Structural bioinformatics

Generalized Born radii computation using linear models and neural networks

Saida Saad Mohamed Mahmoud [1,2] and Gennaro Esposito,[3] and Giuseppe Serra [1,*] and Federico Fogolari [1,*]

[1]Department of Mathematics, Informatics and Physics, University of Udine, 33100, Italy
[2]Faculty of Science, Cairo University, 1 Gamaa Street, 12613, Giza, Egypt
[3]Science and Math Division, New York University at Abu Dhabi, PO Box 129188, Abu Dhabi, United Arab Emirates

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Implicit solvent models play an important role in describing the thermodynamics and the dynamics of biomolecular systems. Key to an efficient use of these models is the computation of Generalized Born (GB) radii, which is accomplished by algorithms based on the electrostatics of inhomogeneous dielectric media. The speed and accuracy of such computations is still an issue especially for their intensive use in classical molecular dynamics. Here, we propose an alternative approach that encodes the physics of the phenomena and the chemical structure of the molecules in model parameters which are learned from examples.

Results: GB radii have been computed using i) a linear model and ii) a neural network. The input is the element, the histogram of counts of neighbouring atoms, divided by atom element, within 16 Å. Linear models are ca. 8 times faster than the most widely used reference method and the accuracy is higher with correlation coefficient with the inverse of "perfect" GB radii of 0.94 vs 0.80 of the reference method. Neural networks further improve the accuracy of the predictions with correlation coefficient with "perfect" GB radii of 0.97 and ca. 20% smaller root mean square error.

Availability: We provide a C program implementing the computation using the linear model, including the coefficients appropriate for the set of Bondi radii, as supplementary material. We also provide a Python implementation of the neural network model with parameter and example files in the supplementary material as well.

Contact: federico.fogolari@uniud.it, giuseppe.serra@uniud.it

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Implicit solvent models have a long history. Indeed the dielectric constant of materials, used first by Faraday to describe their electrostatic behaviour, is the first implicit representation of matter electrostatics (Whittaker, 1910). Models for ionic solutions based on the Poisson equation and the Boltzmann distribution were introduced more than one century ago, and further used in the following decades to model polyionic molecules, including proteins, their ionization and their solvation free energy (Fogolari *et al.*, 2002). A field where implicit solvent models have been particularly fruitful is represented by molecular dynamics (MD) simulations which are widely used to understand the behaviour of biomolecules. In many applications, MD simulations are used to generate a conformational ensemble for a given system, rather than to get kinetic informations. In explicit solvent simulations, it is quite normal that 90% of the computational time is spent on simulating the solvent, which is often simply not considered in the analysis of the results. These well known considerations have led to the development of implicit solvent models suitable for MD simulations. An additional advantage is that solvation free energy (including entropy) is simply described by implicit solvent models (Fogolari *et al.*, 2018).

Molecular dynamics simulations based on the Poisson-Boltzmann

equation implicit solvent representation, have indeed ben proposed. However this approach is characterized by an intensive computational load and by problems related to the numerical solutions on spatial grids (Sharp, 1991; Niedermeier and Schulten, 1992; Gilson *et al.*, 1995; David *et al.*, 2000; Fogolari *et al.*, 2003). Mostly in the nineties, Generalized Born (GB) models have been developed, with different ways to compute the GB radii which are the key parameters of the model (Constanciel and Contreras, 1984; Still *et al.*, 1990; Hawkins *et al.*, 1995, 1996; Qiu *et al.*, 1997). Solvation and interaction energies are computed based on generalized Born radii (Bashford and Case, 2000; Onufriev and Case, 2019).

Theoretical "perfect" GB radii (Onufriev *et al.*, 2002), computed using the reference Poisson-Boltzmann equation (PBE) model, may be approximated very accurately using surface integrals, but this possibility is ruled out for use in MD simulations, because of the computational cost. For this reason, surface (or equivalently) volume integrals are usually approximated using sums over neighbouring atoms.

The model by Onufriev, Bashford and Case (Onufriev *et al.*, 2004) (OBC) for computing GB radii is one of the most used ones and will be considered here as the reference for comparison, because it recapitulates many aspects of other models.

The accuracy of this and other models in reproducing "perfect" GB radii, is still low, as shown below. Following a similar approach, but targeting $1/r_{GB}^3$ instead of $1/r_{GB}$, a significant improvement was achieved by the AR6 method by Onufriev and co-workers at a similar computational load (Aguilar *et al.*, 2010). Corrections to a similar model resulted in the successful GB-Neck2 method (Mongan *et al.*, 2007c).

A more heuristic approach, named FACTS and implemented in the simulation program CHARMM (Brooks *et al.*, 2009), incorporates both the idea of volume occupation by surrounding atoms and the asymmetry of the distribution (Haberthur and Caflisch, 2008). The latter approach reaches higher accuracy with limited amount of computation compared to the OBC or similar analytical approaches.

The search for fast and accurate methods to compute GB radii (and possibly their derivatives with respect to atomic positions) is however still an open issue.

Recently, machine learning approaches have been used to construct multibody coarse grained potentials (Zhang *et al.*, 2018; Wang *et al.*, 2019) or to learn atomistic force-fields (Behler, 2016). With their flexibility machine learning methods are expected to achieve the highest accuracy given a large body of training data. The GB implicit solvent model is a peculiar form of coarse graining and therefore the above works share common aspects with the approach adopted here. The main difference is that here the solute is still described by a high quality force field whereas only the solvation part is modeled, adopting the widely used GB functional form for solvation free energy.

Our aim is to obtain a method to compute the GB radii from the conformation that is faster and more accurate than available methods, a task much less complex (and perhaps less accurate as well) than learning conformation dependent effective potentials.

Here we use linear models and neural networks, as alternative functional forms, to compute GB radii from histograms of distances and we show that better accuracy can be obtained with lesser or comparable amount of computation with respect to current widely employed methods.

2 Methods

Generalized Born radii

The basis for most approaches to computation of GB radii is the Coulomb field approximation (CFA) which assumes that the electric displacement $\vec{D}$ may be approximated by the Coulombic field $\vec{E}$ as if the medium were homogeneous (Bashford and Case, 2000; Onufriev and Case, 2019; Bardhan, 2008):

$$\vec{D} = \epsilon_0 \epsilon_r \vec{E} = \frac{1}{4\pi} \frac{q}{r^2} \tag{1}$$

where q is the source charge, ϵ_0 is the vacuum permittivity, ϵ_r is the relative dielectric constant of the solvent and r is the distance from the source charge where the $\vec{D}$ is computed.

Let us consider the single charge q_i of the atom i with van der Waals radius a_i embedded in a molecular volume V_{in}. The energy U of the system is computed by integrating the electrostatic energy density:

$$U = \frac{1}{2} \int_V \vec{E} \cdot \vec{D} dV \tag{2}$$

The solvation energy ΔG_i of point charge q_i embedded in a molecular structure is computed therefore as the difference between the self energy in vacuum and in solvent in the CFA approximation which leads to the integral over the solvent volume V_{ext}:

$$\Delta G_i = \frac{\epsilon_0}{2} \left(1 - \frac{1}{\epsilon_r}\right) \frac{q_i^2}{(4\pi\epsilon_0)^2} \int_{V_{ext}} \frac{1}{r^4} dV \tag{3}$$

This solvation energy is equated with the known solvation energy of the same charge embedded at the center of a sphere of radius α_i:

$$\Delta G_i = \left(1 - \frac{1}{\epsilon_r}\right) \frac{1}{8\pi\epsilon_0} \frac{q_i^2}{\alpha_i} \tag{4}$$

thus defining α_i as a purely geometric quantity:

$$\frac{1}{\alpha_i} = \frac{1}{4\pi} \int_{V_{ext}} \frac{1}{r^4} dV \tag{5}$$

It is further convenient to consider the integral over the whole volume external to the isolated atom, i.e. from infinite to its van der Waals radius:

$$\frac{1}{a_i} = \frac{1}{4\pi} \int_{V_{r>a_i}} \frac{1}{r^4} dV \tag{6}$$

and rewrite equation 5 as:

$$\frac{1}{\alpha_i} = \frac{1}{a_i} - \frac{1}{4\pi} \int_{V_{in, r>a_i}} \frac{1}{r^4} dV \tag{7}$$

Most approaches, save for those explicitly integrating over the volume or equivalently, using the divergence theorem, over the surface (Ghosh *et al.*, 1998; Fogolari *et al.*, 2012, 2013), replace the integral over the molecular volume surrounding atom i by a discrete sum over atoms' contributions (Still *et al.*, 1990; Qiu *et al.*, 1997; Hawkins *et al.*, 1995, 1996; Mongan *et al.*, 2007b; Tjong and Zhou, 2007):

$$\frac{1}{\alpha_i} = \frac{1}{a_i} - \sum_j f(a_i, a_j, r_{ij}) \tag{8}$$

Following Hawkins, Cramer and Truhlar (Hawkins *et al.*, 1995, 1996), the atomic radii are modified and scaled in the terms entering the sum. Furthermore the method of Onufriev, Bashford and Case subjects the summation to processing in order to smooth the dependence of the results on the summation (Onufriev *et al.*, 2004). The function in eq. 8 is parametrized as to avoid double counting due to overlap of atoms.

More accurate approaches estimate $\left(\frac{1}{\alpha_i}\right)^3$ rather than $\frac{1}{\alpha_i}$ (Grycuk, 2003; Mongan *et al.*, 2007a; Tjong and Zhou, 2007; Fogolari *et al.*, 2012) based on the exact solutions of the Poisson equation for a spherical boundary in the limit of infinite external dielectric constant, rather than on the Coulomb Field approximation.

Reference Generalized Born radii

For comparison the Onufriev, Bashford and Case (OBC) GB radii computation was used with the formulae and parameters reported in the NAMD program (Kale *et al.*, 1999) user guide and in the program files. This implementation might be slightly different from that reported in the original paper (Onufriev *et al.*, 2004).

Rather than comparing the "perfect" GB radii directly with the computed GB radii, the inverse of the Generalized Born radii were considered as the target to assess the quality of the estimation provided by the different methods, because proportional to self-solvation energy. It has been shown before that the "perfect" GB radii may be computed very accurately by a surface integral approach (Fogolari *et al.*, 2012, 2013; Izadi *et al.*, 2018). To this end the molecular surface was generated using the program MSMS (Sanner *et al.*, 1996), read in a customized version of the program Bluues and GBR6 radii were calculated by numerical integration (Fogolari *et al.*, 2012).

The set of radii used was the Bondi radii set (Bondi, 1964) as implemented in the program NAMD (Kale *et al.*, 1999) which assigns radii 1.2, 1.5, 1.55, 1.70, 1.8 Å to H, O, N, C and S, respectively. Polar hydrogens were assigned a radius of 1.3 Å. With minor differences the same set of radii was implemented in versions 4 and 5 of the GROMACS simulation software package (Berendsen *et al.*, 1995).

The top500H (Lovell *et al.*, 2003) curated dataset, which includes 500 non redundant protein structures obtained by X-ray crystallography with resolution better than 1.8 Å and with few deviations from ideal geometry, was used and overall 1.7 million "perfect" GB radii were calculated.

Generalized Born energies and forces

Once GB radii have been computed, it is possible to compute system energies, that include solvation energies, using pairwise summation. In the following, to make notation less clumsy, we indicate GB radii $r_{GB,i}$ by α_i.

$$\Delta G = U + \Delta G^{solv} = \frac{1}{4\pi\epsilon_0}\sum_{i>j}\frac{q_i q_j}{r_{ij}} \quad (9)$$

$$-\frac{1}{8\pi\epsilon_0}(1-\frac{1}{\epsilon_{out}})\sum_{i,j}\frac{q_i q_j}{\sqrt{r_{ij}^2+\alpha_i\alpha_j\exp^{\frac{-r_{ij}^2}{4\alpha_i\alpha_j}}}} \quad (10)$$

The equation reproduces the correct limiting behaviour for r_{GB} radii in the very large and very small distance regimes and provides a smooth transition between the two regimes.

The GB force acting on atom j is minus the derivative of the solvation energy with respect to the position of atom j:

$$\vec{F}_j = -\nabla_{\vec{r}_j}\Delta G(\{\alpha_k(\vec{r}_l)\},\{\vec{r}_l\}) = \quad (11)$$

$$= -\nabla_{\vec{r}_j}\Delta G(\{\alpha_k\},\{\vec{r}_l\}) + \quad (12)$$

$$-\sum_k \frac{\partial\Delta G(\{\alpha_k\},\{\vec{r}_l\})}{\partial\alpha_k}\nabla_{\vec{r}_j}\alpha_k(\{\vec{r}_l\}) \quad (13)$$

where the explicit and implicit dependence on atomic position is indicated. The derivative of the Born radii are further obtained by the chain rule:

$$\nabla_{\vec{r}_j}\alpha_k = \frac{\partial\alpha_k}{r_{kj}}\times\vec{\nabla}_{\vec{r}_j}r_{kj} \quad (14)$$

Learning Generalized Born radii from examples

Here, we represent the environment of each atom through descriptors and we explore the possibility of learning the relationship of these with the Generalized Born radius of the same atom. For the sake of clarity we refer to the atom for which the GB radius is computed as the screened atom, whereas all neighbouring atoms are referred to as the screening atoms.

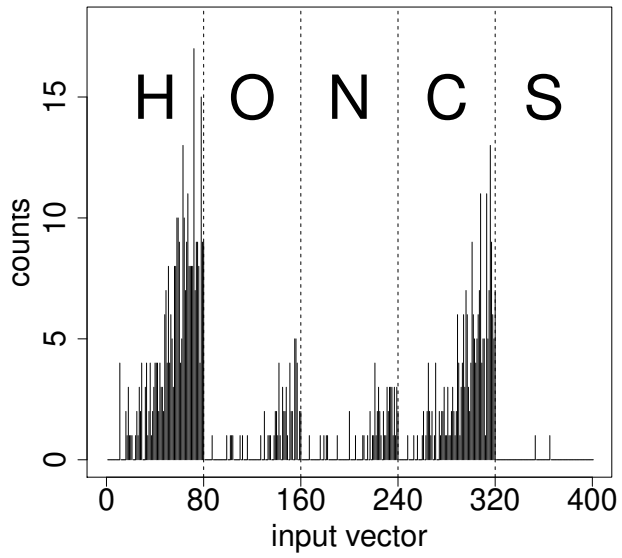

Fig. 1. An example of the input vector for a carbon atom. The five 80-bin histograms corresponding to H, O, N, C, S neighbours counts at distances less than 16 Å are indicated in the plot.

For each screened atom we consider all neighbour screening atoms within 16.0 Å (larger than typical cutoffs used in simulation). We divide all screening atoms according to element and compute for each element (i.e. H, N, C, O or S) the histogram of distances, i.e. the counts of screening atoms occurring in 0.2 Å-wide bins from 0 to 16.0 Å.

For each atom therefore five (one for each element) 80-bin histograms are computed. Since the effect of screening atoms depends also on the element of the reference atom, we consider separately each element, so that for each element we use 400 predictive variables.

In summary for each atom a categorical variable, to indicate the element of the atom, and 400 counts of neighbouring atoms are the input to predict the inverse of the Generalized Born radius. An example of the input vector for a carbon atom is reported in Figure 1. As can be seen the vector is partitioned in five 80-components sections relative to neighbours of a given atomic element. Each component represents a distance of the screening atoms from the screened atom. Thus, components 1 to 80 represents the counts of screening hydrogens at distances from 0 to 16 Å from the screened atom, components 81 to 160 represents the counts of screening oxygens at distances from 0 to 16 Å from the screened atom, etc... The number of counts in all sections increases in principle up to the boundary of the protein with the square of the corresponding distances.

The section relative to sulphur atoms is mostly poorly populated due to the low number of sulphur atoms in protens.

As mentioned above, we chose the inverse of the GB radius as the target because self-solvation energies are porportional to this quantity.

Generalized Born radii and forces by multilinear regression

A linear model is built to predict the inverse of the "perfect" GB radii. The inverse of the "perfect" GB radius is fitted by multilinear regression on the components of the aggregate histogram of the neighbouring atoms according to atom element:

$$\frac{1}{\alpha_k} = c_0^{e(k)} + \sum_{ij} c_i^{e(k),e(j)} n_{i,e(j)} \quad (15)$$

where $e(k)$ is the element of screened atom k, and $n_{i,e(j)}$ is the number of screening atoms of element $e(j)$ in the distance bin i.
The coefficients $c_i^{e(k),e(j)}$ may be understood as the discretization of the function of the distance d_{kj} representing the linear effect of element $e(j)$ on the inverse of the Generalized Born radius of element $e(k)$.
In principle the coefficients should represent the effective volume of atoms scaled by the inverse of the fourth power of the distance. In practice fitting the strong correlations in particular at short distances leads also to negative coefficients. Similarly the intercept of the regression should in principle be the inverse of the van der Waals atomic radius, but in practice as discussed in the results section is definitely smaller.
Several count values, e.g. those at very short distances, are zero and therefore their coefficients are not defined. These values were linearly interpolated from other coefficients using the package imputeTS (Moritz and Bartz-Beielstein, 2017).
The number of operations that must be performed for each GB radius to be predicted from the histogram of distances is 799 (400 multiplications and 399 sums). Due to the linearity of the model the GB radius may be computed directly from the distances with 3 operations (modulo, multiplication and sum) for each screening atom.
Generalized Born forces are obtained from equations 13 and 14. Since the inverse of the Generalized Born radii is a linear combination of counts in bin distances, the derivative of the Generalized Born radii is approximated by a linear combination of counts in bin distances using the chain rule:

$$\frac{\partial \alpha_k}{\partial r_{kj}} = -{\alpha_k}^2 \frac{\partial \frac{1}{\alpha_k}}{\partial r_{kj}} \quad (16)$$

$$= -{\alpha_k}^2 \sum_i d_i^{e(k),e(j)} \delta_{i,\mathrm{bin}(r_{kj})} \quad (17)$$

where $\mathrm{bin}(r_{kj})$ is the index of the bin of distance r_{kj}.
Consistent with the intepretation of coefficients $c_i^{e(k),e(j)}$ as the discretization of a continuous function, the coefficients $d_i^{e(k),e(j)}$ are defined as:

$$d_1^{e(k),e(j)} = 0 \quad (18)$$

$$d_i^{e(k),e(j)} = \frac{c_{i+1}^{e(k),e(j)} - c_{i-1}^{e(k),e(j)}}{2\Delta d} \quad i = 2, ..., N-1 \quad (19)$$

$$d_N^{e(k),e(j)} = 0 \quad (20)$$

where Δd is the distance bin width, and N is the number of bins. Setting the first and last coefficients to zero has no consequence because the first bins are never populated due to steric repulsion of atoms, and the last ones are close to cutoff distance where effects are negligible.

Implementation in Molecular Dynamics software

A preliminary test of the methods described here was performed within the software GROMACS v. 5.1.2 (Van Der Spoel *et al.*, 2005), by suitable modifications to data structures related to GB functions and by modifications to the routines that implement GB radii and their derivatives calculation. No attempt was done at this stage to parallelize the code. Since the code is rather complex, checks were performed by printing the GB radii and solvation energies and forces (by difference with vacuum computation using the same structures) both for the linear model and for the reference OBC model, in order to verify that no artifacts were introduced by the modifications to the code. The typical correlation between OBC and linear model solvation forces was always found to be larger than 0.8.
The purpose of this implementation is to test in practice running time and overall accuracy.
We have simulated the protein barnase (1727 atoms) for 50 ns at 310 K using the OBC and the linear GB models. All bonds were restrained by the LINCS algorithm, a cutoff of 1.6 nm was used for all non-bonded terms. GB radii were recalculated every 10 steps. The simulations used a stochastic dynamics integrator at 310 K with time constant 0.1 ps. Using larger time constants resulted in occasional LINCS failures.
The accuracy of the simulation was judged by the time evolution of the backbone RMSD from the starting minimized structure. This is a sensitive parameter which is able to detect excessive restraining on structures, or inaccuracies in forces.
The efficiency of the linear and OBC GB models was assessed on short (50 ps) MD runs where GB radii and derivatives were recalculated at each step, in order to exclude that the differences could be due to the combined effect of conformational changes and cut-offs.

Generalized Born radii using neural networks

A neural network model has been implemented to predict the inverse of the "perfect" GB radii. The proposed network consists of five hidden layers with linear activation function REctified Linear Unit (ReLU) (Nair and Hinton, 2010) for each layer and one hidden layer (the last one) with Sigmoid activation function (Sibi *et al.*, 2013). Each layer consists of 32 neurons except the last layer that has only one neuron to predict the target value. A scheme of the network is shown in Figure 2.
The network was trained using the same input as for the multilinear regression. In practice there are five data sets, one for each screened element H, C, N, S, O respectively. The environment of each screened atom is represented by a 400-element vector, containing five 80-bin histograms corresponding to each screening element H, O, N, C, S with neighbors counts at distances less than 16 Å. The input vector for a screened carbon atom is shown in Figure 1. Thus, five identical neural networks (one for each screened element) are trained with their relative data sets. Each input consists of batch of N vectors of size 400. For each set we train and test the same proposed model. We fit a neural network model for each element and evaluate the model on the test data set for each element. The number of training examples that is utilized in one iteration in the model is 512. We used as optimization algorithm the Root Mean Square Propagation (RMSprop) (Hinton *et al.*, 2012)(URL: http://www.cs.toronto.edu/ hinton/coursera/lecture6/lec6.pdf).

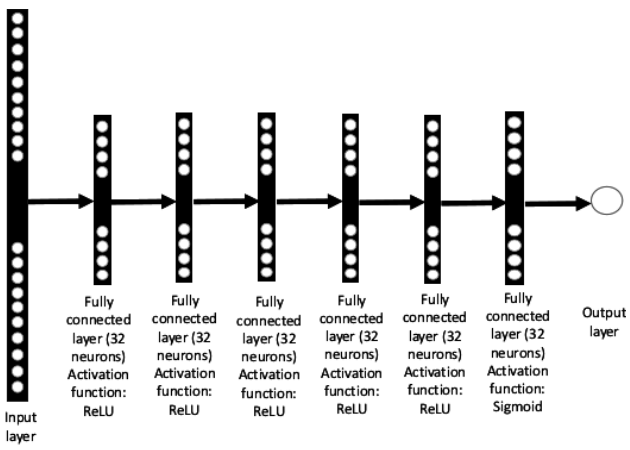

Fig. 2. The schematization of the proposed neural network.

The neural network model for each element, built to predict the inverse of the GB radius from the counts of neighbors, reflects the influence of each other atom on the atom to be predicted. The number of operations performed for each prediction is 35840. Most operations are elementary

operations (multiplication, sum, max, exponential, division).
We tried many different models that are different in number of layers, number of batch size, optimization algorithm and activation function. We compared their results until we got the best result for the architecture described above. To deal to the overfitting issue we adopt in our approach the Early stopping strategy (Prechelt, 1998). In fact, this strategy provides an effective way to evaluate how many epochs can be run before the neural network model begins to overfit. In particular, it compares at every epoch the neural network performance in both training and valuation sets and the learning process in order to preserve the generalization of the network.

Test set

The accuracy of the predictions are tested on the non-redundant, representative set of proteins used by Tjong and Zhou (Tjong and Zhou, 2007) after exclusion of the proteins similar to those used for fitting the multilinear model. The criterion for accepting the structure in the testing set was that the expectation value of the alignment score of its sequence with the sequences of the training set proteins was larger than 0.05. The resulting test set includes 32407 atoms to be predicted.

Results and discussion

Generalized Born radii by multilinear regression

The linear model built to predict the inverse of the GB radius from the counts of neighbours reflects the influence of each other atom on the atom to be predicted. The self contribution which should be equal to the inverse of the van der Waals radius is coded in the intercept of the model which corresponds to radii smaller than expected, because of extensive overlap of atoms.
Thus, the self screening radius, which is obtained from the intercepts of the linear model for each element, is 1.04 Å for H, 1.41 Å for O, 1.34 Å for N, 1.42 Å for C and 1.27 Å for S.
To compare predictions with reference "perfect" GB radii, all predicted GB radii larger than 16 Å were reset to 16 Å and those smaller than van der Waals radii were reset to van der Waals radii.
The coefficients for the five screening elements H, O, N, C , S (for each screened atom) are reported in Figure 3. Although the general behaviour is as expected, i.e. the coefficients are mostly negative and decrease fastly with the distance corresponding to the bins, there are also deviations. There are occasional short interatomic distances in the input vectors which result in coefficients based on few if not just one count. These will have little consequences because these cases, due to inaccuracies in the determined structure, in the added hydrogens or to the natural variance of bond lengths and interatomic distances, are extremely rare. There are, however, also positive values which are determined by a large number of counts, which reflect the way the linear model fits the extensive correlation of the effect of neighbouring atoms. This is well seen for the bins corresponding to short distances. With increasing distances the number of counts increases and all effects are averaged.
It is reassuring that the coefficients for larger distances are well fitted by the fourth inverse power of the distance, showing that the approach recovers the expected screening dependence on the distance.
The correlation among coefficients for different elements (both screened and screening) is in general poor, but it greatly increases for screening atoms when only larger distances (say greater than 6.0 Å) are taken into consideration.
The results are reported in Table 1. The correlation coefficients are for all elements over 0.90, and the root mean square error (RMSE) is less than 0.1 for all elements. These figures appear significantly better than for the reference method (Table 2). The implementation in NAMD might slightly

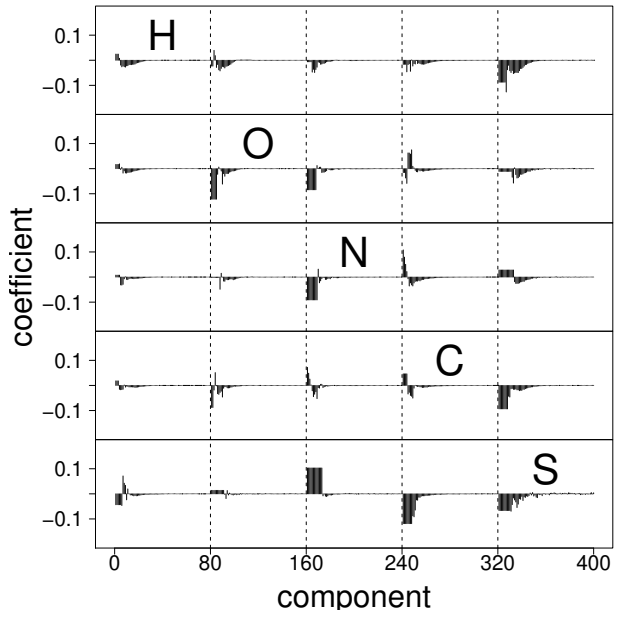

Fig. 3. Coefficients for the five 80-bin histograms corresponding to predicted H, O, N, C and S radii (on each row). The coefficients refer, in groups of 80 coefficients, to neighbouring H, O, N, C and S atoms.

Table 1. Summary of the linear model

	n. train	cc train	RMSE train	n. test	cc test	RMSE test
H	837537	0.93	0.067	16031	0.94	0.093
O	162615	0.94	0.051	3246	0.95	0.059
N	139732	0.94	0.037	2812	0.95	0.046
C	545894	0.94	0.036	10252	0.94	0.046
S	3976	0.94	0.039	66	0.91	0.054
all	1689754	0.94	0.055	32407	0.94	0.074

Correlation coefficients (cc) and root mean square error (RMSE) are computed on the inverse of the Generalized Born radii

differ from the original paper, so the performance of the original method could be slightly better than the one reported here.
The results are plotted in Figure 4 where the quality of the predicted values may be appreciated. It is apparent that for all elements there is a region, just below the inverse of the van der Waal radius where the predicted values tend to be lower (i.e. the predicted radii are larger) than the "perfect" values. This is attributable to crevices in the molecular structures where the binning is not able to catch the details of the histogrammed distribution of distances between atoms. This was confirmed visually on few tens of the largest deviations between predictions and reference values.
Another apparent feature in Figure 4 is that estimated inverse GB radii tend to be larger than the inverse "perfect" GB radii, at large values of the latters. This may be attributed to the CFA approximation, indeed this feature is absent for the GBR6 estimation of GB radii.

Generalized Born radii by neural networks

The neural network model, used to predict the inverse of the GB radius from the counts of neighbors, builds all correlations due to chemical structure in the parameters of the network and goes obviously beyond the linear model employed in the previous subsection.
The complexity of the connections in the network (Figure 2) should be in

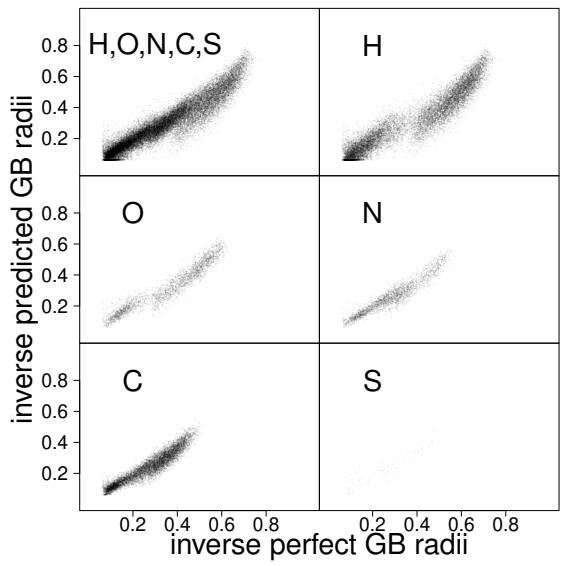

Fig. 4. The inverse of the GB radii predicted using multilinear regression versus the inverse of "perfect" GB radii. The predictions for different elements are grouped together in the top left panel and shown ungrouped in the other panels, as indicated by the element symbol in each panel.

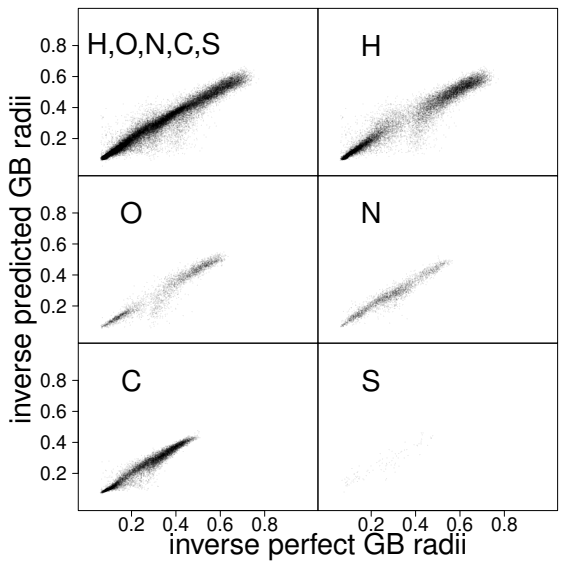

Fig. 5. The inverse of the GB radii predicted using the neural network versus the inverse of "perfect" GB radii. The predictions for different elements are grouped together in the top left panel and shown ungrouped in the other panels, as indicated by the element symbol in each panel.

Table 2. Summary of the results obtained by the OBC model

	n. test	cc test	RMSE test
H	16031	0.83	0.146
O	3246	0.94	0.096
N	2812	0.88	0.057
C	10252	0.75	0.076
S	66	0.86	0.126
all	32407	0.80	0.117

Correlation coefficients (cc) and root mean square error (RMSE) are computed on the inverse of the Generalized Born radii

principle able to describe accurately relations of the input data with the quantities to be predicted. The architecture of the network was chosen by trial for optimal performance. This should provide a limit to the predictability of the output based on the information provided by the input. The results are reported in Table 3. As expected, the neural network outperforms the linear model both in terms of correlation coefficients and RMSE with respect to the true values. The better performance of the neural network is apparent when the inverse of the GB radii are plotted versus the inverse of the "perfect" GB radii (Figure 5). For both the multilinear regression and the neural network models the RMSE for test vs. train dataset is much larger for hydrogen atoms. Largest errors are found close to surface crevices. Since in both methods it is the mean square errors which is minimized, parameters are tuned as to reduce large errors, adapting in a sense the parameters to the training set. Being more flexible, neural networks largely improve predictions. The possibility that this is due to overfitting is ruled out because this is observed for the test set of proteins which are completely independent of those used for training the network. Hydrogen atoms, endowed with the smaller radii, are most sensitive to the problems at crevices. Using larger van der Waals radii greatly hampers this problem resulting in smaller errors. This is observed for instance when testing the multinear regression model with the radii set used by FACTS which results in lower errors. Here we focused on the Bondi radii set which is widely used and, coupled with few modifications with the GB-Neck2 model, resulted in succesful folding of small proteins (Nguyen *et al.*, 2014). It is worth noting that there are also sets of radii optimized to reproduce solvation forces which are definitely larger than those of the Bondi sets of radii (Swanson *et al.*, 2005, 2007).

Table 3. Summary of the neural network model

	n. train	cc train	RMSE train	n. test	cc test	RMSE test
H	837537	0.97	0.047	16031	0.97	0.078
O	162615	0.97	0.042	3246	0.97	0.045
N	139732	0.96	0.039	2812	0.96	0.038
C	545894	0.97	0.038	10252	0.97	0.034
S	3976	0.92	0.058	66	0.92	0.047
all	1689754	0.97	0.044	32407	0.97	0.060

Correlation coefficients (cc) and root mean square error (RMSE) are computed on the inverse of the Generalized Born radii using neural network model

Accuracy and efficiency for MD simulations

The computation of GB radii and forces has been implemented in one of the latest versions of GROMACS still maintaining the feature of GBSA simulations.

The first concern was that the computed forces were not suffering from the discretization of distances in finite intervals. To this end we performed the same tests used to asses the accuracy of forces according to various GB models (Fogolari *et al.*, 2015), and compared computed forces with the reference Poisson-Boltzmann forces.

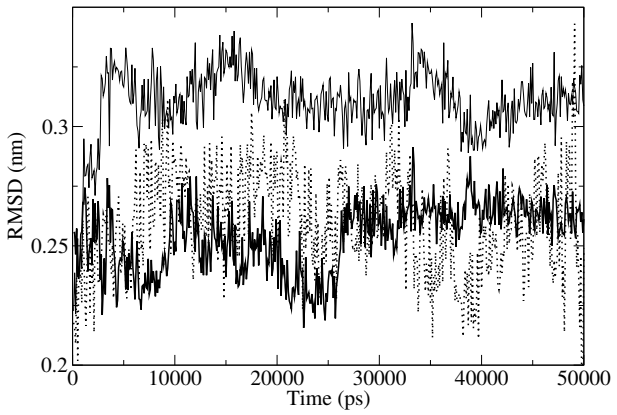

Fig. 6. Root mean square deviation from starting structure for the linear GB model MD simulation (thick continuous line), for the GB OBC model (thick dotted line) and for a vacuum simulation (upper thin continuous line).

The mean absolute error on each component of the solvation force by the linear GB model for 32407 atoms in the 18 protein structures of the Tjong and Zhou test set (non-dependent on the training set) is 0.73 kcal/(mol Å^2) vs. 0.78 kcal/(mol Å^2) of the reference OBC GB model. As we noted previously (Fogolari *et al.*, 2015) the GB OBC model estimation of solvation forces is fairly accurate.

The computation time in the absence of cutoff scales exactly quadratically with the number of atoms of the system.

In order to have a useful assessment of the linear GB model efficiency for MD simulations, we performed first a short MD simulation of the protein barnase (110 residues) where GB radii and derivatives were recalculated at each MD step. We performed also a reference vacuum simulation. The simulation is short (50 ps) so that the simulated structure is approximately the same for all three simulations, because the conformation impacts the computational time. The computational time due to the GB part of the simulation is obtained by taking the difference of the CPU time of the GB simulations and subtracting the CPU time of the vacuum simulation. In this way we obtain the relative efficiency of the linear GB model with respect to the GB OBC model, as far as GB radii and derivatives computation is concerned, which is 8.1. The linear GB model in this setting takes 2.3 times the time of a vacuum simulation.

Since recalculation of GB radii and derivatives is typically performed only every 20fs (in our setting every 10 2fs steps), we performed the same simulation under these conditions.

The relative efficiency of the linear GB model over the GB OBC model was reduced, as expected, to just 1.9, whereas the running time of the linear model compared to the vacuum running time is almost the same, i.e. 2.2. The latter fact confirms that the linear model is extremely efficient.

It must be noted however that running time includes, for all GB algorithms, computations which do not involve calculation of GB radii and their derivatives. It is these computations that slow simulation roughly by a factor 2 compared to vacuum calculations (which is still more efficient than simulating surrounding waters for specific systems).

As a last test of accuracy we simulated the protein barnase for 50 ns with the linear GB model and the GB OBC model. We compared the RMSD of the protein from the starting energy minimized structure. Inaccuracies in estimated forces would likely result in large deviations of the structure and/or in absence of fluctuations. The results reported in Figure 6 show the the linear GB model is able to produce the typical fluctuations observed in implicit solvent simulations and maintain the structure close to the native starting one.

3 Conclusions

Here we presented the application of multilinear regression and neural networks model for computation of Generalized Born radii. The input for both models consists in a categorical variable (the element of the atom to be predicted) and a five 80-component vectors (one for each element H, C, N, O and S) with counts of neighbours.

Multilinear regression improves significantly the accuracy of the computation, with respect to "perfect" GB radii, in comparison with one of the currently most used models. At the expenses of more computation, the neural network is able to further improving the performance highlighting perhaps the limits attainable with a simple input like the one chosen.

The multilinear regression model is suited also for force calculation, as the derivatives of the GB radii can be easily computed by finite differences (and interpolation), whereas the non linearity of the relation between input and output prevents such calculation for the neural network model.

We have compared the forces computed in this way with the reference Poisson-Boltzmann forces showing that the accuracy is slightly better than one of the most used methods.

We have tested GB radii and their derivative calculation using the multinear regression by an ad hoc implementation in one of the most popular MD simulation programs. Based on the simulation parameters the efficiency is 2 to 8 times better than the reference GB method.

Overall the two models described here, which are available as supplementary material, provide an alternative to approaches based on the physics of solvation, optimized to reproduce accurately GB radii, and can thus be used for implementing fast calculation of GB radii and their derivatives and provide a useful reference for other alternative methods.

Funding

This work was partly supported by the University of Udine through PRID2017 (Project PRONANO).

References

Aguilar, B., Shadrach, R., and Onufriev, A. V. (2010). Reducing the Secondary Structure Bias in the Generalized Born Model via R6 Effective Radii. *J Chem Theory Comput*, **6**, 3613–3630.

Bardhan, J. P. (2008). Interpreting the Coulomb-field approximation for generalized-Born electrostatics using boundary-integral equation theory. *J Chem Phys*, **129**, 144105.

Bashford, D. and Case, D. A. (2000). Generalized Born models of macromolecular solvation effects. *Annu Rev Phys Chem*, **51**, 129–152.

Behler, J. (2016). Perspective: Machine learning potentials for atomistic simulations. *J. Chem. Phys.*, **145**(17), 170901.

Berendsen, H. J. C., van der Spoel, D., and van Drunen, R. (1995). Gromacs: A message-passing parallel molecular dynamics implementation. *Comp. Phys. Comm.*, **91**, 43–56.

Bondi, A. (1964). van der Waals volumes and radii. *J. Phys. Chem.*, **68**, 441–451.

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., and Karplus, M. (2009). CHARMM: the biomolecular simulation program. *J Comput Chem*, **30**, 1545–1614.

Constanciel, R. and Contreras, R. (1984). Self consistent field theory of solvent effects representation by continuum models: Introduction of

desolvation contribution. *Theor. Chim. Acta*, **65**, 1–11.

David, L., Luo, R., and Gilson, M. K. (2000). Comparison of generalized Born and Poisson models: energetics and dynamics of HIV protease. *J. Comput. Chem.*, **21**, 295–309.

Fogolari, F., Brigo, A., and Molinari, H. (2002). The Poisson-Boltzmann equation for biomolecular electrostatics: a tool for structural biology. *J. Mol. Recogn.*, **15**, 377–392.

Fogolari, F., Brigo, A., and Molinari, H. (2003). Protocol for MM/PBSA molecular dynamics simulations of proteins. *Biophys. J.*, **85**, 159–166.

Fogolari, F., Corazza, A., Yarra, V., Jalaru, A., Viglino, P., and Esposito, G. (2012). Bluues: a program for the analysis of the electrostatic properties of proteins based on generalized Born radii. *BMC Bioinformatics*, **13 Suppl 4**, S18.

Fogolari, F., Corazza, A., and Esposito, G. (2013). Generalized Born forces: surface integral formulation. *J Chem Phys*, **138**, 054112.

Fogolari, F., Corazza, A., and Esposito, G. (2015). The accuracy of generalized born forces. In W. Rocchia and M. Spagnuolo, editors, *Computational Electrostatics for Biological Applications: Geometric and Numerical Approaches to the Description of Electrostatic Interaction Between Macromolecules*, pages 143–155. Springer International Publishing, Cham.

Fogolari, F., Corazza, A., and Esposito, G. (2018). Free energy, enthalpy and entropy from implicit solvent end-point simulations. *Front. Mol. Biosci.*, **5**, 11.

Ghosh, A., Rapp, C. S., and Friesner, R. A. (1998). Generalized Born model based on a surface integral formulation. *J. Phys. Chem. B*, **102**, 10983–10990.

Gilson, M. K., McCammon, J. A., and Madura, J. D. (1995). Molecular dynamics simulation with a continuum electrostatic model of the solvent. *J. Comput. Chem.*, **16**, 1081–1095.

Grycuk, T. (2003). Deficiency of the Coulomb-field approximation in the generalized Born model: An improved formula for Born radii evaluation. *J. Chem. Phys.*, **119**, 4817–4826.

Haberthur, U. and Caflisch, A. (2008). FACTS: Fast analytical continuum treatment of solvation. *J Comput Chem*, **29**, 701–715.

Hawkins, G. D., Cramer, C. J., and Truhlar, D. G. (1995). Pairwise solute descreening of solute charges from a dielectric medium. *Chem. Phys. Lett.*, **246**, 122–129.

Hawkins, G. D., Cramer, C. J., and Truhlar, D. G. (1996). Parametrized models of aqueous free energies of solvation based on pairwise descreening of solute atomic charges from a dielectric medium. *J Phys Chem*, **100**, 19824–19839.

Hinton, G., Srivastava, N., and Swersky, K. (2012). [Coursera] Neural networks for machine learning lecture 6.

Izadi, S., Harris, R. C., Fenley, M. O., and Onufriev, A. V. (2018). Accuracy Comparison of Generalized Born Models in the Calculation of Electrostatic Binding Free Energies. *J Chem Theory Comput*, **14**, 1656–1670.

Kale, L., Skeel, R., Bhandarkar, M., Brunner, R., Gursoy, A., Krawetz, N., Phillips, J., Shinozaki, A., Varadarajan, K., and Schulten, K. (1999). NAMD2: greater scalability for parallel molecular dynamics. *J. Comp. Phys.*, **151**, 283–312.

Lovell, S. C., Davis, I. W., Arendall, W. B., de Bakker, P. I., Word, J. M., Prisant, M. G., Richardson, J. S., and Richardson, D. C. (2003). Structure validation by C_α geometry: ϕ, ψ and C_β deviation. *Proteins*, **50**, 437–450.

Mongan, J., Svrcek-Seiler, W. A., and Onufriev, A. (2007a). Analysis of integral expressions for effective Born radii. *J Chem Phys*, **127**, 185101.

Mongan, J., Simmerling, C., McCammon, J. A., Case, D. A., and Onufriev, A. (2007b). Generalized Born model with a simple robust molecular volume correctio n. *J. Chem. Theory Comp.*, **3**, 156–169.

Mongan, J., Simmerling, C., McCammon, J. A., Case, D. A., and Onufriev, A. (2007c). Generalized Born model with a simple robust molecular volume correction. *J. Chem. Theory Comp.*, **3**, 156–169.

Moritz, S. and Bartz-Beielstein, T. (2017). imputeTS: Time Series Missing Value Imputation in R. *R Journal*, **9**, 207–218.

Nair, V. and Hinton, G. E. (2010). Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 807–814.

Nguyen, H., Maier, J., Huang, H., Perrone, V., and Simmerling, C. (2014). Folding simulations for proteins with diverse topologies are accessible in days with a single physics-based force field and implicit solvent. *J. Am. Chem. Soc.*, **136**, 13959–13962.

Niedermeier, C. and Schulten, K. (1992). Molecular dynamics simulations in heterogeneous dielectrics and Debye-Huckel media: application to the protein bovine pancreatic trypsin inhibitor. *Mol. Simul.*, **8**, 361–387.

Onufriev, A., Case, D. A., and Bashford, D. (2002). Effective Born radii in the generalized Born approximation: the importance of being perfect. *J Comput Chem*, **23**, 1297–1304.

Onufriev, A., Bashford, D., and Case, D. A. (2004). Exploring protein native states and large-scale conformational changes with a modified generalized Born model. *Proteins: Struct, Func, Gen*, **55**, 383–394.

Onufriev, A. V. and Case, D. A. (2019). Generalized Born implicit solvent models for biomolecules. *Annu Rev Biophys*, **48**, 275–296.

Prechelt, L. (1998). Early stopping-but when? In G. Montavon, G. Orr, and K.-R. Mueller, editors, *Neural Networks: Tricks of the trade.*, pages 55–69. Springer, Berlin.

Qiu, D., Shenkin, P., Hollinger, F., and Still, W. (1997). The GB/SA continuum model for solvation. a fast analytical method for the calculation of approximate Born radii. *J. Phys. Chem.*, **101**, 3005–3014.

Sanner, M., Spehner, J.-C., and Olson, A. (1996). Reduced surface: an efficient way to compute molecular surfaces. *Biopolymers*, **38**, 305–320.

Sharp, K. A. (1991). Incorporating solvent and ion screening into molecular dynamics using the finite-difference Poisson-Boltzmann method. *J. Comput. Chem.*, **12**, 454–468.

Sibi, P., Jones, S. A., and Siddarth, P. (2013). Analysis of different activation functions using back propagation neural networks. *J. Theor. Appl. Inform. Technol.*, **47**, 1264–1268.

Still, W. C., Tempczyk, A., Hawley, R. C., and Hendrickson, T. (1990). Semianalytical treatment of solvation for molecular mechanics and dynamics. *J. Am. Chem. Soc.*, **112**, 6127–6129.

Swanson, J. M. J., Adcock, S. A., and McCammon, J. A. (2005). Optimized radii for Poisson–Boltzmann calculations with the AMBER force field. *J. Chem. Theory Comp.*, **1**, 484–493.

Swanson, J. M. J., Wagoner, J. A., Baker, N. A., and McCammon, J. A. (2007). Optimizing the Poisson dielectric boundary with explicit solvent forces and energies: lessons learned with atom-centered dielectric functions. *J. Chem. Theory Comp.*, **3**, 170–183.

Tjong, H. and Zhou, H. X. (2007). GBr^6: a parametrization free, accurate, analytical generalized Born method. *J. Phys. Chem.*, **111**, 3055–3061.

Van Der Spoel, D., Lindahl, E., Hess, B., Groenhof, G., Mark, A. E., and Berendsen, H. J. C. (2005). Gromacs: Fast, flexible, and free. *J. Comp. Chem.*, **26**, 1701–1718.

Wang, J., Olsson, S., Wehmeyer, C., Perez, A., Charron, N. E., de Fabritiis, G., Noe, F., and Clementi, C. (2019). Machine learning of coarse-grained molecular dynamics force fields. *ACS Central Science*, **5**, 755–767.

Whittaker, E. T. (1910). *A history of the theories of aether and electricity: from the age of Descartes to the close of the nineteenth century*. Longmans, Green and Co., London, UK.

Zhang, L., Han, J., Wang, H., Car, R., and E, W. (2018). Deepcg: Constructing coarse-grained models via deep neural networks. *J. Chem. Phys.*, **149**, 034101.