric indices are employed to compensate for the systematic errors of the standard GB formula for proteins. Unfortunately, the model neither alleviates the cost of GB term estimation, nor applies to species other than proteins.

The key idea of this work was to use the fast, fitted linear function of atom-specific geometric and topological indices for accurate but quick *prediction* of atomic *EBR* and accessible surface (SA) fractions *F*. We captured the dependence of EBRs and Fs on atomic type and geometry of its neighborhood. A set of atom-centered topological indices is aimed at describing the connectivity-induced degree of steric hindrance, in completion to geometric descriptors (rendering the actual occupancy of the surrounding space by atoms not bonded to the central one). Multilinear regression has then been used to weigh the importance of the connectivity-based *versus* geometry-based indices in order to reproduce the standard EBR – as obtained by numeric integration of equation (1) – and F values, respectively.

The ability to reproduce standard EBRs is a necessary and, as will be shown, successfully reached milestone, but not the final one. Further efforts to fine-tune the GB/SA model directly with respect to experimental data[47] and, if needed, with respect to more accurate PB-based EBRs, will be undertaken in perspective. The ultimate success criterion is ability to reproduce experimental observations – focusing the fitting process specifically on reproducing hydration energies is not sufficient: the empirical GB/SA term must be rendered compatible[48] with the underlying force field engine of the simulations.

2. Methods.

The approach advocated in this work is graphically illustrated in the Figure 1 below:

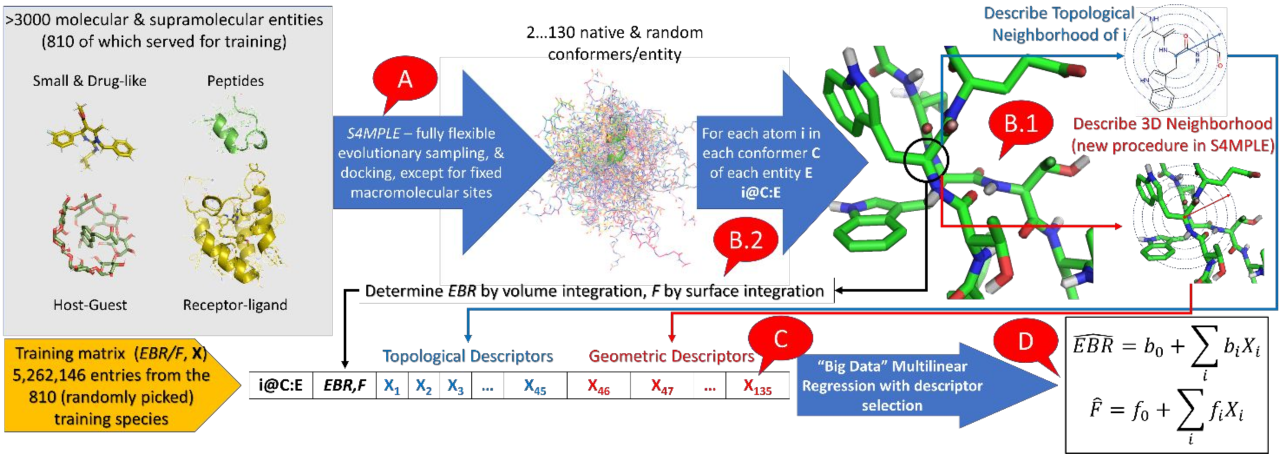

>3000 molecular & supramolecular entities (810 of which served for training)
Small & Drug-like
Peptides
Host-Guest
Receptor-ligand
A
S4MPLE – fully flexible evolutionary sampling, & docking, except for fixed macromolecular sites
2...130 native & random conformers/entity
For each atom i in each conformer C of each entity E i@C:E
B.2
B.1
Describe Topological Neighborhood of i
Describe 3D Neighborhood (new procedure in S4MPLE)
Determine EBR by volume integration, F by surface integration
Training matrix (EBR/F, X) 5,262,146 entries from the 810 (randomly picked) training species
Topological Descriptors
Geometric Descriptors
C
i@C:E
EBR,F
X_1
X_2
X_3
...
X_{45}
X_{46}
X_{47}
...
X_{135}
"Big Data" Multilinear Regression with descriptor selection
D
$\widehat{EBR} = b_0 + \sum_i b_i X_i$
$\hat{F} = f_0 + \sum_i f_i X_i$

Figure 1: Flowchart of the fitting procedure of the empirical models for atomic EBR and F value estimators as linear functions of topological and geometric descriptors. (A) Random conformers were generated with S4MPLE, and native geometries added to the conformer pools of each compound. (B) For each compound (entity *E*), each of its conformers *C* is iteratively visited. For every atom, its descriptors are computed (B.1) and then reference EBR and F values are obtained by numeric integration (B.2). This results (C) in the obtention of the explained variable-descriptor matrices, leading to final models (D).

2.1 Used Compounds and their Preparation

Structures considered here span the largest possible structural diversity – from fragment-like neutral and charged compounds with measurable solvation free energies (a literature set[49]), to small proteins typically employed in folding simulations (PDB[50] codes 1L2Y, 1LE1, 1CHL, 1VII), to drug-like compounds reported in ChEMBL[51] and their docked poses into the sites of proteins they were experimentally assessed against, to host-guest systems (a series of ligands binding to β-cyclodextrin). On the overall – with some organic compounds being represented both as stand-alone species and docked into (rigid) protein sites – 2310 different molecular and supramolecular entities were considered.

Organic molecules (including the β-cyclodextrin host) were standardized according to the rules of our ChemAxon-powered[52] web server http://infochim.u-strasbg.fr/webserv/VSEngine.html and prepared for docking or stand-alone sampling according to the standard S4MPLE protocol[53].

Note that the sampling of the free ligand is part of the S4MPLE docking process. For fragment-like and drug-like molecules, one (for small and rigid compounds) to 135 (for the most flexible compound) S4MPLE-generated conformers were employed. Above-mentioned polypeptides and small proteins were represented by their native geometries, plus a set of ~100 random decoy geometries produced by S4MPLE. Note that, as S4MPLE systematically energy-minimizes structures obtained by evolutionary operators, these decoy geometries are clash-free and respect covalent constraints. Obviously faulty geometries were not included: it is irrelevant whether their solvation energies can be accurately reproduced or not.

For protein-ligand docking, poses generated in previous work[53] for randomly selected ligands of the following targets were exploited here: CHEMBL1827 (Phosphodiesterase V, with 240 randomly picked ligands), CHEMBL1865 (Histone deacetylase, 136 ligands), CHEMBL203 (Epidermal growth factor receptor, 914 ligands), CHEMBL204 (Thrombin, with 632 ligands) and CHEMBL227 (Angiotensin receptor II, with 204 ligands). Note that the ligand subsets exploited here are all taken from the ChEMBL ligand series used in the cited work, with the exception of CHEMBL204, Thrombin, for which the 632 random picks were taken from the DUD binder/decoy set. Preparation of protein active sites was thus described previously[53]. At this point it is important to emphasize that S4MPLE is designed to work with a predefined cutoff of 12 Å and therefore entire residues of the (rigid) protein that were clearly too far to interact with the ligand placed in contact with active site "hot spots[17]" were cut out – completely removed – of the protein site .mol2 files. The removed parts would have significantly impacted on EBRs of protein site and ligand atoms, as discussed in Introduction. Yet, the present goal is not (yet) to reproduce physically accurate EBR values, but to prove that our empirical approach can successfully mimic the volume integral-based values, equally affected by the "cuts". Herein employed geometries for protein-ligand complexes were taken from the diverse pool of less than 200 stable poses selected by S4MPLE from the entire pool of poses, in terms of contact fingerprint diversity. The study only considered the atoms of the ligands plus the few protein "hot spot" atoms for EBR/F monitoring, since in the rigid sites the numerous "frozen" protein atoms at large distances from the ligand would have systematically returned near-constant EBR and constant F values, generating extremely large output files with little noteworthy information.

Empty protein sites were also used, both with their PDB-imported geometries and with S4MPLE energy-minimized structures (full flexibility), and all their atoms were included in the study.

In host-guest docking calculations, the only differences to the previously mentioned S4MPLE docking protocol are that (1) the "site" β-cyclodextrin is not a protein, so it had to be explicitly assigned Gasteiger/ChemAxon partial charges, and GAFF parameters, (2) hydrogen atoms of the host were designed as flexible, in order to allow -OH groups to reorient in view of hydrogen bonding with the guest. In addition, the 88 host-guest complexes were also subjected each to 2 ns of MD simulation with S4MPLE, considering full flexibility of the host β-cyclodextrin, and 200 equally spaced frames of the trajectory were also included in the study, in addition to docking-generated conformer families.

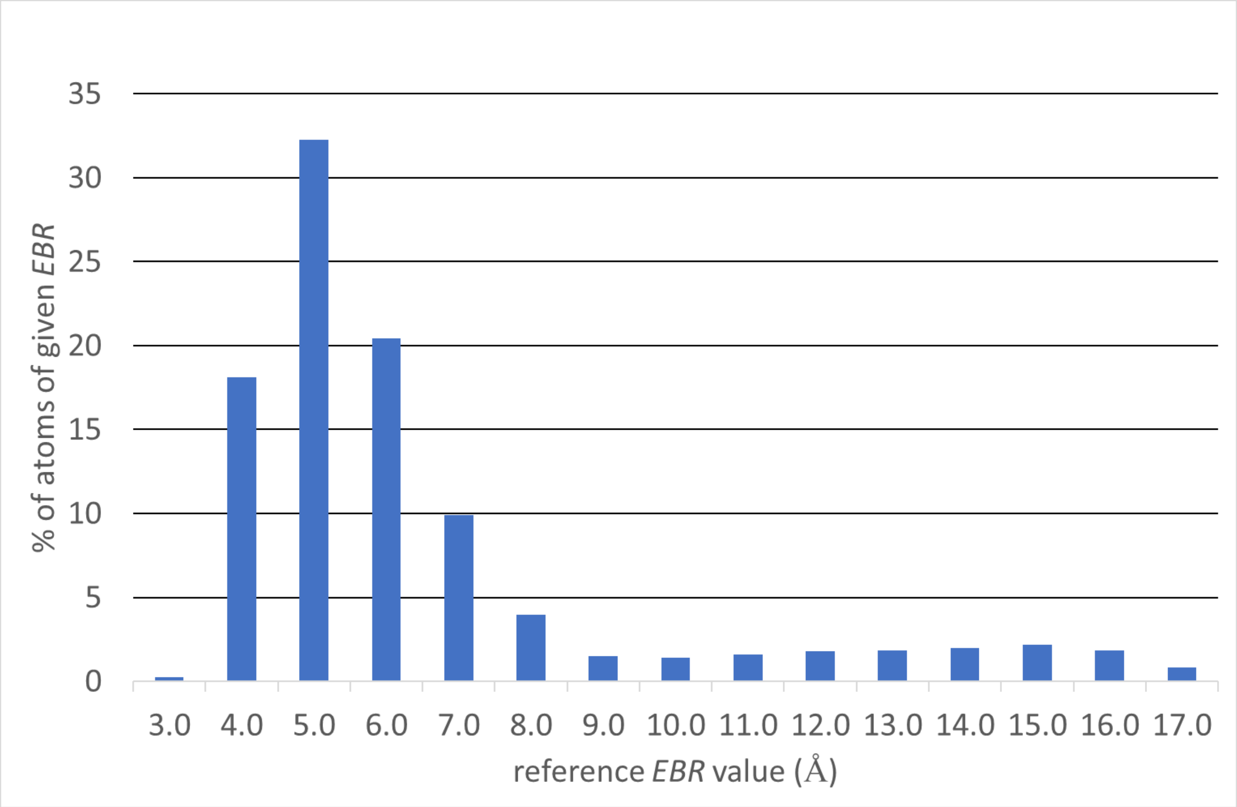


Figure 2: Distribution of reference values, calculated according to equation (1), over the entire pool of considered compounds and molecular complexes

The distribution of the reference values of effective Born radii (Figure 2) highlights the heterogeneous and diverse nature of the monitored species, with a majority of free and docked ligands (contributing EBR values in the 3-8 Å range), and fewer macromolecules, contributing high EBR values. No experimental data is associated to the molecules exploited by this work. The above bias is an expression of the ad hoc selection of compounds – mostly for historical reasons, these being systems that were already studied and sampled. Yet, the selection is large enough to support random splitting into training and test sets. Random splitting did not fail to leave out representatives of each compound class for external validation – all of small proteins, protein-ligand complexes, host-guest complexes, ligands and fragment-like molecules were present in the test sets, in proportions matching the training set. The split was random, but the outcome has been checked and would have been manually corrected to ensure the above requirement – yet, that was not necessary.

2.2 Descriptors of Atom Neighborhoods

In the following, indices designed to capture the relevant topological and geometric information about each atom of a molecular system will be introduced. The final descriptor of an atom *i* will be a vector $\boldsymbol{X_i}$ (**bold** setting for vectors) having as components the various below-defined indices. It is convenient to consider vector $\boldsymbol{X_i}$ as a concatenation of several specific vectors: $\boldsymbol{X_i} = (\boldsymbol{T_i}, \boldsymbol{PT_i}, \boldsymbol{D_i}, \boldsymbol{O_i}, \ldots)$, each standing for a series of related index values – some of topological, and some of geometric nature, *vide infra*. In order to emphasize that these are descriptors of *atoms*, not of molecules, the index *i* associated to the vector highlights the association of the vector to given atom *i* in some molecule *m.* The scalar components of a vector $\boldsymbol{X_i}$ will be denoted as X_{ik}.

2.2.1 Topological Indices

Each atom *i* is characterized by eight topological indices. The first four, T_{ik} *(k=1..4)* represent inverse power sums of its topological distances τ_{ij} to all other (connected) atoms *j* in the system,

$$T_{ik} = \sum_{j \neq i} \tau_{ij}^{-k} \ \ (k = 1..4) \qquad (2)$$

while the last four *(k=5..8)* weight the contribution of neighbors *j* by their atomic masses M_j.

$$T_{ik} = \sum_{j \neq i} M_j \tau_{ij}^{-(k-4)} \ \ (k = 5..8) \qquad (3)$$

Furthermore, every atom *i* is additionally characterized by its “one-hot-encoded” AMBER/GAFF potential type vector $\boldsymbol{PT_i}$. Its element PT_{ik} equals one if atom *i* is of potential type *k*, and zero otherwise. Potential type *k* refers to a list of the 37 most ubiquitous AMBER/GAFF potential types, found to occur at least 2500 times in the considered molecular systems. Atoms of types not among these 37 most ubiquitous ones therefore have $PT_{ik}=0$ □ $k=1..37$. This vector allows machine learning to fit systematic potential type-dependent offsets to calculated EBR_i and F_i values, if needed.

2.2.2 Geometric Descriptors

A first set of four geometric indices D_{ik} $(k=1..4)$ represent inverse power sums of actual interatomic distances, employing equation (2) with actual interatomic distances d_{ij} instead of their topological counterparts τ_{ij}. Since S4MPLE may visit non-physical conformational space zones, including physically impossible geometries with overlapping atoms, any d_{ij} values below the sum of covalent radii of concerned atoms was reset to this minimum, in order to avoid pathologically high D values causing instability of machine learning by multilinear regression. This safeguard was also applied to all other geometric descriptors, *vide infra*.

Next, a set of twelve "shell occupancy counts" O_{ir} represent the (fuzzily counted) number of atoms populating the twelve successive spherical shells of radii r=1 to 12 Å, centered on atom i. The fixed 12 Å range was chosen to match the fixed nonbonded interaction cutoff value in S4MPLE. Note that although protein sites were "trimmed" using a 12 Å cutoff, residues having at least one atom within cutoff distance to interaction hot spots were fully kept – therefore, atoms further away than 12 Å may be numerous in ligand-protein complexes. All are generically counted within the last shell $r=12$, irrespective of their actual distance.

For each atom $j \neq i$, if the interatomic distance d_{ij} happens to be an integer r, then atom j will be fully counted as resident of sphere r (and any atoms with d_{ij}>12 are counted in shell #12). Otherwise, atom j will be shared between shells $r=[d_{ij}]$ ([] stands for the truncation – floor – operator) and $r+1$, proportionally to the actual d_{ij} value. Considering $\delta_{ij}=d_{ij}-[d_{ij}]$, $O_{i(r+1)}$ will be incremented by δ_{ij} and O_{ir} increased by $(1- \delta_{ij})$.

Eventually, a more complete set of spherical shell descriptors employs a different fuzzy-logics counting scheme: an atom j is associated to shell r centered on atom i with a weight of:

$$w_{ij}^{r} = \frac{r^2}{r^2 + \left(d_{ij}^2 - r^2\right)^2} \qquad (4)$$

empirically chosen to peak at 1.0 when $r=d_{ij}$ and to conveniently decrease as r and d_{ij} diverge. Unlike in above-defined descriptors of O type, all atoms j (marginally) contribute to all shells around i. In this formalism, occupancy levels of the spherical shells are taken as – both plain and (square of) atom radius-weighed – sums of w values, respectively:

$$S_{ir} = \sum_{j \neq i} w_{ij}^r \, ; SR_{ir} = \sum_{j \neq i} w_{ij}^r R_j^2 \qquad (5)$$

Above, atomic radii R_j are equal to van der Waals radii plus a user-specified offset parameter R_w, typically corresponding to the “radius of a water molecule” used to offset the dielectric boundary away from the van der Waals surface. On one hand, EBR and F values depend on the choice of R_w, while on the other R_w will implicitly impact several components of the X vector, such as the ***SR*** terms of equation (5), and others introduced below. Furthermore, both R_w as such and the solvation radius R_i of the central atom were also included in vector $\boldsymbol{X}$, as explicit components. Therefore, as $\boldsymbol{X}$ explicitly contains the user-chosen R_w value and R_w-dependent terms, it is expected that the fitted coefficients of the machine-learned relationship *EBR=f(**X**)* will be independent of R_w. Input of an $\boldsymbol{X}$ vector generated at given R_w into the machine-learned model should directly return the EBR and F values expected at that particular R_w value.

The latest types of contributions to the atomic descriptor vector $\boldsymbol{X}$ monitor whether the atoms *j* surrounding atom *i* within a given spherical shell *r* are homogeneously distributed across the spherical shell, or rather clustered together. In the latter case, they’d form a compact bulk occupying a delimited area on the spherical shell, leaving the rest of it solvent-accessible. This degree of homogeneity can be estimated by considering the unit vectors $\boldsymbol{u}_{ij} = (\boldsymbol{x}_j - \boldsymbol{x}_i)/d_{ij}$ defining the relative position of atom *j* with respect to “center” atom *i*, with $\boldsymbol{x}$ denoting the absolute position vectors of the atoms, and d_{ij} the interatomic distance. As the degree of association of atom *j* to a spherical shell *r* is w_{ij}^r, the average of unit vectors over a spherical shell *r* is $\sum_{j \neq i} w_{ij}^r \boldsymbol{u}_{ij} / \sum_{j \neq i} w_{ij}^r$. The ||norm|| of this vector, as given in equation (6), may approach 1.0 (if all atoms *j* relevant to that shell are close, overlapping neighbors of each other – leaving most of the shell accessible to solvent), or drop to 0.0 if atoms *j* are homogeneously spread over the spherical shell. Thus, it is a useful metric of the degree of homogeneity. These norms, as well as the equivalent expressions enacting the above-introduced atom-radius-based weighing scheme, compose two specific vectors ***U*** and ***UR*** contributing to ***X****:*

$$U_{ir} = \left\| \sum_{j \neq i} w_{ij}^r \boldsymbol{u}_{ij} / \sum_{j \neq i} w_{ij}^r \right\| ; UR_{ir} = \left\| \sum_{j \neq i} w_{ij}^r R_j^2 \boldsymbol{u}_{ij} / \sum_{j \neq i} w_{ij}^r R_j^2 \right\| \qquad (6)$$

Therefore, knowing that ***S*** and ***SR*** are fuzzy counts of solvent-displacing atoms in sphere shell *r*, while ***U*** and respectively ***UR*** describe how effectively they cover the entire shell, it was postulated that terms like ***B*** and ***BR*** below may be useful descriptors of the effective degree of solvent-displacement at shell *r*:

$$B_{ir} = S_{ir}(1 - U_{ir})\ ;\ BR_{ir} = SR_{ir}(1 - UR_{ir}) \qquad (7)$$

All of ***S, SR, U, UR, B*** and ***BR*** were added to vector ***X***, leaving it up to the descriptor selection protocol of the machine learning procedure to pick the most useful. Herewith, vector ***X*** totalizes 135 components, characterizing atom *i* in terms of its successive neighborhoods.

2.3 Generation of Training and Test Data Matrices

For each compound in the considered set, atomic radii were assigned to standard van der Waals radii R^{vdW}_i plus the R_w offset. Four R_w options were considered: 0.9, 1.1, 1.4 (the typical water molecule radius value) and 1.6 Å. At each R_w choice, the solute-solvent interface is defined, for each of the sampled geometries of that molecule, by the "solvent-accessible" surface at (R^{vdW}_i + R_w). For every atom *i* in this conformation, the reference EBR_i is obtained by volume integration according to equation (1), noting that *no* distance cutoff is applied during the volume integration. Integration proceeds by generating a cubic grid with a spacing of 0.2 Å around the molecule and selecting the grid points that fall inside the solute's dielectric boundary. The reference fraction of the solvent-accessible sphere, F_i, is calculated by placing a predefined "mask" of 500 points homogeneously covering the atom sphere and counting those that are not buried by overlapping spheres. The procedures for reference EBR and F calculations are implemented in S4MPLE.

The descriptor vector of every atom *i* in the current conformation is calculated as above described. Thus, *(EBR, **X**)* and respectively *(F, **X**)* explained variable-descriptor pairs have been generated for all atoms in all conformations of all molecules, at all considered R_w values. These data were then randomly dispatched into training and test pools. Entries pertaining to all atoms of a same molecule at given R_w were always dispatched together – either all in training, or all in test. Under no circumstances were some geometries of one molecule used for training, with remaining kept for testing. However, entries at one R_w value may be assigned to training, while the corresponding entries (of the same atoms, over the same pool of conformers) at different R_w are kept out for testing (in order to verify how well the model manages to capture the implicit impact of R_w).

Out of the 2310 considered species, 810 randomly picked contributed to the training set, in association to one or several of the three considered R_w values. Therefore, the 810 compounds are represented by 1031 distinct conformer pools. Each pool has a variable number of sampled geometries and features a variable number of atoms. Per total, training sets consisted of 5,262,146 *(EBR,* $\boldsymbol{X}$*)* and respectively *(F,* $\boldsymbol{X}$*)* explained variable-descriptor pairs. Remaining 28,222,701 explained variable-descriptor pairs were used for testing. 1495 species of the 2258 present in the test pool were strictly absent from the training set. They form the "external" test set, further on referred as "testX". The other were represented in training but at different R_w values, but the current (compound, R_w combination) that was not used for training. These serve to assess how robust prediction would be if the user would choose R_w values other than used for training. They form the "testw" set.

2.4 Machine Learning Procedure.

An evolutionary algorithm was used to support descriptor selection, by encoding the status (used/ignored) of each of the 135 descriptor terms in a 135-bit "chromosome". The evolutionary algorithm is an adaptation of the script-driven, asynchronous procedure used for generic (hyper)parameter selection for chemoinformatics model construction[54]. Terms (columns) with a status of "ignored" are filtered out from the master file of explained variable-descriptors, prior to input into the regression tool. Even so, the filtered file may totalize several GB of data and therefore cannot be read and stored in memory by the regression tool Fortunately, RAM storage of the input data is not required in order to determine the regression coefficients. In the present "big data" scenario, cross-validation would be too memory-consuming. Therefore, model quality is assessed by a herein designed stochastic external validation protocol. The input file is read line by line, and at each new line a random number is drawn. If its value is larger than an empirical threshold of 0.3, then the input line is used for regression matrix update (coefficient fitting) and otherwise it is uploaded into memory. The peculiar cutoff was chosen such that out of the 5.2M lines of the (Y,$\boldsymbol{X}$) matrix, the fraction loaded into memory is low enough to avoid overflows on the local 16-core workstation. When input is completed, the program proceeds with coefficient calculation, then applies the latter to estimate $\hat{Y}$ for the "external" items in memory.

The resulting R^2 value thus represents the external prediction proficiency of the model. For each chromosome, randomized regression is run five times and the "pessimistic" (worst) R^2 value over the five randomization attempts is conservatively used as a "fitness" criterion for the current chromosome (descriptor selection scheme). Note that *fitted* R^2 values are never estimated, knowing that the use of ~3.5M data items to fit 135 coefficients guarantees an extreme robustness of the regression model (typically 10…20N data items are required[55] in order to robustly fit N coefficients and avoid overfitting artefacts – here, this ratio is not 20, but 25,000). After completion of 5000 evolutionary generations, a consensus model (one for EBR and one for F, respectively) was drawn from so-far best discovered descriptor selection schemes by plain averaging of their linear coefficients.

2.5 Computational Cost Assessment

For each considered system, the CPU user time required for a complete evaluation of the energy function (including the calculation of internal coordinates from atom positions) has been reported, using the *times.tms_utime* record returned by the FreePascal *FPTimes* function of the BaseUnix unit, for (a) the default S4MPLE energy function featuring the simple desolvation pairwise desolvation term *versus* (b) the GB/SA-based energy function based on (b1) the approximate EBR and F values and (b2) reference, numerically integrated EBR and F values, respectively. In clear, (b1) implies reevaluation of all geometric atom descriptors and their linear combination resulting in estimated EBR and F values, whereas (b2) proceeds with full-blow numerical integrations in order to return reference values. Since *times_tms.utime* returns user time in 1/100 s, effective times per energy function calls were estimated by monitoring elapsed time over repeated calls of the energy function, the number of repeats being inversely proportional to the number of nonbonded pairs in the system. Very fast calls in small entities are thus repeated enough times in order to get a robust assessment of time/call. The execution time of the default S4MPLE energy function was taken as baseline, and the slow-down penalty for using (b1) approximate GB/SA and (b2) reference GB/SA was reported as the ratio of respective times.

2.6 Result Evaluation Protocols.

First, it is important to verify how well the estimated $\widehat{EBR}$ and, respectively, $\hat{F}$ match their reference values, for any given atom, over the pool of sampled geometries at given R_w. This is best expressed in terms of *RMSE* values, in Ǻ for EBRs, and dimensionless numbers for the accessible fraction. Determination coefficients $R^2 = 1 - RMSE^2/\sigma^2$ can be calculated, albeit a low R^2 value might be either an expression of model inaccuracy (high *RMSE*) or low intrinsic standard deviation σ of the parameter as a function of the considered sampled geometries at acceptable *RMSE*. The performance over the pool of sampled geometries can thus be tracked for hundreds of thousands of distinct atoms available here. Thus, results are best shown by monitoring the percentage of tracked atoms for which the monitored criterion falls within a given range, *i.e.* construct cumulative density histograms (% of atoms reporting $RMSE \leq x$ or $R^2 \geq x$) within relevant x value ranges.

These histograms will also be generated for various atom subsets, in order to evidence specific trends of quality criteria. Specific *RMSE* distributions over the "testw" atom set will evidence whether the model is able to extrapolate at different R_w values. More challenging, distributions over the "testX" atoms of compounds (not participating at all in training) will evidence the capability of the model to cope with completely novel chemical entities.

Another meaningful classification scheme follows the chemical nature of the molecule systems: small ligands, small foldable proteins, host-guest systems, protein-ligand complexes were specifically scrutinized. Statistics over MD-sampled trajectories of the host-guest systems (*MD* subset) may explain whether the conformational sampling mode is affecting the model quality. Last but not least, specific profiles of atoms pertaining to a given AMBER/GAFF potential type (from arbitrary molecules) may show whether some atom types are systematically predicted better or worse than average.

The comparison can also be taken over the pool of all the atoms in all the conformers of the species – in this case, the R^2 value will be a measure of the propensity of the model to correctly discriminate between "buried" and "accessible" atoms (according to either of the two criteria, EBR or F). This is the strategy used at the model selection step.

Eventually, the fundamental question is whether the solvation energy returned by the GB/SA formula using $\widehat{EBR}$ and $\hat{F}$ estimators is an accurate estimator of the value that would correspond to reference *EBR* and *F*. To this purpose, the total GB/SA energy was taken as the sum of the GB term plus a surface-dependent cavitation/hydrophobicity contribution[1] of 7.2 cal/Å^2 × total accessible surface. The estimator $\widehat{GBSA}$ (based on $\widehat{EBR}$ and $\hat{F}$) was compared to its reference *GBSA* energy on the basis of fixed-slope, free-intercept regression lines $GBSA = 1.0 \times \widehat{GBSA} + C$ over the pools of sampled geometries of each molecular system. Here, *C* is fitted in order to minimize *RMSE.* Indeed, estimated and reference solvation terms may diverge by any arbitrary large constant offset for each species. The only important criterion is to ensure that the solvation energy differences between any two conformers is roughly the same: $\Delta GBSA \approx \Delta\widehat{GBSA}$. $RMSE$ values in kcal/mol and, when pertinent, R^2 values of unit-slope, free-intercept regression lines are thus the best suited indicators here.

3. Results and Discussion

Before going into detailed analysis of the results, the following brief discussion is meant to put this proof-of-concept work into perspective, highlighting the rationale for which this effort was undertaken, its potential benefits and outlining the future steps towards a consistent solvation energy prediction model.

Our ultimate goal is to obtain a fast, QSPR-based GB/SA approximation within the molecular mechanics Hamiltonian powering the S4MPLE tool. This GB/SA approximation shall return solvation energies as a function of conformation. Calibrating these against experimental data is now in progress, and not a concern of the current paper. Note that this is *not* an attempt to establish a direct QSPR model for solvation energies, *i.e.* express solvation energies as a function of molecular descriptors. The latter would correlate the average solvation energies of conformers to averages of structural features present in the molecule – thus fail to capture how this solvation energy depends on geometry.

The targeted GB/SA term is designed to be part of the “fitness score” (energy level) governing the evolutionary process in S4MPLE. It may, as such, be gradient-minimized, since it is a function of geometric descriptors, themselves differentiable functions of atomic coordinates. However, further work will be needed in order to define the best strategy for differentiation of EBR and F values. Note, furthermore, that it will be used to energy-minimize and rank conformations but will not serve to guide evolutionary operators. Simplified potential functions are probably sufficient to attract sampling towards clash-free conformational space zones with putatively favorable contacts. The full GB/SA-driven energy minimization would only be applied to already rather stable geometries, in order to fine-tune the balance of hydrogen bonds, hydrophobic contacts, etc. Or, in S4MPLE Molecular Dynamics is simply a “mutation” operator used to force a conformer into a nearby minimum by simulated annealing – this may as well be achieved with the basic desolvation model already implemented in S4MPLE. GB/SA will only be used to accurately detect and score the new energy well. This interplay of GB/SA energy minimization and scoring *versus* simpler Hamiltonians piloting the stochastic conformation space jumps is also an important further research topic.

The current paper focuses on a precise question at the very beginning of our quest: is it possible to approximate EBR and F values as a linear combination of simple geometric and topological neighborhood? The answer is positive, because it was shown that GB/SA solvation energies calculated with approximate EBR and F terms are excellent approximations of reference values obtained from volume and surface-integral based approaches. This is the key non-trivial result presented here, knowing that this is only a first – but important – step in order to achieve the above-mentioned ultimate goal. There are four potential key advantages of the herein advocated machine-learned approach over previously attempted GB/SA approximations:

- First, the approach is free to choose its geometric descriptors, in an unbiased way. By contrast, state-of-the-art analytical approximations rely on empirically designed functional forms (with, presumably, a lot of human trial-and-error work that was never explicitly published but cannot challenge the volume of descriptor selection hypotheses tested by machine learning).
- Second, the approach introduces a series of coefficients – the descriptor weights – which could be novel "force field parameters" of the solvation term. Their initial values may serve as a starting point in search of optimal setups that maximize agreement with experiment. The ensemble values of the GB/SA term over all representative geometries populated at given temperature *T* should match experimental vacuum-to-water energies of molecules, which provides a first objective for fine-tuning of the herein fitted empirical coefficients. Nor will experimental solvation energies be the only criterion used for GB/SA fine tuning: a multiobjective optimization strategy will be set up, including (1) solvation energy predictions, (2) docking benchmarks (ability to rank potent ligands better than inactives by GB/SA-enhanced AMBER/GAFF energy), (3) peptide folding benchmarks (propensity to rank native-like structures by GB/SA-enhanced AMBER/GAFF energy, *etc*). Such fine tuning will be restricted to relatively few examples of experimentally studies systems – which alone could not have supported the herein described global fitting that led to this "default" setup of the EBR model. In chemoinformatics, this strategy is called "inductive transfer of knowledge[56, 57]": roughly estimating parameters on the basis of related, less accurate or less appropriate but wealthy data, followed by specific fine tuning of some of the

latter against few, but specific experimental end points. Future work will show whether this further fine tuning is the correct strategy to follow, or whether this proof-of-concept will first have to be rerun against PB-modeled "perfect" EBR values, as suggested in some literature studies.

- The third key advantage of the approach over previously published analytical GB approximations is applicability to a wide panel of chemical systems. By contrast, the former were initially dedicated either to small molecules or to proteins and had to be recalibrated in order to compensate for the unavoidable molecule size artefacts. Even so, there is yet no compelling proof that they would also apply to docking problems, where a small ligand approaches a large protein. This approach was from start trained on small proteins, protein-ligand complexes, host-guest complexes, ligands and fragment-like molecules. It was challenged to predict external small proteins, protein-ligand complexes, host-guest complexes, ligands and fragment-like molecules – thus, its very large realm of applicability is established.
- Fourth but not least, note that two rather complex approximation problems – volume integral-based EBR and surface integral-based F values – were reduced to a common framework, using the same set of geometric descriptors – the core effort for both models, descriptor calculation, is shared. Alternatively, "analytical" surfaces and "analytical" Born radii would have to be calculated separately, summing up the effort.

3.1 Model fitting results

The evolutionary model fitting procedure was able to quickly generate many high-quality models for both EBR and F properties. For EBR, 25 top models had "pessimistic" R^2>0.985, whilst F models reached R^2>0.938 (over the 30% randomly left-out training data, see §2.4). The very high scores are foremost an indicator of the high proficiency to discriminate "typically buried" from "typically accessible" atoms, according to their chemical context. Thus, the near-perfect R^2 values – albeit they correspond to a genuine external prediction, *not* to fit values – are necessary, but not sufficient.

Fitting a model against the reciprocal of EBR values ($Y=EBR^{-1}$) might have seemed preferable, since GB energies relate to EBR^{-1}. This had been originally attempted with, however, slightly lower statistical quality criteria. Such a model would be very imprecise in terms of absolute EBR values for strongly buried atoms – or, robust EBR values may find some utility *per se*, as chemoinformatics descriptors of atoms. For these both reasons, the *Y=EBR* scenario was preferred here.

Interestingly, individual top models invariably tend to include a large majority of the available descriptors (typically about 80% of the 135 provided terms are co-opted), which eventually caused all of the terms to enter the EBR consensus model, and nearly all to be present in the F model. If a family of descriptors would have been useless or redundant, bypassing its calculation would have resulted in time gain. On the other hand, these terms were *designed*, to the best of our knowledge, to be relevant, *i.e.* to meaningfully characterize the degree of burial of atoms. As the (training set size)/(explaining variables) ratio is of the order of 25×10^3, there is definitely enough data at hand to support inclusion of all these terms.

This work is merely a proof-of-concept showing that chemoinformatics-inspired predictors of the key GB/SA parameters can be successfully fitted to mimic the "reference" volume integral-based approach. Since, however, the latter is already an empirical, parameterization-prone approach (atomic radii, surface coefficients σ and R_w being typical[58] degrees of freedom) the eventually useful coefficients will need retuning in conjunction with the above-mentioned parameters, in a large-scale challenge to reproduce experimental data. The fitted coefficients are nevertheless given for illustrative purposes, as Supplementary Information (file "ebr+fbur.model", formatted as required for input by S4MPLE). Column 2 of this file reports the concerned descriptor, column 3 representing its weight in the EBR model, while column 4 marks its contribution to the fraction of *buried* area (a positive term in column #4 means that increasing that descriptor value tends to decrease the accessible fraction). In descriptor names, the numeric index refers to the *k* values in defining equations. "RSOLV" represents the solvation radius, "INT" is the free intercept of the equations, "RH2O" the assumed radius of the water molecule, and all the terms starting at entry 105, labeled by AMBER/GAFF potential types represent the potential type bits *PT*.

3.2 How accurately are atomic EBR and F values estimated by the model?

For each molecule, at each considered R_w value, for each atom *i,* the series of its reference EBR values at each sample geometry was compared to the series of the estimates returned by the consensus model, and the *RMSE* and determination coefficient were reported (rigid species for which no more than 2 distinct geometries could be sampled, and *apo* protein sites, represented by the PDB geometry and the energy-minimized structures only, were not part of this analysis). Each such atom pertains either to the training set, or to one of the test sets (testw, testX) as above-defined. Better performance of the model with respect to a subset of atoms amounts to a faster increase of the cumulated density plot.

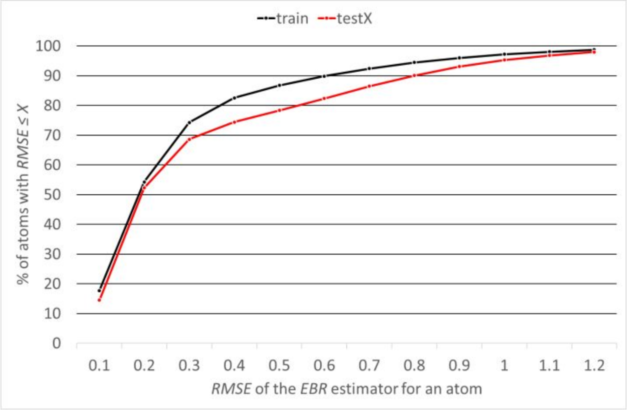


Figure 3: Cumulative density plots of *RMSE* of *EBR* values committed for each atom within specified subsets, over the pool of sampled geometries, at given R_w values.

According to Figure 3, in virtually all cases (>98%), the empirical model approximates volume-integral-based EBR values with *RMSE* errors below 1.2 Ǻ. Expectedly, atoms included in training set tend to be slightly better predicted than atoms in “unknown” species. In order to check whether these *RMSE* values are small when compared to the standard deviations σ of the EBR parameter in response to the change of geometry, cumulative plots of the distribution of determination coefficients $R^2 = 1 - RMSE^2/\sigma^2$ were represented in Figure 4 below.

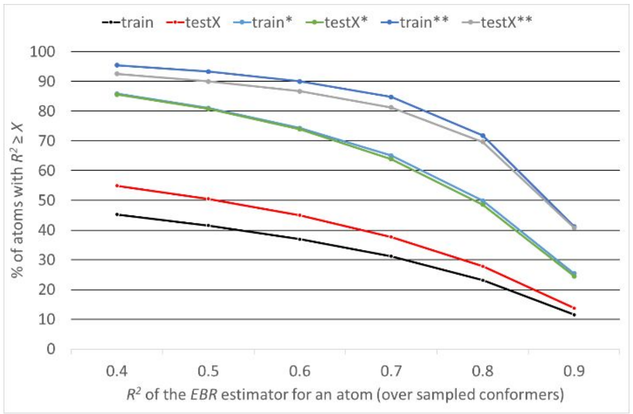


Figure 4: Cumulative density plots of the determination coefficients for atomic EBR estimators, over all atoms in train and testX and, respectively, on subsets including only those atoms with significant variance of the reference EBR values (*: $\sigma > 0.2$; **: $\sigma > 0.5$ Å, respectively)

According to the black curve in Figure 4, only in 45% of cases some reasonable correlation (R^2 >0.4) of estimated *versus* reference EBR values can be found (for training set atoms – intriguingly, the external test set seems to behave marginally better). However, if the analysis is restricted to only the atoms which actually witness some significant variation of their reference EBRs over the sampled geometries (subsets * and **, respectively), then moderate correlation can be shown to occur with very high likelihood, and near-perfect correlation can be observed for a non-negligible fraction of cases, training and external test atoms alike. In other words, the empirical EBR estimator is competently mimicking the variation of reference EBRs – *if* a significant variation of this EBR can be evidenced. In many cases (~50% of all atoms witness a standard deviation of less than 0.2 Å over their pool of sampled geometries, and 76% are below 0.5 Å), the hypothesis of a constant EBR value would have been applicable.

The performance of the estimator of the solvent-accessible fraction F of atomic surface can be directly interpreted in terms of *RMSE*, since F is bound to the [0,1) range (practically, its maximal value is of about 0.8, for terminal atoms). Approximating F with a precision of ±0.05 is an intrinsically good performance, and it concerns ~96% of the atoms, as can be seen from Figure 5. There are no noticeable differences between training and external test atoms in terms of F performance.

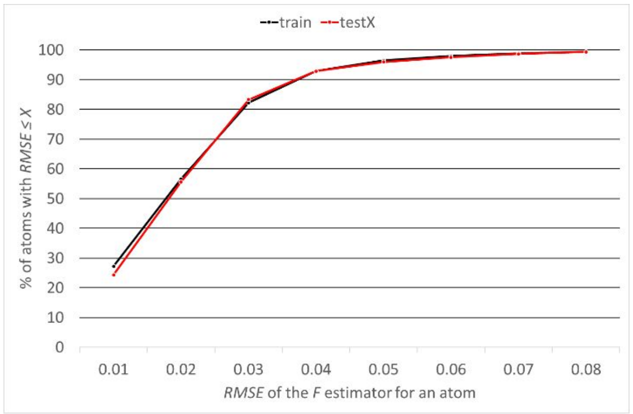


Figure 5: Cumulative density plots of RMSE of F values committed for each atom within specified subsets, over the pool of sampled geometries, at given R_w values.

3.2.1 Compound-specific behavior: not all EBR/F estimation problems are equally difficult.

Protein-ligand docking simulations by S4MPLE causes significant variations of EBR values, and implicitly larger *RMSE* values of the EBR estimator (compare docking-specific Figure 6 below to the global *RMSE* distribution in Figure 3).

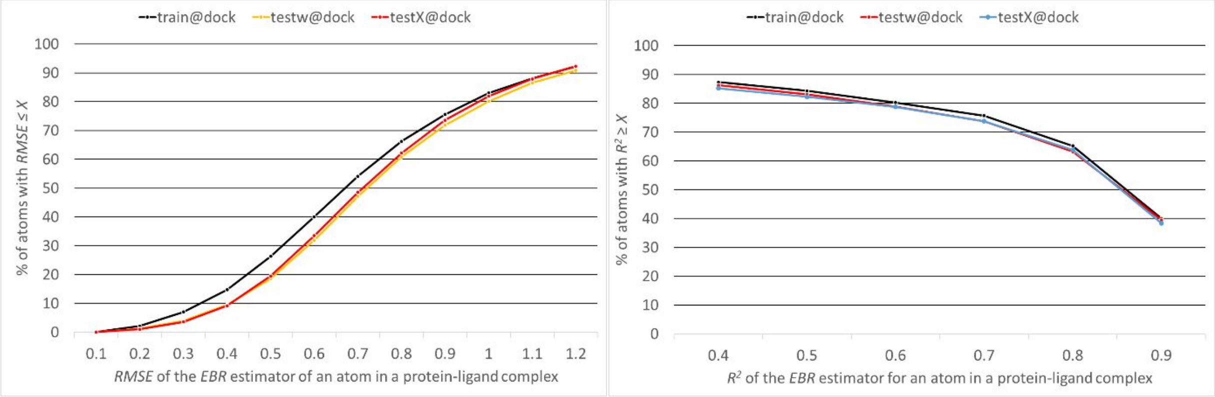


Figure 6: Specific RMSE (left) and R2 (right) cumulative density plots for EBR, over atoms within protein-ligand complexes, within docked poses by S4MPLE in rigid active sites.

However, higher *RMSE* notwithstanding, the EBR model reproduces observed variations rather well, as shown by the right-hand R^2 plot (note – by contrast to Figure 4, reference standard deviations are significant in docking problems, there is no need to filter by σ in order to discard cases where R^2 estimation makes little sense). For the docking problem also, performance over external protein-ligand complexes (within testX) is comparable to the one achieved upon challenging the model to extrapolate at a different R_w value (testw), and altogether very close to the performance over training items. The statistical robustness of the model is herewith established. In terms of surface-accessible fractions, docking problems are however not "special", the docking-specific *RMSE* density plots are undistinguishable from the global ones shown in Figure 5. This is expected, since the depth of the desolvating layer of protein atoms may significantly modulate EBRs, but has no impact on surface burial, caused by closest contacts. By different weighing of the same descriptors, both EBR and F can be approximated – an untrivial result.

Host-guest docking into β-cyclodextrin is significantly less impacted by desolvation than protein-ligand docking. These systems behave more like small drug-like molecules, with limited geometry-dependent variance of EBR values, and thus high accuracies of estimated values – in particular for docking simulations into the rigid, symmetric β-cyclodextrin ring. The Molecular Dynamics simulations however treat the host as a flexible entity – hence (Figure 7), EBR value variations (and implicitly R^2 values, not shown) are higher over MD trajectories than over docking poses.

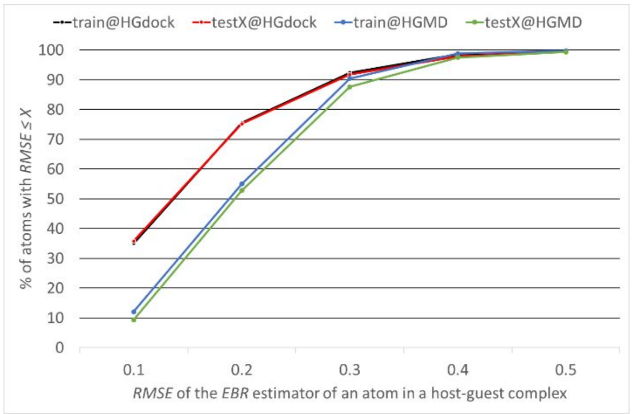


Figure 7: Cumulative density plots of *RMSE* of *EBR* values for host-guest systems, specifically reported for docking poses (HGdock) and MD-sampled (HGMD) geometries.

In terms of the small proteins selected amongst typical subjects of folding simulations (Figure 8), consequent variations of EBR values are also customary – albeit less important than in protein-ligand docking. Like in ligand-protein docking (Figure 6), the model is unlikely to return near-perfect estimations at <0.2 Å. By contrast, its uncertainly will never exceed 1.0 Å. A significant degree of correlation between estimated and reference EBR values is observed for all atoms, in all proteins, albeit weaker than in protein-ligand docking.

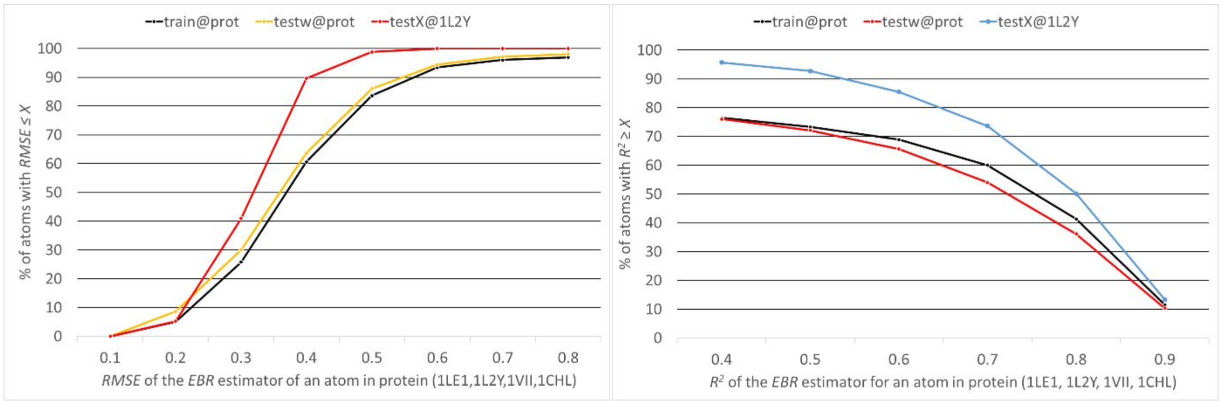


Figure 8: Specific *RMSE* (left) and R^2 (right) cumulative density plots over atoms within small proteins.

In this relatively data-sparse realm of conformational sampling problems, the random pick of training/test items resulted in a single protein being kept within the external validation challenge: the tryptophan cage, 1L2Y. The model works particularly well for this specific molecule – hence the unexpected better performance of the "external test" over training items.

Last bit not least, monitoring of the estimation propensity can also be focused on atoms of specified potential types. Some examples of cumulated density curves for various AMBER and GAFF potential types are shown in Figure 9 (here, atoms are taken irrespectively of their train, testw or testX status).

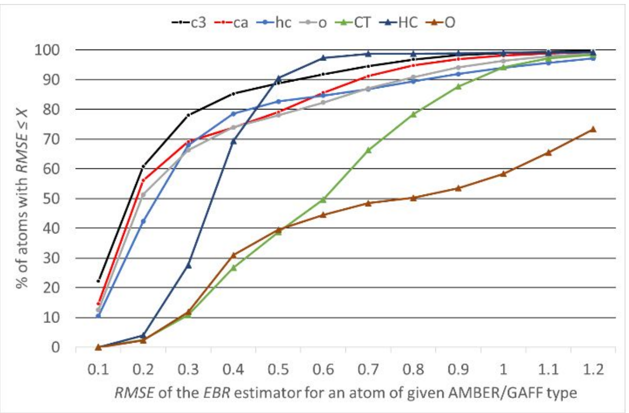


Figure 9: Cumulative density plots of *RMSE* of *EBR* values for atoms of specific potential types.

Ligand atoms are, as expected, consistently well predicted – as they are also present in the "easier" cases of stand-alone ligands. Protein site atoms, however, score the largest errors – but also experience the largest variations of their *EBR*, as already discussed. Terminal atoms – hydrogens and carbonyl oxygens – tend to be more error-prone than atoms with multiple substituents, for which dielectric screening partly comes from the immediate, relative immobile, neighborhood. This is an interesting observation, suggesting, as a perspective, to try fitting of "local" models for specific atom types, finding atom type-specific sets of descriptor weights. This might increase the precision of estimators even further, at no additional computational cost.

There are 84 different AMBER/GAFF potential types present in the studied systems, and 80 of them are represented more than 1000 times in all conformers of all molecules confounded. *RMSE* distributions over individual atoms of given type would lead to overcrowded plots, hence the limited number of case studies in Figure 9. However, *RMSE* and associated determination coefficients R^2 were reported over the entire pool of representative instances of each potential type, sorted from the most often encountered *c3* (5.8 million instances) to *hs* (1000 instances).

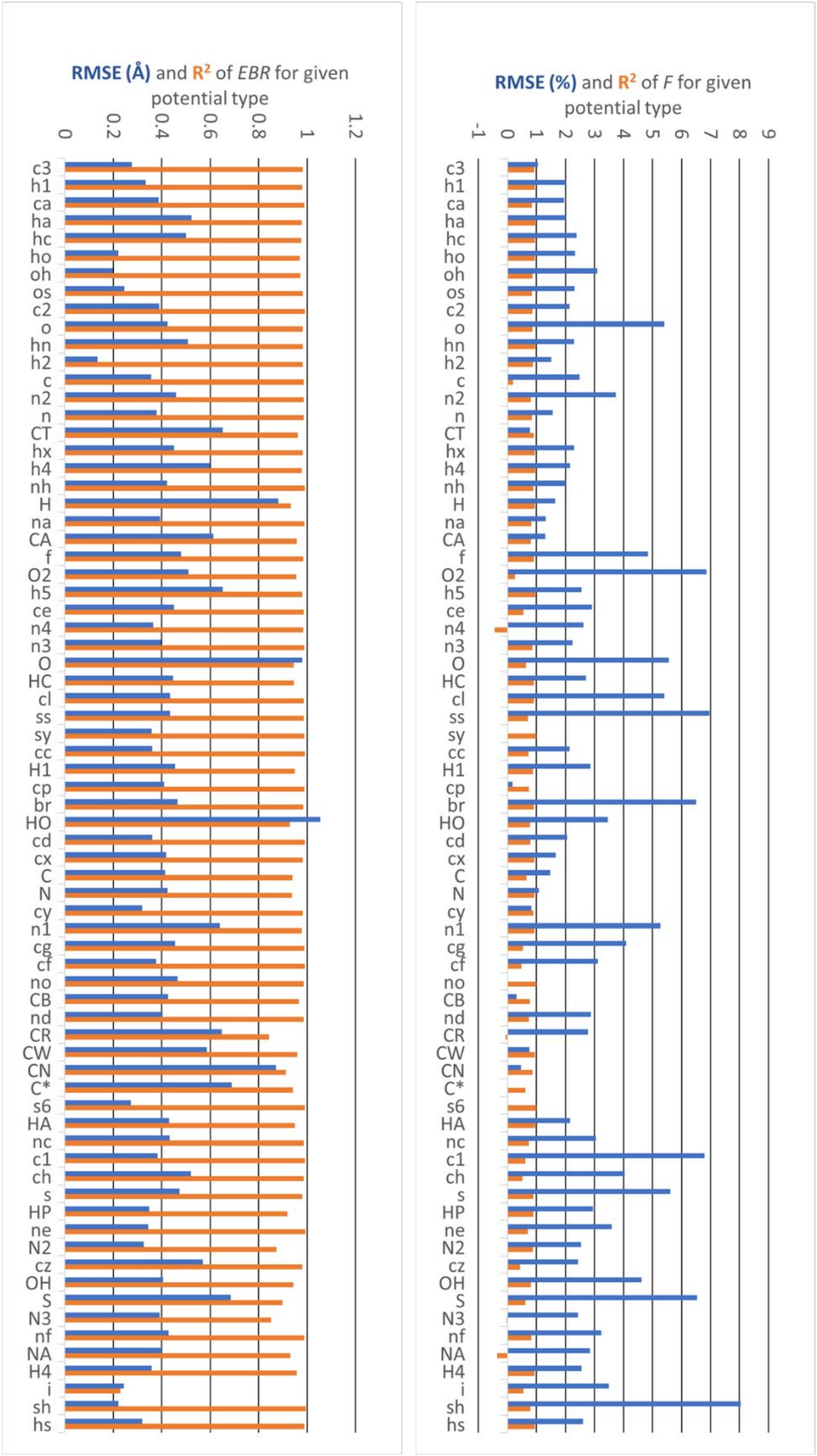


Figure 10: Charts of *RMSE* and R^2 values comparing reference and approximate *EBR* and *F* values over the entire pools of representative instances associated to each potential type.

In terms of *EBR* (left-hand chart in Figure 10) *RMSE* values only exceed 1 Å for the *HO* type, even though the associated R^2=0.93 is very high. This atom type is witnessing very strong *EBR* variations, depending on its exact embedding in macromolecules and on the conformation – in spite of large *RMSE*, this variation is very well mimicked by the empirical models. In the right-hand chart, F values were expressed in % in order to obtain ranges that can be juxtaposed with the (0,1) domain of R^2 values. Note that some potential types correspond to atoms always connected to bulky shielding neighbors: they are intrinsically buried irrespectively of geometry, and correctly predicted as such. In these cases, R^2 is not defined as there is no variation of F to be reproduced (it was set to 1.0, since the model behaves properly in these cases). Unsurprisingly, the highest *RMSE* values coincide for the potential types of atoms connected to one (-H, -Halogen, =O) or two (-O-, -S-) covalent neighbors, for these are the atom types most prone to switch from solvent-accessible to buried. High *RMSE* (never exceeding 8%) is however matched by high R^2 values.

3.3 How accurately is GB/SA energy estimated on the basis of modeled EBR and F values?

As explained in §2.6, that herein reported *RMSE/R*^{2} values ignore systematic offsets of energy levels: they correspond to the unit-slope free-intercept regression line $GBSA = 1.0 \times \widehat{GBSA} + C$. Out of all simulated molecular systems, in 99.3% of cases the root-mean-squared error of the GB/SA term estimated by modeled EBR and F values lies below the 3 kcal/mol mark with respect to reference GB/SA energies. As force-field-based energy values are hardly expected to be accurate within a few kcal/mol, these results clearly show that the chemoinformatics-based model is well suited to be used in this context. For 48% of the training molecules and 41% of the external test molecules, this *RMSE* value is of less than 0.25 kcal/mol. In protein-ligand complexes, such near-perfect fit is slightly less often seen, but nevertheless occurs in 20…30% of complexes. In 100% of complexes (including those of the external test set), the *RMSE* is lower than 3 kcal/mol (Figure 11).

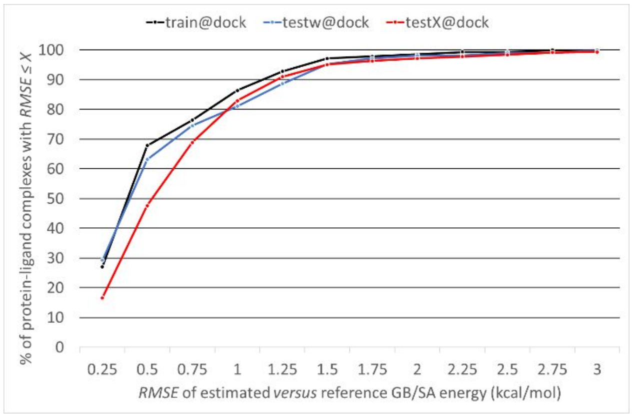


Figure 11: Specific *RMSE* cumulative density plots for the GB/SA energy, over atoms within protein-ligand complexes, within docked poses by S4MPLE in rigid active sites.

Estimated GB/SA energies significantly correlate with the reference terms, as exemplified below for both protein-ligand complexes and MD-simulated host-guest complexes.

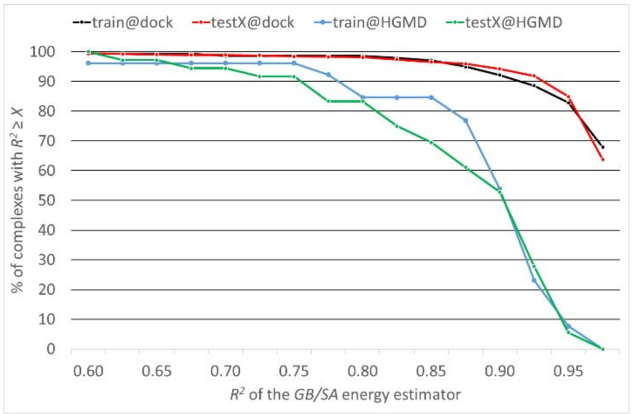


Figure 12: Cumulative density plots of the correlation coefficient R^2 between EBR/F model-based and reference GB/SA energy values, over protein-ligand complexes (@dock) and respectively host-guest complexes sampled by Molecular Dynamics (@HGMD)

Notice (Figure 12) that correlation in terms of energies is much stronger than the actual correlation of EBR values as such (compare to Figure 6, right plot). Energy correlation over evolutionary algorithm-generated docking poses of protein-ligand complexes is perfect in almost 70% of cases. For MD-generated trajectories of host-guest systems, perfect correlation is harder to achieve as MD-driven energy fluctuations are much finer than those observed in non-physical conformational space sampling. Results are, nevertheless, quite encouraging.

Out of the more than 2300 herein studied molecules and complexes, the largest ever *RMSE* deviation between reference and model-based GB/SA energies was observed at R_w=1.1 Ǻ for the villin headpiece peptide 1VII, herein represented by 77 conformers (one native and 76 S4MPLEgenerated random geometries). This *RMSE* value is of 7.04 kcal/mol. However, this corresponds to a near-perfect correlation, as shown in Figure 13. Model-based estimates of GB/SA energies are systematically higher by 21.7 kcal/mol with respect to their reference counterparts, but this constant offset has no consequence on the quality of this near-perfect correlation (R^2=0.985, at unit slope). If a dielectric boundary offset value of R_w=1.4 Ǻ is chosen (herewith weakening any Born terms), correlation is improved (*RMSE*=5.6 kcal/mol, R^2=0.988). Following what seems to be a general trend of the herein approximated magnitudes, the largest *RMSE* errors typically appear in cases with largest fluctuations of reference values, and in the context of the latter turn out to be perfectly acceptable, as proven by the high levels of ensuing degrees of correlation.

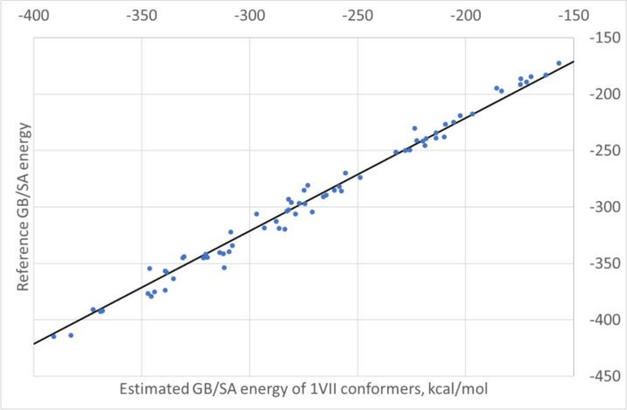


Figure 13: Correlation of reference *versus* model-based estimator of GB/SA energy values of the villin headpiece conformers, the "worst" of all study cases in terms of *RMSE*. The shown "trend line" of slope 1.0 corresponds to *RMSE*=7 kcal/mol, but R^2=0.985

3.4 Computational Cost Assessment

The impact of an algorithm change in a molecular mechanics approach on computational speed is not straightforward to predict. This is not only a matter of the number of floating point operations required, but also of memory management and depends on the peculiar architecture of the code, the used compiler, etc. Therefore, we do not see any way to achieve any technically meaningful comparison between this approximate GB/SA implementation and alternative, cited approaches. Even if we would take the (large) effort to implement those approaches into S4MPLE, the local implementation would significantly differ from the initial one. S4MPLE is fully object-oriented: internal coordinates are objects, managing their own methods to estimate energy contributions. While this ensures readability of the code, it is very difficult to predict how changes in that architecture will impact execution times. During the development of S4MPLE, the accent was never set on code optimization but rather on maintaining an as large as possible domain of applicability. Therefore, any comparison with professionally optimized "standard" MD codes makes, in our eyes, little sense. Furthermore, S4MPLE is not a MD program and has completely different constraints in terms of pair list update strategies, for example. Also, this work provides a unified framework providing fast access to both EBR and surface area values – the costly descriptor calculation is thus twice as rentable, compared to an EBR-focused approximative approach introducing novel terms only for this unique purpose.

The question we wish to tackle here is thus a practical one: by how much slower will energy evaluation become upon the use of approximate GB/SA model, compared to the default energy function using a pairwise desolvation term that is proportional to the sum of squares of the charges of the two atoms, divided by the fourth power of separating distance. For comparison, the same question was asked with respect to the reference, integral-based GB/SA model. Results are shown in Figure 14, where the CPU user time/default S4MPLE energy function evaluation is reported on *X*, and the slow-down factors due to replacing the default desolvation term by the approximate (blue) and respectively integral-based (orange) GB/SA methods are plotted on *Y*. All these times cover the operations needed to return an energy values from atomic coordinates, *i.e.* include the computer effort to update internal coordinates, geometric descriptors or volume/surface integrals, respectively, update EBR and F values and eventually calculate energy. Valence, torsional and Coulomb/van der Waals term estimation effort is the same in all functions, but this offset was not separately evaluated – the interesting indicator being the global slow-down factor. As expected, molecular systems are ordered along the *X* axis by growing size, with default S4MPLE energy estimation times spanning seven orders of magnitude. (Sometimes overlapping) zones can be assigned to each of the monitored classes of chemical entities – from fragment-like molecules to large protein binding sites of thousands of atoms. Interestingly, a consensus emerges throughout the realm of ligands, ligand-cyclodextrin complexes and up to the smaller proteins (1LE1): the approximate GB/SA-based energy model is roughly ten times slower that the default S4MPLE energy term. However, the explicit, integral-based approach would have been more than 100× slower.

Two marked exceptions can be noticed. First, for larger systems (proteins), the slow-down ratio of the approximate GB/SA model tends to decrease (linearly, in the double-log-scale plot) with the default computer cost, meaning that geometric descriptor calculation tends to become (relatively) more effective in larger compounds. This rather good news was unexpected, and a technical explanation would be hard to find. By contrast, integral-based effort slowdown linearly increases, in a mirror-like fashion (see the trapezoidal zone of "Protein Sampling").

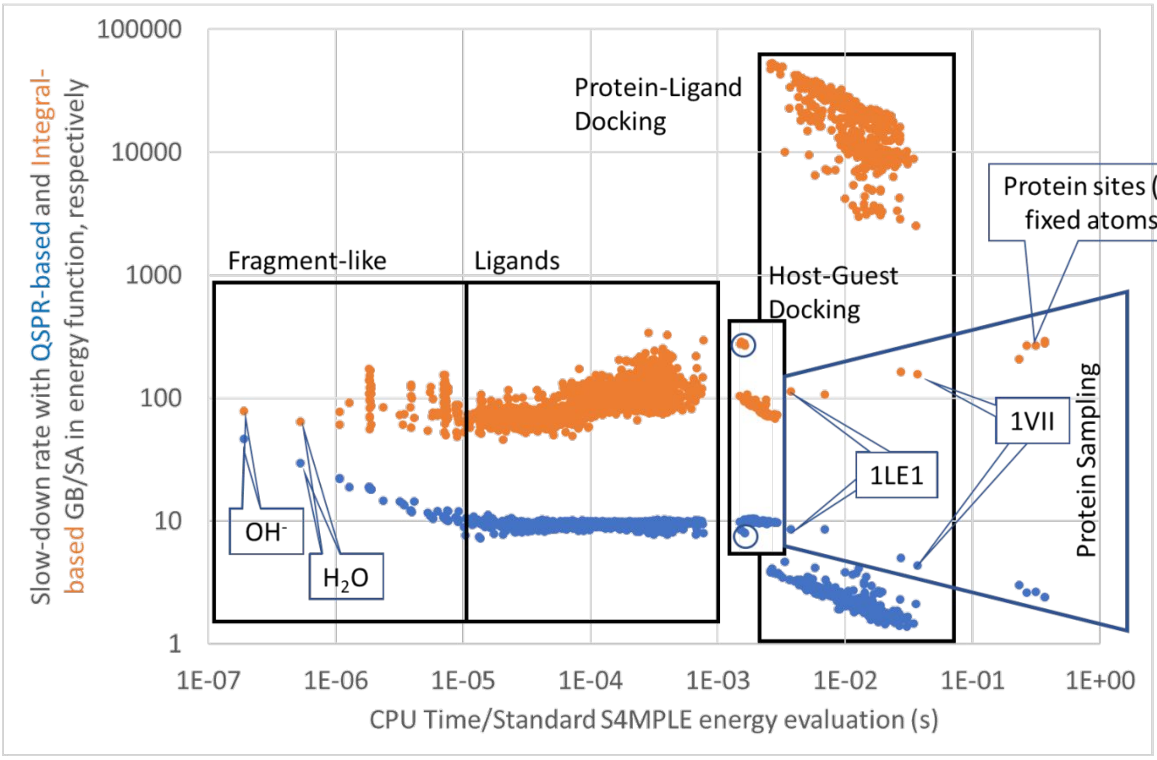


Figure 14. Slow-down rates of energy function calculation triggered by using approximate QSAR-based (blue) and respectively integral-based (orange) GB/SA terms instead of the default desolvation term of S4MPLE. Areas specific to different classes of chemical entities were delimited (rectangles, trapezes) and annotated. Extreme dots were annotated by the compounds they correspond to (the four blue dots at bottom right correspond to protein sites – connector was left out for clarity) Circled spots represent complexes resulting from semi-flexible docking into cyclodextrin.

The most notable differences occur however for docking problems, in which active sites are fixed. The approximate GB/SA fares very well in docking: the code was conceived such as to precompute the contributions from fixed atom pairs to the geometric descriptors. At give geometry, the program only has to update these values with contributions of pairs including at least one mobile atom. By contrast, the integral approach was *not* optimized in such a way (it might have, but this effort was not undertaken for a procedure only meant to provide reference EBR values). This effect is also visible, to a lesser extent, in host-guest docking: while in the MD-based simulations of host-guest complexes all atoms were free (these are the linearly scattered points within the host-guest zone), docking simulations into a rigid-frame cyclodextrin (allowing only for the mobility of polar H) occupy the circled spots (blue) below and (orange) above.

The observed slow-down factors do however not imply that GB/SA-driven S4MPLE simulations will necessarily be 2…10 times longer than with the current tool. S4MPLE genetic operators are prone to often visit non-physical conformational space zones, in which the default energy function, or even simpler energy scores (covalent terms and repulsive van der Waals only) need to be used in order to relax the geometry back to some clash-free and valence constraint-compliant conformer. The GB/SA energy landscape will exclusively be used to further refine the latter; hence its computational overhead will depend on the ruggedness of the explored energy landscape, in a way too early to envisage at this point.

4. Conclusions & Perspectives

This is a proof-of-concept study, advocating an original chemoinformatics-based approach to the fast estimation problem of the GB/SA solvation term and the herein required Effective Born Radii and accessible surface areas. The so-far privileged strategy in the field consisted in finding simplified (but still mathematically complex) analytical expressions to mimic the expected results of the prohibitively time-consuming integral equations. Yet, it had to accept fitable coefficients entering those expressions, and realized that their fitting was dependent on the nature of targeted molecules or complexes. By contrast, the chemoinformatics approach fully assumes its empirical nature, and judges the utility of empirical descriptors as explaining variables of a molecular or atomic property by nothing else but the statistical impact they bring when made available to machine learning. Without needing to guess what the debatably best analytical form for approximations of complex integral expressions may look like, we designed simple and intuitively appealing topological and geometric descriptors of atoms and their successive neighborhoods, then showed that these descriptors capture all the needed information to explain how atomic EBR and F values depend on molecular geometries. Furthermore, the models were calibrated and validated in a compound class-independent manner, spanning >2300 distinct species, from butane to large protein-ligand complexes and host-guest systems, within a large training set (and even larger external test sets).

The approach is by at least one orders of magnitude faster than the integral method and was specifically designed to be particularly effective for docking into (predominantly) rigid protein sites. For this category of problems, the total energy function takes 2…5 times more than the current (fast but inaccurate) evaluation.

Both EBR and F monitor the degree of burial of an atom within the low dielectric "interior" of a molecule – but describe this burial from distinct perspectives. The EBR is sensitive to how deeply the atom is buried, whereas F simply monitors the extent to which the atomic sphere participates at the solute-solvent interface. In spite of this difference, the designed descriptors are able to successfully approximate both magnitudes, by adapting their linear coefficients. Most important, GB/SA energy values estimated on the basis of modeled EBR and F were found to be very close to reference values (with EBRs and Fs from numerical integration). Average imprecision (*RMSE*) was found to be conveniently low – including in thousands of species never seen at the training stage. Even with external test compounds, in 99.4% the root-mean-squared error in GB/SA energy committed by the estimator would be below 3 kcal/mol, and in 83% of the cases below 1 kcal/mol. Furthermore, higher *RMSE* values systematically appear in atoms subject to the strongest variance of their properties, rendering them acceptable in that particular context (as proven by very high ensuing correlation coefficient values). The method is therefore particularly well suited for non-physical sampling techniques, like the evolutionary procedure in S4MPLE, since they trigger large-scale modifications of geometry and hence cover wider energy ranges than molecular dynamics. Nevertheless, MD-based fluctuations are also well approximated by the methodology. The ability to generate "Big Data" for the calibration of these EBR and F models implicitly addressed the problem of its applicability domain – the method was trained to work for a wide variety of systems and behaves correctly throughout the molecular size range. Fragment-like and drug-like ligands have EBR values and GB/SA energies which change little as a function of geometry, so near-constant and rather accurate values are consistently returned by our approach. Due to inclusion of topological atom descriptors, the approach generalizes previous attempts to learn atom-specific constant EBR values as a function of their chemical environment. However, unlike cited SMARTS-based approaches, unpractical for macromolecules and impossible to refer to for docking problems, this approach goes beyond the intrinsic working hypothesis that ligand EBR values can be regarded as atom-type specific constants. On the contrary, their burial into the active site upon docking is very well captured by the proposed geometric descriptors.

Note that the resulting EBR and F models, based on differentiable geometric descriptors, are in principle also differentiable with respect to atomic coordinates. However, further work is needed to assess whether explicit evaluations of such derivatives would be useful, or whether EBR and F values may be assumed to be practically constant during gradient-based energy minimizations.

Some atom types – in particular protein atom types – were shown to be more difficult to predict than others. Therefore, in the future, local models (based on same descriptors, but with locally fitted coefficients) could be tailor-made for specific potential types. Enough "big data" could be easily generated to this purpose, and there is no penalty to "switch" to atom type-dependent sets of coefficients when applying the linear models to calculate EBRs or Fs.

Eventually, fitted coefficients should be seen as nothing else than an extension of the AMBER/GAFF force field, and should be validated in terms of the global propensity of the resulting solvation-aware molecular Hamiltonian. These might be directly fine-tuned in order to bring conformational sampling results using the GB/SA term in closest agreement to experiment – again, including as many as possible experimental endpoints (predicting solvation energies, ranking of potent ligands ahead of inactives and decoys in docking benchmarks, recognizing native-like folds of structured peptides, *etc*). Alternatively, if needed, equivalent EBR models could be trained such as to approximate the more accurate Poisson-Boltzmann-derived radii. So far, the herein supported observation that EBR and F magnitudes are machine-learnable, and that linear models can describe EBRs and Fs with high accuracy, is a strong incentive to continue this work.

5. Acknowledgements

The High Performance Computing (HPC) Center of the Strasbourg University is acknowledged for technical support.

6. Supporting Information Available:

The huge amount of data used in this work cannot be supplied as a standard downloadable Supplementary Material file. All compounds are available upon request, as .mol2 and .sdf files. The pools of geometries are so far in S4MPLE-specific files but could be reverted to standard format upon request. S4MPLE is accessible on our web server (http://infochim.u-strasbg.fr/spip.php?rubrique152). The file "ebr+fbur-model.txt" is provided in supplementary material (see explanations in §3.1).

7. References

1. Still, W. C.; Tempczyk, A.; Hawley, R. C.; Hendrickson, T., Semianalytical treatment of solvation for molecular mechanics and dynamics. *Journal of the American Chemical Society* **1990**, 112, 6127-6129.

2. Onufriev, A. V.; Izadi, S., Water models for biomolecular simulations. *Wiley Interdisciplinary Reviews-Computational Molecular Science* **2018**, 8, e1347.

3. Kleinjung, J.; Fraternali, F., Design and application of implicit solvent models in biomolecular simulations. *Current Opinion in Structural Biology* **2014**, 25, 126-134.

4. Chen, J. H.; Brooks, C. L.; Khandogin, J., Recent advances in implicit solvent-based methods for biomolecular simulations. *Current Opinion in Structural Biology* **2008**, 18, 140-148.

5. Onufriev, A. V.; Case, D. A., Generalized Born Implicit Solvent Models for Biomolecules. *Annual Review of Biophysics, Vol 48* **2019**, 48, 275-296.

6. Born, M., Volumen und Hydratationswärme der Ionen. *Zeitschrift für Physik* **1920**, 1, 45-48.

7. Onufriev, A.; Case, D. A.; Bashford, D., Effective Born radii in the generalized Born approximation: The importance of being perfect. *Journal of Computational Chemistry* **2002**, 23, 1297-1304.

8. Purisima, E. O.; Sulea, T., Solvation Models: Theory and Validation. *Current Pharmaceutical Design* **2014**, 20, 3266-3280.

9. Wang, E. C.; Weng, G. Q.; Sun, H. Y.; Du, H. Y.; Zhu, F.; Chen, F.; Wang, Z.; Hou, T. J., Assessing the performance of the MM/PBSA and MM/GBSA methods. 10. Impacts of enhanced sampling and variable dielectric model on protein-protein Interactions. *Physical Chemistry Chemical Physics* **2019**, 21, 18958-18969.

10. Horvath, D., van Belle, D., Lippens, G., Wodak, S.J., Development and Parametrization of Continuum Solvent Models. I. Models based on the Boundary Element Method. *J. Chem. Phys.* **1996**, 104, 6679-6695.

11. Lu, B.; Zhang, D.; Mccammon, J. A., Computation of electrostatic forces between solvated molecules determined by the Poisson-boltzmann equation using a boundary element method. *The Journal of chemical physics* **2005**, 122, 214102.

12. Onufriev, A.; Bashford, D.; Case, D. A., Modification of the generalized Born model suitable for macromolecules. *Journal of Physical Chemistry B* **2000**, 104, 3712-3720.

13. Brieg, M.; Wenzel, W., Power Born: A Barnes-Hut Tree Implementation for Accurate and Efficient Born Radii Computation. *Journal of Chemical Theory and Computation* **2013**, 9, 1489-1498.

14. Nguyen, H.; Roe, D. R.; Simmerling, C., Improved Generalized Born Solvent Model Parameters for Protein Simulations. *Journal of Chemical Theory and Computation* **2013**, 9, 2020-2034.

15. Hoffer, L.; Horvath, D., S4MPLE - Sampler For Multiple Protein-Ligand Entities: Simultaneous docking of several entities. *J Chem Inf Model* **2012**, 53, 88-102.

16. Hoffer, L.; Renaud, J.-P.; Horvath, D., In Silico Fragment-Based Drug Discovery: Setup and Validation of a Fragment-to-Lead Computational Protocol Using S4MPLE. *J. Chem. Inf. Model.* **2013**, 53, 836-51.

17. Hoffer, L.; Chira, C.; Marcou, G.; Varnek, A.; Horvath, D., S4MPLE-Sampler for Multiple Protein-Ligand Entities: Methodology and Rigid-Site Docking Benchmarking. *Molecules (Basel, Switzerland)* **2015**, 20, 8997-9028.

18. Hawkins, G. D.; Cramer, C. J.; Truhlar, D. G., Parametrized models of aqueous free energies of solvation based on pairwise descreening of solute atomic charges from a dielectric medium. *Journal of Physical Chemistry* **1996**, 100, 19824-19839.

19. Onufriev, A.; Bashford, D.; Case, D. A., Exploring protein native states and large-scale conformational changes with a modified generalized born model. *Proteins-Structure Function and Bioinformatics* **2004**, 55, 383-394.

20. Onufriev, A. V.; Sigalov, G., A strategy for reducing gross errors in the generalized Born models of implicit solvation. *Journal of Chemical Physics* **2011**, 134, 15.

21. Anandakrishnan, R.; Daga, M.; Onufriev, A. V., An n log n Generalized Born Approximation. *Journal of Chemical Theory and Computation* **2011**, 7, 544-559.

22. Aguilar, B.; Onufriev, A. V., Efficient Computation of the Total Solvation Energy of Small Molecules via the R6 Generalized Born Model. *Journal of Chemical Theory and Computation* **2012**, 8, 2404-2411.

23. Onufriev, A. V.; Izadi, S., Water models for biomolecular simulations. *Wiley Interdisciplinary Reviews-Computational Molecular Science* **2018**, 8.

24. Marenich, A. V.; Cramer, C. J.; Truhlar, D. G., Generalized Born Solvation Model SM12. *Journal of Chemical Theory and Computation* **2013**, 9, 609-620.

25. Brown, R. A.; Case, D. A., Second derivatives in generalized born theory. *Journal of Computational Chemistry* **2006**, 27, 1662-1675.

26. Gaillard, T.; Simonson, T., Pairwise Decomposition of an MMGBSA Energy Function for Computational Protein Design. *Journal of Computational Chemistry* **2014**, 35, 1371-1387.

27. Grant, J. A.; Pickup, B. T.; Sykes, M. J.; Kitchen, C. A.; Nicholls, A., The Gaussian Generalized Born model: application to small molecules. *Physical Chemistry Chemical Physics* **2007**, 9, 4913-4922.

28. Gallicchio, E.; Levy, R. M., AGBNP: An analytic implicit solvent model suitable for molecular dynamics simulations and high-resolution modeling. *Journal of Computational Chemistry* **2004**, 25, 479-499.

29. Liu, H. Y.; Kuntz, I. D.; Zou, X. Q., Pairwise GB/SA scoring function for structure-based drug design. *Journal of Physical Chemistry B* **2004**, 108, 5453-5462.

30. Forouzesh, N.; Izadi, S.; Onufriev, A. V., Grid-Based Surface Generalized Born Model for Calculation of Electrostatic Binding Free Energies. *Journal of Chemical Information and Modeling* **2017**, 57, 2505-2513.

31. Arthur, E. J.; Brooks, C. L., Parallelization and Improvements of the Generalized Born Model with a Simple sWitching Function for Modern Graphics Processors. *Journal of Computational Chemistry* **2016**, 37, 927-939.

32. Zhang, B. F.; Kilburg, D.; Eastman, P.; Pande, V. S.; Gallicchio, E., Efficient Gaussian Density Formulation of Volume and Surface Areas of Macromolecules on Graphical Processing Units. *Journal of Computational Chemistry* **2017**, 38, 740-752.

33. Tanner, D. E.; Phillips, J. C.; Schulten, K., GPU/CPU Algorithm for Generalized Born/Solvent-Accessible Surface Area Implicit Solvent Calculations. *Journal of Chemical Theory and Computation* **2012**, 8, 2521-2530.

34. Huang, H.; Simmerling, C., Fast Pairwise Approximation of Solvent Accessible Surface Area for Implicit Solvent Simulations of Proteins on CPUs and GPUs. *Journal of Chemical Theory and Computation* **2018**, 14, 5797-5814.

35. Mongan, J.; Simmerling, C.; McCammon, J. A.; Case, D. A.; Onufriev, A., Generalized Born model with a simple, robust molecular volume correction. *Journal of Chemical Theory and Computation* **2007**, 3, 156-169.

36. Wojciechowski, M.; Lesyng, B., Generalized born model: Analysis, refinement, and applications to proteins. *Journal of Physical Chemistry B* **2004**, 108, 18368-18376.

37. Nguyen, D. D.; Wei, G. W., The Impact of Surface Area, Volume, Curvature, and Lennard-Jones Potential to Solvation Modeling. *Journal of Computational Chemistry* **2017**, 38, 24-36.

38. Brieg, M.; Setzler, J.; Albert, S.; Wenzel, W., Generalized Born implicit solvent models for small molecule hydration free energies. *Physical Chemistry Chemical Physics* **2017**, 19, 1677-1685.

39. Lee, K. H.; Chen, J. H., Optimization of the GBMV2 implicit solvent force field for accurate simulation of protein conformational equilibria. *Journal of Computational Chemistry* **2017**, 38, 1332-1341.

40. Katkova, E. V.; Onufriev, A. V.; Aguilar, B.; Sulimov, V. B., Accuracy comparison of several common implicit solvent models and their implementations in the context of protein-ligand binding. *Journal of Molecular Graphics & Modelling* **2017**, 72, 70-80.

41. Zhang, H. Y.; Yin, C. H.; Yan, H.; van der Spoel, D., Evaluation of Generalized Born Models for Large Scale Affinity Prediction of Cyclodextrin Host-Guest Complexes. *Journal of Chemical Information and Modeling* **2016**, 56, 2080-2092.

42. Hoffer, L.; Saez-Ayala, M.; Horvath, D.; Varnek, A.; Morelli, X.; Roche, P., CovaDOTS: In Silico Chemistry-Driven Tool to Design Covalent Inhibitors Using a Linking Strategy. *Journal of Chemical Information and Modeling* **2019**, 59, 1472-1485.

43. Case, D. A.; Darden, T. A.; Cheatham, T. E.; Simmerling, C. L.; Wang, J.; Duke, R. E.; Luo, R.; Walker, R. C.; Zhang, W.; Merz, K. M.; Roberts, B.; Hayik, S.; Roitberg, A.; Seabra, G.; Swails, J.; Goetz, A. W.; Kolossvary, I.; Wong, K. F.; Paesani, F.; Vanicek, J.; Wolf, R. M.; Liu, J.; Wu, X.; Brozell, S. R.; Steinbrecher, T.; Gohlke, H.; Cai, Q.; Ye, X.; Hsieh, M.-J.; Cui, G.; Roe, D. R.; Mathews, D. H.; Seetin, M. G.; Salomon-Ferrer, R.; Sagui, C.; Babin, V.; Luchko, T.; Gusarov, S.; Kovalenko, A.; Kollman, P. A., AMBER 12. *University of California* **2012**.

44. Wang, J. M.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A., Development and testing of a general amber force field. *Journal of Computational Chemistry* **2004**, 25, 1157-1174.

45. Zhang, W.; Hou, T. J.; Xu, X. J., New Born radii deriving method for generalized Born model. *Journal of Chemical Information and Modeling* **2005**, 45, 88-93.

46. Oshima, H.; Kinoshita, M., A Highly Efficient Hybrid Method for Calculating the Hydration Free Energy of a Protein. *Journal of Computational Chemistry* **2016**, 37, 712-723.

47. Li, J. N.; Abel, R.; Zhu, K.; Cao, Y. X.; Zhao, S. W.; Friesner, R. A., The VSGB 2.0 model: A next generation energy model for high resolution protein structure modeling. *Proteins-Structure Function and Bioinformatics* **2011**, 79, 2794-2812.

48. Robinson, M. K.; Monroe, J. I.; Shell, M. S., Are AMBER Force Fields and Implicit Solvation Models Additive? A Folding Study with a Balanced Peptide Test Set. *Journal of Chemical Theory and Computation* **2016**, 12, 5631-5642.

49. Lee, S.; Cho, K. H.; Lee, C. J.; Kim, G. E.; Na, C. H.; In, Y.; No, K. T., Calculation of the Solvation Free Energy of Neutral and Ionic Molecules in Diverse Solvents. *Journal of Chemical Information and Modeling* **2011**, 51, 105-114.

50. RCSB Protein Data Bank. http://www.rcsb.org/pdb/

51. Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P., ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Research* **2011**, 40, D1100-D1107.

52. ChemAxon, In; 2008.

53. Zhenin, M.; Bahia, M. S.; Marcou, G.; Varnek, A.; Senderowitz, H.; Horvath, D., Rescoring of docking poses under Occam's Razor: are there simpler solutions? *Journal of Computer-Aided Molecular Design* **2018**, 32, 877-888.

54. Horvath, D.; Brown, J.; Marcou, G.; Varnek, A., An Evolutionary Optimizer of libsvm Models. *Challenges* **2014**, 5, 450-472.

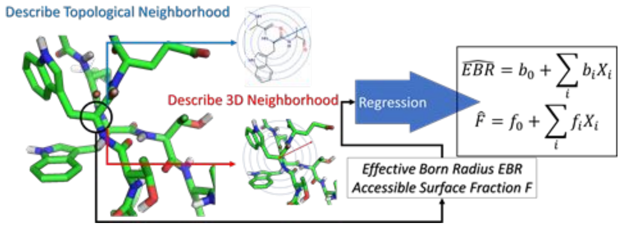
55. Harrell, F. E., *Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis*. Springer New York: 2013.

56. Brown, J. B.; Okuno, Y.; Marcou, G.; Varnek, A.; Horvath, D., Computational chemogenomics: Is it more than inductive transfer? *J. Comput.-Aided Mol. Des.* **2014**, 28, 597-618.

57. Varnek, A.; Gaudin, C. d.; Marcou, G.; Baskin, I.; Pandey, A. K.; Tetko, I. V., Inductive Transfer of Knowledge: Application of Multi-Task Learning and Feature Net Approaches to Model Tissue-Air Partition Coefficients. *J. Chem Inf. Model.* **2009**, 49, 133-144.

58. Horvath, D.; Lippens, G.; vanBelle, D., Development and parametrization of continuum solvent models .2. A unified approach to the solvation problem. *Journal of Chemical Physics* **1996**, 105, 4197-4210.

8. Table of Contents Graphic


Describe Topological Neighborhood
Describe 3D Neighborhood
Regression
$\widehat{EBR} = b_0 + \sum_i b_i X_i$
$\hat{F} = f_0 + \sum_i f_i X_i$
Effective Born Radius EBR
Accessible Surface Fraction F