Comparative Molecular Field Analysis using Molecular Integral Equation Theory

Samiul M. Ansari and David S. Palmer*

Department of Pure and Chemistry, University of Strathclyde, Thomas Graham Building, 295 Cathedral St, Glasgow, Scotland G1 1XL, UK

E-mail: david.palmer@strath.ac.uk

Phone: (+44)141 548 4178

Abstract

Recently, Güssregen et al. used solute–solvent distribution functions calculated by the 3D Reference Interaction Site Model (3DRISM) in a 3D-QSAR model to predict the binding affinities of serine protease inhibitors; this approach was referred to as Comparative Analysis of 3D RISM MAps (CARMa). [*J. Chem. Inf. Model.*, **2017**, *57*, 1652-1666] Here we extend this idea by introducing *probe atoms* into the 3DRISM solvent model in order to directly capture other molecular interactions in addition to those related to hydration/dehydration. Benchmark results for six different protein-ligand systems show that CARMa models trained on probe atom descriptors gives consistently more accurate predictions than CoMFA, and other common QSAR approaches.

Introduction

The premise of quantitative structure-activity relationships (QSAR) is that a compound's molecular structure can be used to determine its macroscopic properties, such as binding

affinity and pIC_{50}. A QSAR is derived by using experimental data to learn a statistical relationship between the physical property of interest (e.g. pIC_{50}) and molecular descriptors calculable from a simple computational representation of the molecule. The QSAR must accurately model the training data and generalize to correctly predict activities for molecules outside the representative training set.[1] A large number of QSAR methods have been described in the literature using various classes of descriptors. For the prediction of physicochemical properties, 1D and 2D descriptors that can be calculated quickly without knowledge of molecular conformation are often considered to be satisfactory (e.g. counts of functional groups, graph indices, etc)[2,3] However, for modelling protein-ligand systems, where ligand conformation influences the strength of binding interactions, 3D (or 4D) descriptors are usually preferred.[1,4–7]

One of the most widely used 3D-QSAR methods is Comparative Molecular Field Analysis (CoMFA), which was proposed by Cramer *et al.* in 1988.[7] CoMFA establishes a uniform grid encompassing a series of pre-aligned molecules. Electrostatic and Lennard-Jones potential energies are then calculated between a positively charged carbon atom probe, located at each vertex of the grid, and each of the molecules embedded within.[7] The resulting electrostatic and "steric" fields are used as input for partial-least-squares regression models. Since its first publication, CoMFA has been cited in thousands of published articles and used in numerous drug discovery programs.[8,9] Several extensions to the CoMFA methodology have been proposed, of which the highest profile is comparative molecular similarity indices analysis (CoMSIA).[10,11]

Although CoMFA is widely used, it relies on a relatively simple representation of molecular interactions, which does not explicitly account for solvation/desolvation effects that can dramatically influence protein-ligand binding. Since CoMFA was first proposed, advances in theory, algorithms and computer power mean that there are now many fast and accurate methods to model molecular solvation effects. Some success has been acheived using numerical simulation (e.g. Monte Carlo or molecular dynamics simulations) to compute solute-

solvent descriptors for QSAR models,[12] but such methods are computationally expensive and subject to sampling errors that reduce the signal-to-noise ratio in the modelling dataset. Integral equation theory approaches are of particular interest for QSAR modelling because they allow solute-solvent distributions and solvation thermodynamics to be computed at a fraction of the cost of explicit solvent numerical simulations and with no sampling error.[13–15] The most widely used of these methods are the 1D and 3D Reference Interaction Site Models proposed by Chandler et al.[16] and Beglov and Roux,[17–19] respectively. Accurate predictions of hydration free energy and Caco-2 permeability have previously been reported using QSAR models based on 1D RISM molecular descriptors.[20] Recently, Güssregen et al. proposed the Comparative Analysis of 3D RISM Maps (CARMa) methodology, which uses solute–solvent distribution functions calculated by 3DRISM to replace the electrostatic or steric fields in CoMFA.[21] This approach was shown to give accurate predictions of binding affinities for a series of serine protease inhibitors, but tests on other systems have not yet been published.[21]

The purpose of this article is two-fold. Firstly, we propose an extension to the CARMa methodology. CARMa uses a statistical mechanics solvent model to capture solvation effects, but does not directly model the electrostatic and steric effects probed by CoMFA. Solving the 3D RISM equations for a solvent comprising CoMFA probes in aqueous solution addresses this issue and results in predictions that are more accurate than either CoMFA or the original CARMa model; in what follows, this approach is referred to as CARMa(electrolyte) whenever a distinction needs to be made with the original CARMa method. Secondly, we provide an extensive benchmark of both CARMa and CARMa(electrolyte) models over six different protein-ligand systems and compare their accuracy to previously published CoMFA and 3D-QSAR results. The influence of 3DRISM algorithmic parameters, such as the 3D RISM bridge-functional and grid-size, on the prediction accuracy are systematically investigated.

Theory

The method proposed here uses density distribution functions calculated by the 3D reference interaction site model (3D-RISM) as input. We begin with a brief description of the relevant parts of the standard 3D RISM theory before outlining our approach to QSAR predictions.

3D-RISM

3D-RISM[19,22–24] is a theoretical method for modelling solution phase systems based on classical statistical mechanics. The 3D-RISM equations relate 3D intermolecular *solvent site - solute* total correlation functions ($h_\alpha(\mathbf{r})$), and direct correlation functions ($c_\alpha(\mathbf{r})$) (index α corresponds to the solvent sites):[19,24]

$$h_\alpha(\mathbf{r}) = \sum_{\xi=1}^{N_{solvent}} \int_{R^3} c_\xi(\mathbf{r}-\mathbf{r}')\chi_{\xi\alpha}(|\mathbf{r}'|)d\mathbf{r}', \tag{1}$$

where $\chi_{\xi\alpha}(r)$ is the bulk solvent susceptibility function, and $N_{solvent}$ is the number of sites in a solvent molecule (see 1). The solvent susceptibility function $\chi_{\xi\alpha}(r)$ describes the mutual correlations of sites ξ and α in solvent molecules in the bulk solvent. It can be obtained from the solvent intramolecular correlation function ($\omega^{solv}_{\xi\alpha}(r)$), site-site radial total correlation functions ($h^{solv}_{\xi\alpha}(r)$) and the solvent site number density (ρ_α): $\chi_{\xi\alpha}(r) = \omega^{solv}_{\xi\alpha}(r) + \rho_\alpha h^{solv}_{\xi\alpha}(r)$ (from here onwards we imply that each site is unique in the molecule, so that $\rho_\alpha = \rho$ for all α).[24] In this work, these functions were obtained by solution of the RISM equations of the solvent.[24,25]

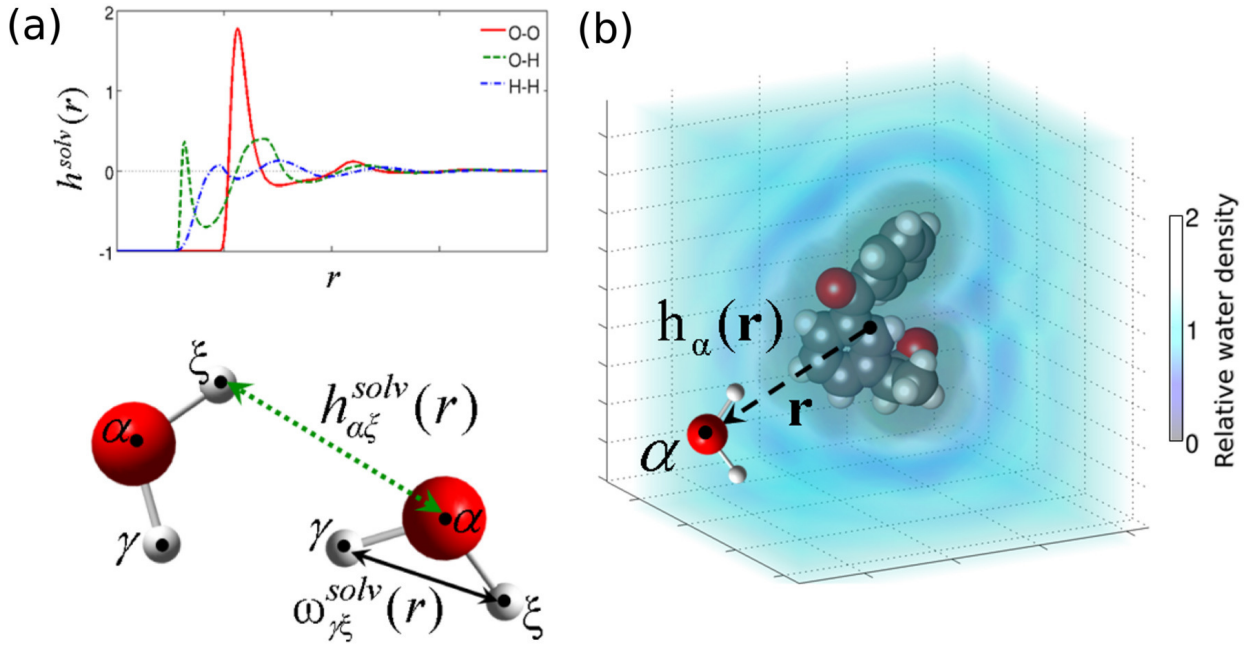


Figure 1: Correlation functions in the 3D-RISM approach. (a) Site-site intramolecular ($\omega^{solv}_{\gamma\xi}(r)$) and intermolecular ($h^{solv}_{\alpha\xi}(r)$) correlation functions between sites of solvent molecules. The graph shows the radial projections of water solvent site-site density correlation functions: oxygen-oxygen (OO, red solid), oxygen-hydrogen (OH, green dashed) and hydrogen-hydrogen (HH, blue dash-dotted); (b) Three-dimensional intermolecular solute-solvent correlation function $h_\alpha(\mathbf{r})$ around a model solute (diclofenac). This figure is based on Figure 1 from our earlier work.[14]

In order to calculate $h_\alpha(\mathbf{r})$ and $c_\alpha(\mathbf{r})$, $N_{solvent}$ approximate *closure* relations must be introduced. Here two forms of closure relationship were tested: the Kovalenko and Hirata (KH) closure, which is also referred to as the partial series expansion order 1 (PSE-1), or partial-linearised hypernetted chain (PLHNC) closure;[26] the PSE-3 closure.[27] The KH closure is:

$$h_\alpha(\mathbf{r}) = \begin{cases} \exp(\Xi_\alpha(\mathbf{r})) - 1 & when \quad \Xi_\alpha(\mathbf{r}) \leq 0 \\ \Xi_\alpha(\mathbf{r}) & when \quad \Xi_\alpha(\mathbf{r}) > 0 \end{cases} \tag{2}$$

where $\Xi_\alpha(\mathbf{r}) = -\beta u_\alpha(\mathbf{r}) + h_\alpha(\mathbf{r}) - c_\alpha(\mathbf{r})$, $u_\alpha(\mathbf{r})$ is the 3D interaction potential between the solute molecule and α solvent site, $\beta = 1/k_B T$, k_B is the Boltzmann constant, and T is the temperature. The PSE-3 closure is:[27,28]

$$h_\alpha(\mathbf{r}) = \begin{cases} \exp(\Xi_\alpha(\mathbf{r})) - 1 & when \quad \Xi_\alpha(\mathbf{r}) \leq 0 \\ \sum_{i=0}^{n} (\Xi_\alpha(\mathbf{r}))^i / i! - 1 & when \quad \Xi_\alpha(\mathbf{r}) > 0 \end{cases} \tag{3}$$

The 3D interaction potential between the solute molecule and α site of solvent ($u_\alpha(\mathbf{r})$, Equation 2) is estimated as a superposition of the site-site interaction potentials between solute sites and the particular solvent site, which depend only on the absolute distance between the two sites. We use the common form of the site-site interaction potential represented by the long-range electrostatic interaction term and the short-range term (Lennard-Jones potential).[29] 3DRISM distribution functions computed using the PSE-3 closure have been labelled in the following text with a superscript, e.g. $g_O^{PSE-3}(r)$; all other calculations were performed using the KH closure.

3D-RISM–CARMa

Two different classes of functions were tested as input to CARMa analyses: *solvent density distribution functions*, $g(r)$, which represent the local solvent density at grid points around the solute; *solvation free energy density functions*, which indicate the local contribution to the excess chemical potential of the solute (further details below).

Solvent Density Distribution Functions

Solving the 3D RISM equations gives a solvent density distribution function, $g(r)$, for each interaction site (atom) in the solvent. For water, the $g(r)$ functions are identical for the two hydrogen atoms because of molecular symmetry. Four different $g(r)$ functions were tested as input to CARMa: (i) *water density distribution functions*, $g_O(r)$ or $g_H(r)$, computed for pure aqueous solvent; (ii) *solvent-probe density distribution functions*, $g_{C+}(r)$ or $g_{C-}(r)$, obtained by solving the 3DRISM equations with 0.1 M "C+" and 0.1 M "C-" probe atoms as co-solvents in aqueous solution. The C+ and C- probes are positively or negatively charged sp^3 carbon atoms with Lennard-Jones parameters taken from the general Amber forcefield

(GAFF).[30] $g_O(r)$ and $g_H(r)$ functions were also obtained from the simulations in 0.1 M C+/C- (aq), but these were not used in CARMa models because they were observed to be highly correlated with $g_O(r)$ and $g_H(r)$ functions computed in pure water, respectively (e.g. to two decimal places, $R = 1.00$ and $R = 1.00$, respectively, for the steroid dataset with grid-spacing of 3.O Å). By contrast the $g_{C+}(r)$ and $g_{C-}(r)$ functions are only weakly correlated with the $g_O r$,$g_H r$, and SFED functions computed in pure water ($0.01 \leq |R| < 0.6$ for the steroid dataset with grid-spacing of 3.O Å).

Solvent Free Energy Density

Within the framework of the RISM theory there exist several approximate functionals that allow one to analytically obtain values of the solvation free energy from the total $h_\alpha(\mathbf{r})$ and direct $c_\alpha(\mathbf{r})$ correlation functions.[27,31,32] These can be derived analytically from the appropriate 3DRISM closure relationship.

The PSE-3 free energy functional is given by:

$$\Delta G_{hyd}^{PSE-3} = \Delta G_{hyd}^{HNC} - k_B T \sum_{\alpha=1}^{N_{solvent}} \rho_\alpha \int_V \left[\Theta[h_\alpha(\mathbf{r})] \frac{\Xi_\alpha(\mathbf{r})^{n+1}}{(n+1)!} \right] d\mathbf{r} \tag{4}$$

where ρ_α is the number density of solvent sites α, Θ is a Heaviside step function, and ΔG_{hyd}^{HNC} is the solvation free energy calculated using the hypernetted-chain functional, which is given by:[33]

$$\Delta G_{hyd}^{HNC} = k_B T \sum_{\alpha=1}^{N_{solvent}} \rho_\alpha \int_V \left[\frac{1}{2} h_\alpha^2(\mathbf{r}) - \frac{1}{2} h_\alpha(\mathbf{r}) c_\alpha(\mathbf{r}) - c_\alpha(\mathbf{r}) \right] \tag{5}$$

The KH free energy functional is given by:

$$\Delta G_{hyd}^{KH} = k_B T \sum_{\alpha=1}^{N_{solvent}} \rho_\alpha \int_{R^3} \left[\frac{1}{2} h_\alpha^2(\mathbf{r}) \Theta(-h_\alpha(\mathbf{r})) - \frac{1}{2} h_\alpha(\mathbf{r}) c_\alpha(\mathbf{r}) - c_\alpha(\mathbf{r}) \right] d\mathbf{r} \tag{6}$$

where ρ_α is the number density of solvent sites α, and Θ is the Heaviside step function:

$$\Theta(x) = \left\{ \begin{array}{l} 1 \text{ for x}>0 \\ 0 \text{ for x}<0 \end{array} \right\} \tag{7}$$

Both the KH and PSE-3 solvation free energy functionals can be written in a compact form as:

$$\Delta G_{solv} = \int_0^\infty w(\boldsymbol{r}) dr \tag{8}$$

where the integrand functionals combine the N total and direct correlation functions of a single solute into a single function of $\boldsymbol{r}$, which we refer to as the solvation free energy density, SFED. Thus for the KH free energy functional, SFED is defined as:

$$w_{KH}(\mathbf{r}) = k_B T \sum_{\alpha=1}^{N_{solvent}} \rho_\alpha \left[\frac{1}{2} h_\alpha^2(\mathbf{r}) \Theta(-h_\alpha(\mathbf{r})) - \frac{1}{2} h_\alpha(\mathbf{r}) c_\alpha(\mathbf{r}) - c_\alpha(\mathbf{r}) \right] \tag{9}$$

As indicated by Equation 8, integrating $w_{KH}(\mathbf{r})$ over all space returns ΔG_{hyd}^{KH}. Therefore, the grid points in the function give the local spatial contribution to the total solvation free energy of the solute. One argument for using SFED as opposed to $g(\mathbf{r})$ in a QSAR model is that it inherently includes information from all of the 3D-RISM functions, i.e. for pure water, SFED is a composite of $g_H(\mathbf{r})$, $g_H(\mathbf{r})$,$c_H(\mathbf{r})$, and $c_O(\mathbf{r})$. Also, the SFED functions asymptotically approach zero at shorter solute-solvent distances than $g(r)$, which reduces the number of redundant descriptors in the CARMa analysis.

The SFED function computed in 0.1 M C+/C- (aq) was not used in CARMa models because it was observed to be highly correlated with the SFED function computed in pure water (e.g. $R = 1.00$ for the steroid dataset with grid-spacing of 3.O Å).

Grids

When the 3DRISM equations are solved numerically, both the local solvent density (as given by $g(\mathbf{r})$) and the solvation free energy density ($w(\mathbf{r})$) are represented on discrete grids. In

principle, the values of these functions at specific grid points could be used directly as input to the CARMa models. Since 3D-RISM calculations are normally carried out on a large grid with a relatively small grid spacing (0.3-0.5 Å), however, this would lead to many redundant variables making the numerical data sets too large to be processed easily. A simple solution would be to solve the 3D RISM calculations on a small and coarse grid, but this would reduce the accuracy of the obtained density distribution functions. Instead, in this study, all 3D-RISM calculations were performed on a large and fine grid (>50 $Å^3$ grid with a 0.5 Å spacing). The size of the grids used to represent the 3D-RISM distribution functions were then reduced to a standard size by removing layers of each grid face as appropriate (using custom Python scripts). To provide a further filter to remove some of the unnecessary variables, we tested two different approaches: (i) mapping the 3D-RISM results onto a coarser grid; (ii) selecting only those grid points that were within a distance, d, from the solute. The latter method increased computational expense without improving prediction accuracy and, therefore, is not discussed further. Prior to statistical modelling, we also removed all variables that had a variance of zero. Further variable selection was carried out using standard statistical methods (namely a genetic algorithm and Random Forest, as described later).

Statistical and Machine Learning Algorithms

To derive the predictive CARMa models, two different methods of regression were considered: Partial-Least-Squares (PLS) and Random Forest (RF). A genetic algorithm was also tested to select input variables for the PLS model.

Partial-Least-Squares Regression

Partial least squares (PLS) is a method for linear regression that has been widely used in many different fields of research, including chemistry, biology, econometrics and social science. The PLS algorithm finds a linear regression model by projecting both the dependent and independent variables into a new mathematical space in which the covariance in the data

structure can be explained by a small number of latent variables. As such PLS regression has some similarity to principal component regression, but the latent variables are selected for their ability to explain the variance in the dependent variable as well as in the independent variables. The algorithms used for PLS regression have been explained elsewhere.[34]

Genetic Algorithm

A genetic algorithm was used to select an optimal subset of descriptors for the PLS model. Genetic algorithms are commonly used to solve both constrained and unconstrained optimization problems using a selection approach based on biological evolution.[35] Here the genetic algorithm continuously modifies a population of chromosomes, in which each chromosome is a bit string that indicates whether each variable (grid point from 3DRISM distribution function) should be included or omitted from the PLS regression model.[36] The RMSE for 3-fold cross-validation was used as a fitness function to score each chromosome. Over successive generations, the population "evolves" toward an optimal solution.[37]

Random Forest

Random Forest is a method for classification and regression which was introduced by Breiman and Cutler.[38] The method is based upon an ensemble of regression trees, from which the prediction of a continuous variable is provided as the average of the predictions of all trees. Each tree is grown from a separate bootstrap sample of the training data using the CART algorithm.[38] During tree growth, the branches continue to be subdivided while the minimum number of observations in each leaf is greater than a predetermined value. The descriptor selected for branch splitting at any fork in any tree is not selected from the full set of possible descriptors but from a randomly selected subset of predetermined size. There are three possible training parameters for Random Forest: *ntree* - the number of trees in the Forest; *mtry* - the number of different variables tried at each split; and *nodesize* - the minimum node size below which leaves are not further subdivided. The bootstrap sample

used during tree growth is a random selection with replacement from the molecules in the dataset. The molecules that are not used for tree growth are termed the *out-of-bag* sample. Each tree provides a prediction for its out-of-bag sample, and the average of these results for all trees provides an in situ cross-validation called the out-of-bag validation.

Methods

QSAR Data Sets

Six datasets were selected to benchmark the CARMa predictions. Firstly, the 21 steroids selected by Cramer et al. were used to provide a direct comparison between CARMa and CoMFA.[7,9] Optimized and aligned structures for all 21 molecules were taken from Coates et al.;[9] these files resolve some errors in the way that the structures were reported by Cramer *et al.*[9] Secondly, five pIC_{50} data sets published by Sutherland et al. were used to compare CARMa to a wide-range of 3D-QSAR methods (including CoMFA, COMSIA, etc). The compounds with literature references, aligned molecular structures, and grid parameters for field based QSAR are all described by Sutherland *et al.*[1] Briefly, the datasets are: **ACE dataset** – 114 angiotensin converting enzyme (ACE) inhibitors with pIC_{50} values ranging between 2.1 - 9.9;[39] **AchE dataset** – 111 acetylcholinesterase (AchE) inhibitors with pIC_{50} values ranging between 4.3 - 9.5;[40] **BZR dataset** – 163 ligands for the benzodiazepine receptor (BZR) with pIC_{50} values ranging between 5.5 - 8.9;[41] **COX2 dataset** – 322 cyclooxygenase-2 (COX2) inhibitors with pIC_{50} values ranging between 4.0 - 9.0;[42] **DFHR dataset** – 397 dihydrofolate reductase (DHFR) inhibitors with pIC_{50} values ranging between 3.3 - 9.8.[43] Sutherland *et al* used a "cherry picking" with maximum dissimilarity algorithm to assign 33% of the dataset to the test set and the remaining compounds to the training set.[44,45] To allow a direct comparison with Sutherland's results, we have used the same aligned molecular conformations and the same training and test sets.

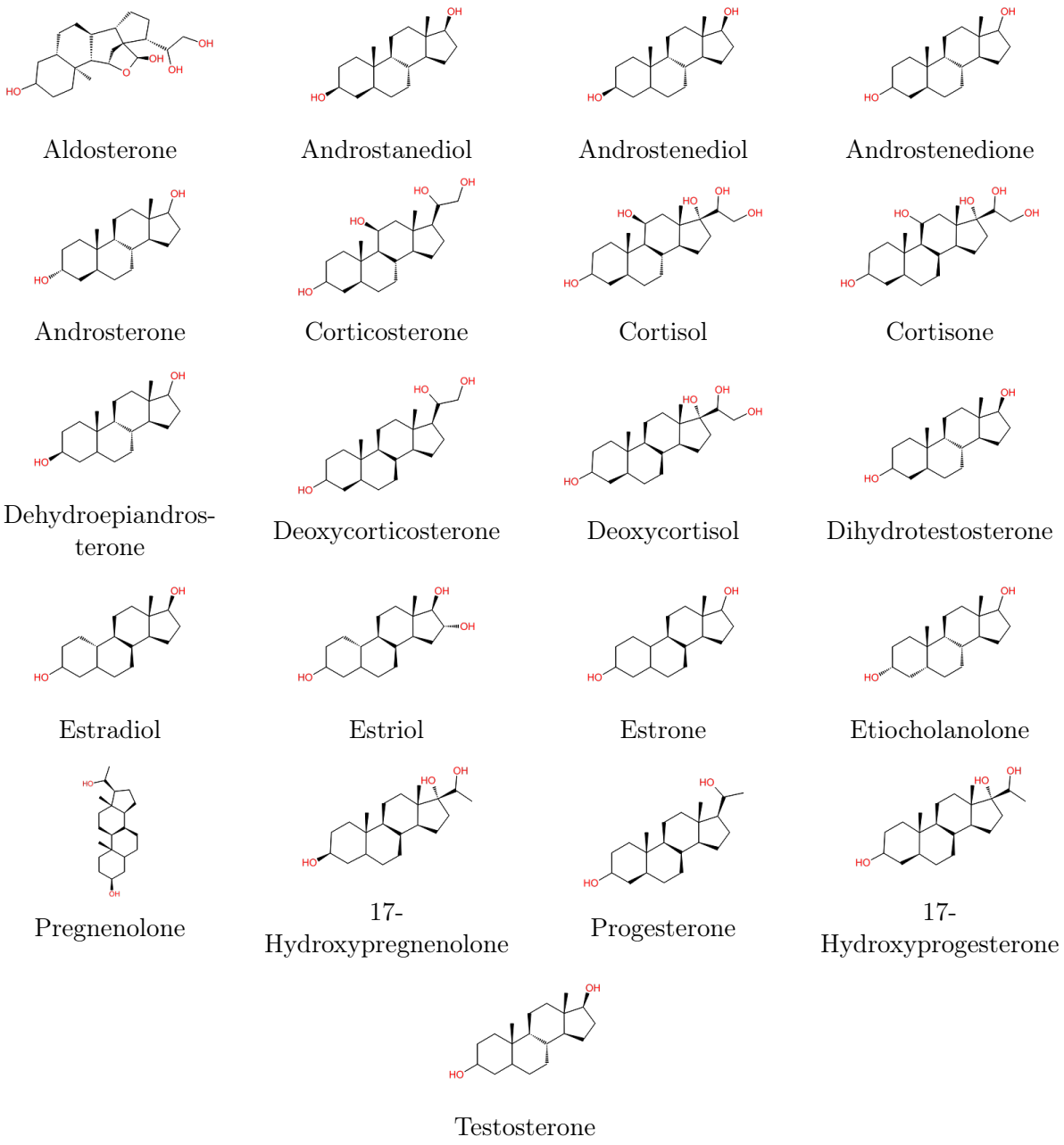


Figure 2: A depiction of steroids training set.

3D-RISM

The 3DRISM calculations were performed using AmberTools16.[46] Ligand structures were obtained from the articles by Coates et al.[9] and Sutherland et al.[1] and were used without modification. Lennard-Jones parameters and atomic partial charges for the ligands were

taken from the General Amber Force Field (AMBER-GAFF).[30] The KH closure was used for solution of the 3D-RISM equations unless otherwise stated. The linear grid spacing in each of the three directions was 0.5 Å. We employed the MDIIS iterative scheme,[47] where we used 5 MDIIS vectors, MDIIS step size of 0.7, residual tolerance of 10^{-10} in the L2 norm of the difference between the $g(r)$ functions for two subsequent solutions of 3D RISM iterations. All calculations were carried out at 298 K.

Solvent susceptibility functions required as input to 3D-RISM were calculated using dielectrically consistent 1D-RISM[48] with the KH closure. The grid size for 1D-functions was 0.025 Å, which gave a total of 16384 grid points. We employed the MDIIS iterative scheme, where we used 20 MDIIS vectors, MDIIS step size of 0.3, and residual tolerance of 10^{-12} in the L2 norm of the difference between the $g(r)$ functions for two subsequent solutions of RISM iterations. The solvent model was: (i) 0.1M C+/C- (aq) for the calculation of $g_{C+}(r)$ and $g_{C-}(r)$, or (ii) pure water for the calculation of all other distribution functions (e.g. $g_O(r)$, $g_O^{PSE-3}(r)$, $g_H(r)$, and SFED). As mentioned previously, we did not build CARMa models for the $g_O(r)$ or $g_H(r)$ functions obtained from the simulations in 0.1M C+/C- (aq) because each one was highly correlated with the same function calculated in pure water ($R = 1.00$ to two decimal places). We used the Lue and Blankschtein version of the SPC/E model of water (MSPC/E).[49] This differs from the original SPC/E water model[50] by the addition of modified Lennard-Jones (LJ) potential parameters for the water hydrogen, which were altered to prevent possible divergence of the algorithm.[51–54] The Lorentz-Berthelot mixing rules were used to generate the solute-water LJ potential parameters.[55] The following LJ parameters (for water hydrogen) were used to calculate the interactions between solute sites and water hydrogens: $\sigma_{H_w}^{LJ} = 1.1657$Å$ and $\epsilon_{H_w}^{LJ} = 0.0155$ kcal/mol.

CARMa

CARMa models were setup and trained using a combination of bespoke Python and R scripts.

Partial-Least Squares

Partial-Least Squares regression models were trained using the *pls* library[56] in the R statistical computing environment.[57] All PLS models were trained with 3 latent variables, which was selected as the optimal balance between model size and prediction accuracy based on consideration of the residual error sum of squares and the percentage of variance explained. It is also the same number of latent variables that was used in the original studies on CoMFA and CARMa methods.[7,21]

Random Forest

Random Forests were trained with the *randomForest* library[58] in the R statistical computing environment,[57] using standard parameters: $mtry = N/3$, $nodesize = 5$, and $ntree = 500$, where N is the number of input variables and $mtry$ is rounded down to the nearest integer. There is extensive evidence in the literature that the Random Forest algorithm is insensitive to training parameters,[59,60] so that variation of $mtry$ between 40 and N, of $ntree$ from 250 upward, and of $nodesize$ in the region 5 to 10 has little effect on prediction accuracy. As has been done previously, we use these standard Random Forest parameters without further optimization.[59,60]

Computational Expense

The CARMa calculations reported here were performed using a quad-core, 3.4GHz Intel Core i5 iMac desktop with 16GB RAM (late 2013, operating system version 10.12.2). The most time-consuming step in making a prediction with a pre-trained 3D-RISM–CARMa model is solving the 3D-RISM equations; the remaining steps require negligible computational expense. For the molecules considered here, none of the 3DRISM calculations took longer than 10 minutes.

Results

Steroid dataset

The steroids dataset consists of 21 compounds with corticosteroid-binding globulins (CBG) binding affinity data. Cramer et al. report a $q^2 = 0.734$ for leave-one-out cross-validation of a CoMFA model,[9] which represents a relatively accurately prediction of the CBG binding affinity data.

Table 1: Steroids leave-one-out cross-validation statistics (q^2) using CARMa with various descriptors and grid spacings.

Grid Spacing (Å)	$g_O(r)$	$g_O^{PSE3}(r)^a$	$g_H(r)$	*SFED*[b]	$g_{C-}(r)^c$	$g_{C+}(r)^d$	*CoMFA*
			PLS				
1.0	0.84	0.85	0.84	0.68	0.84	0.84	-
1.5	0.86	0.86	0.85	0.67	0.85	0.84	-
2.0	0.84	0.84	0.83	0.69	0.85	0.84	0.73
2.5	0.81	0.81	0.85	0.74	0.83	0.83	-
3.0	0.85	0.86	0.85	0.67	0.83	0.84	-

[a] Partial Series Expansion-3 closure; [b] Solvation Free Energy Density; [c] sp^3 Carbon probe atom with -1 charge; [d] sp^3 Carbon probe atom with +1 charge.

A total of 30 different PLS models were trained for the steroid dataset (5 different 3D RISM grid spacings × 6 different 3D RISM distribution functions). The complete set of statistics ($R^2, RMSE, \sigma, bias$ for training and cross-validation) are presented in the Supporting Information. Table 1 presents q^2 values for LOO-CV for all 30 PLS models. Several different trends are evident in Table 1. Firstly, the choice of bridge functional used to solve the 3D RISM equations (KH or PSE-3) does not significantly influence the results. The q^2 values for CARMa models built on $g_O^{KH}(r)$ or $g_O^{PSE3}(r)$ are nearly identical for all grid sizes. A similar conclusion was reached in previous work that used PLS models trained on 1D RISM descriptors to predict hydration free energy and Caco-2 permeability.[20] (Since converging the 3DRISM equations using the PSE-3 closure is occasionally problematic, the five datasets discussed next were modelled using the KH closure). Secondly, for this dataset, the PLS models trained on solvation density distributions ($g_O(r)$, $g_O^{PSE3}(r)$, $g_H(r)$, $g_{C-}(r)$ and

$g_{C+}(r)$) perform better than those trained on solvation free energy density (SFED). Thirdly, there is no obvious trend between the various grid spacings. Although finer grids might be expected to lead to more accurate models, this is not evident in the data, which suggests that some redundancy is present in the finer grids.

Figure 3 shows the cross-validated predictions obtained for PLS models trained on $g_O(r)$ distribution functions represented on a 2 Å grid; the same grid spacing used in the CoMFA models. The CARMa model explains more of the variance in the experimental data than the CoMFA model, as exemplified by $q^2 = 0.84$ for CARMa compared to $q^2 = 0.73$ for CoMFA. The residual cross-validated error in the CARMa model ($RMSE = 0.46$) is predominantly due to random error ($\sigma = 0.45$) with a relatively small systematic error ($bias = 0.09$).

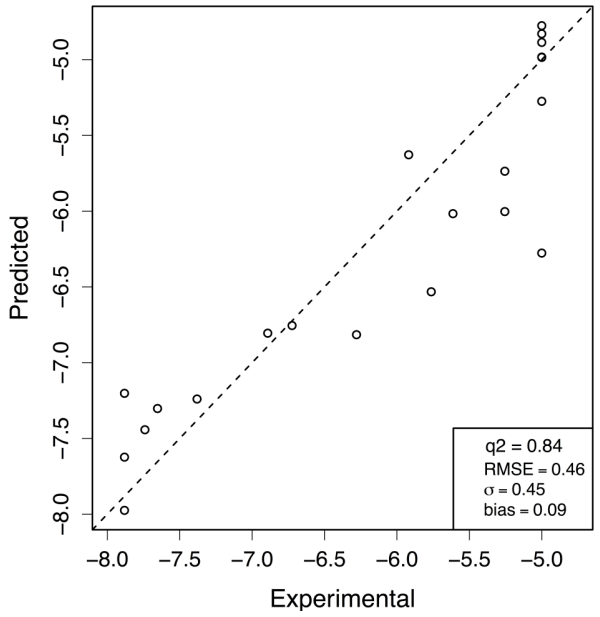
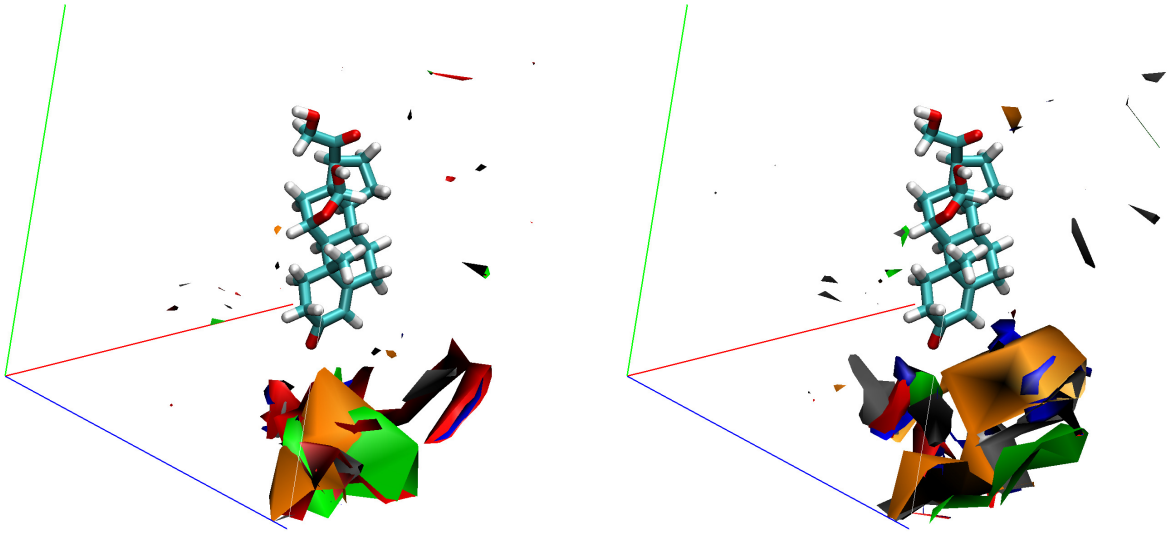


Figure 3: Correlation graphs of leave-one-out cross-validation (LOO-CV) for PLS models using the $g_O(r)$ distribution data at 2.0 Å grid spacing.

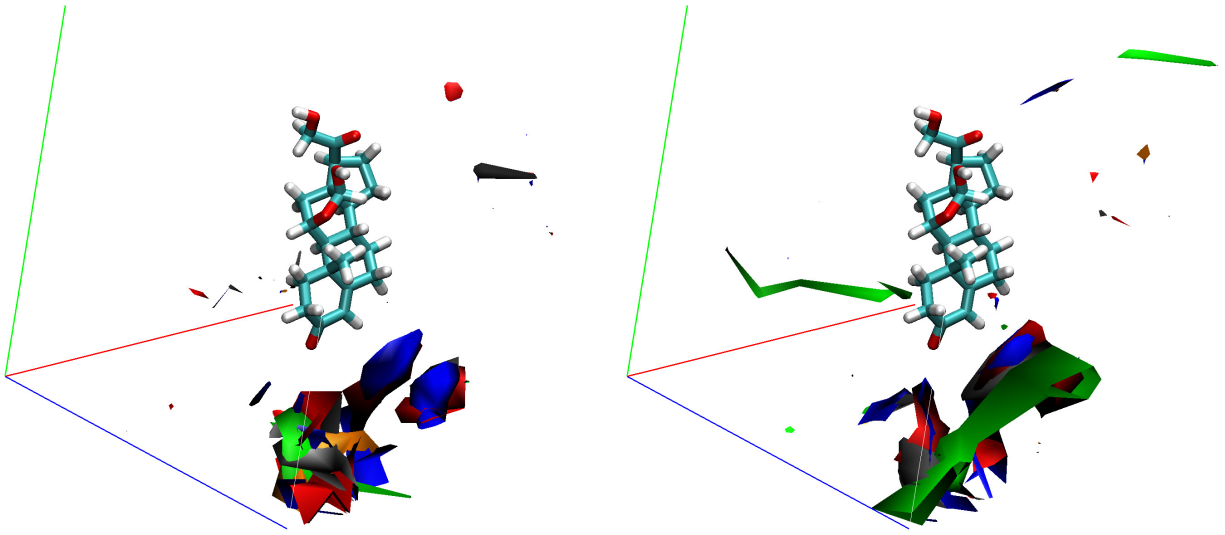
Importance



(a) $g_O(r)$ distribution importance. (b) $g_H(r)$ distribution importance.

Figure 4: Aldosterone is shown with PLS importance of $g_O(r)$ and $g_H(r)$ distributions at grid spacings 1.0 (*blue*), 1.5 (*red*), 2.0 (*grey*), 2.5 (*orange*) and 3.0 (*green*). The graphics show 10% of the most important regions for the PLS models.

The total contribution that each input variable made to the PLS latent variables was used as a metric to assess its importance to the model. Figure 4 depicts the most important 10% of the $g_O(r)$ and $g_H(r)$ functions as assessed from the PLS models. There is little difference between the $g_O(r)$ and $g_H(r)$ descriptor models, which is perhaps not surprising given that oxygen and hydrogen atoms are covalently bonded in water. In Figures 4a and 4b, the regions highlighted are located by the terminal cyclohexane (ring A) of the steroids for all grid spacings. A similar trend is observed in the importance graphics for the $g_{C-}(r)$ and $g_{C+}(r)$ probe atom distributions (see Figure 5), but here the distributions seem to be more localised in space in comparison to those for $g_O(r)$ and $g_H(r)$ (Figure 4).



(a) $g_{C-}(r)$ distribution importance. (b) $g_{C+}(r)$ distribution importance.

Figure 5: Aldosterone is shown with PLS importance of $g_{C-}(r)$ and $g_{C+}(r)$ distributions at grid spacings 1.0 (*blue*), 1.5 (*red*), 2.0 (*grey*), 2.5 (*orange*) and 3.0 (*green*). The graphics show 10% of the most important regions for the PLS models.

pIC_{50} Data Sets

To further validate the methodology, CARMa models were developed to predict pIC_{50} values for five datasets collated by Sutherland et al.[1] In each case, the training/testing datasets and aligned molecular structures selected by Sutherland et al. were used to provide a direct comparison to their CoMFA and 3D-QSAR results. Three different regression methods were considered: PLS, GA-PLS and RF. (The GA-PLS and RF algorithms were not used in the previous section because they can not be reliably trained on smaller datasets).

In total, 450 different CARMa models were considered (5 3DRISM fields $\times$ 6 grid spacings $\times$ 3 regression methods $\times$ 5 datasets). All of the results are compiled in Table 2 (training dataset) and Table 3 (testing dataset). As before, since correlation coefficients (q^2 or R^2) and predictive errors ($RMSE$) were found to be highly correlated for these datasets, only the correlation coefficients are presented in Tables 2 and 3, but all other statistics (RMSE, σ, bias) are provided in the Supporting Information; presenting q^2 statistics here also permits a

direct comparison to previously published results. The "-" entries in Tables 2 and 3 indicate that training PLS or RF models on 3D RISM fields with a 0.5 Å grid spacing was found to be prohibitively computationally expensive. The best predictions for the external test set are summarised in Table 4.

Table 2: Leave-one-out cross-validation statistics (q^2) for 5 pIC_{50} data sets using CARMa with various descriptors and grid spacings. In bold are the best models and each dataset using the PLS, GA-PLS and RF models.

GS^a (Å)	$g_O(r)^b$			$g_H(r)^c$			$SFED^d$			$g_{C-}(r)^e$			$g_{C+}(r)^f$		
	PLS	GA-PLS	RF	PLS	GA-PLS	RF	PLS	GA-PLS	RF	PLS	GA-PLS	RF	PLS	GA-PLS	RF
ACE															
0.5	-	0.702	-	-	0.698	-	-	0.666	-	-	0.714	-	-	0.709	-
1.0	0.640	0.759	0.683	0.629	0.751	0.692	0.622	0.695	0.682	0.649	0.752	0.675	0.649	0.747	0.695
1.5	0.621	0.780	0.696	0.621	0.777	0.676	0.620	0.731	0.684	0.648	0.774	0.692	0.658	0.754	0.675
2.0	0.619	0.743	0.679	0.642	***0.795***	0.681	0.619	0.732	0.687	0.626	0.758	0.695	0.637	0.741	0.680
2.5	0.634	0.672	0.697	0.609	0.722	0.701	0.638	0.743	0.696	**0.670**	0.752	0.707	0.658	0.716	0.702
3.0	0.633	0.685	0.698	0.623	0.717	0.693	0.617	0.685	0.688	0.631	0.672	**0.717**	0.659	0.714	0.692
AchE															
0.5	-	0.564	-	-	0.579	-	-	0.363	-	-	0.572	-	-	0.559	-
1.0	0.490	0.639	0.399	0.497	0.629	0.402	0.276	0.459	0.411	0.493	0.634	0.411	0.487	0.629	0.431
1.5	0.482	0.639	0.427	0.492	0.643	0.426	0.261	0.491	0.426	0.483	0.641	0.383	0.488	***0.682***	0.412
2.0	0.455	0.644	0.395	0.487	0.640	0.414	0.271	0.550	0.418	0.446	0.604	0.403	0.499	0.659	0.398
2.5	**0.508**	0.565	0.412	0.504	0.633	0.392	0.269	0.487	0.436	0.463	0.584	0.395	0.430	0.601	0.392
3.0	0.444	0.623	0.407	0.493	0.590	0.417	0.240	0.446	0.424	0.449	0.627	0.426	0.502	0.628	**0.449**
BZR															
0.5	-	0.314	-	-	0.309	-	-	0.214	-	-	0.310	-	-	0.306	-
1.0	0.232	0.389	0.314	0.235	0.373	0.306	0.117	0.323	0.281	0.232	0.397	0.367	0.211	0.369	0.363
1.5	0.230	0.416	0.309	0.234	0.397	0.296	0.114	0.365	0.286	0.238	0.439	0.337	0.224	0.442	**0.378**
2.0	0.217	0.404	0.353	0.223	0.390	0.309	0.111	0.366	0.298	0.234	0.431	0.356	0.228	0.426	0.341
2.5	0.202	0.456	0.304	0.240	0.390	0.310	0.116	0.410	0.279	**0.243**	0.416	0.354	0.208	0.419	0.346
3.0	0.210	0.417	0.297	0.247	***0.458***	0.328	0.098	0.391	0.263	0.219	0.427	0.311	0.179	0.406	0.336
COX2															
0.5	-	0.460	-	-	0.451	-	-	0.357	-	-	0.462	-	-	0.457	-
1.0	0.426	0.483	0.462	0.421	0.477	0.457	0.276	0.397	0.416	0.427	0.503	0.438	0.426	0.508	0.447
1.5	0.420	0.494	0.457	0.416	0.498	0.473	0.266	0.412	0.405	0.421	0.515	0.458	0.426	***0.536***	0.449
2.0	0.421	0.476	0.466	0.424	0.467	**0.473**	0.262	0.385	0.418	0.416	0.480	0.434	0.415	0.473	0.435
2.5	0.419	0.441	0.459	0.397	0.468	0.431	0.281	0.347	0.411	**0.439**	0.460	0.428	0.412	0.462	0.443
3.0	0.424	0.464	0.463	0.412	0.465	0.454	0.320	0.351	0.427	0.427	0.437	0.440	0.422	0.436	0.458
DHFR															
0.5	-	0.619	-	-	0.622	-	-	0.523	-	-	0.633	-	-	0.627	-
1.0	0.581	0.647	0.634	0.586	0.659	0.636	0.488	0.601	0.642	0.591	0.649	0.646	0.594	0.667	**0.644**
1.5	0.574	0.628	0.636	0.585	***0.668***	0.634	0.485	0.586	0.642	0.587	0.657	0.642	0.593	0.663	0.632
2.0	0.578	0.593	0.631	0.576	0.630	0.634	0.487	0.599	0.642	0.587	0.637	0.638	0.580	0.618	0.640
2.5	0.562	0.523	0.626	0.573	0.602	0.651	0.480	0.581	0.643	0.584	0.599	0.639	0.589	0.609	0.642
3.0	0.568	0.623	0.633	0.583	0.626	0.625	0.482	0.530	0.651	0.589	0.550	0.659	**0.598**	0.632	0.624

[a] Grid spacing; [b] Oxygen distribution; [c] Hydrogen distribution; [d] Solvent free energy distribution; [e] Carbon probe distribution with -1 charge; [f] Carbon probe distribution with +1 charge.

Table 3: Test set predictive accuracy statistics (r^2) for 5 pIC_{50} data sets using CARMa with various descriptors and grid spacings. In bold are the best models and each dataset using the PLS, GA-PLS and RF models.

GS[a] *(Å)*	*$g_O(r)$*[b]			*$g_H(r)$*[c]			*SFED*[d]			*$g_{C-}(r)$*[e]			*$g_{C+}(r)$*[f]		
	PLS	GA-PLS	RF	PLS	GA-PLS	RF	PLS	GA-PLS	RF	PLS	GA-PLS	RF	PLS	GA-PLS	RF
ACE															
0.5	-	0.541	-	-	0.565	-	-	0.472	-	-	0.597	-	-	**0.615**	-
1.0	0.558	0.531	0.621	0.571	0.459	0.595	0.447	0.474	0.615	0.611	0.564	**0.636**	0.605	0.582	0.614
1.5	0.532	0.460	0.612	0.561	0.510	0.613	0.440	0.559	0.603	0.605	0.513	0.621	0.585	0.502	0.599
2.0	0.519	0.156	0.592	0.572	0.373	0.559	0.450	0.487	0.575	0.596	0.345	0.621	0.623	0.497	0.589
2.5	0.542	0.481	0.578	0.531	0.467	0.605	0.458	0.507	0.577	***0.638***	0.458	0.625	0.631	0.502	0.616
3.0	0.505	0.395	0.601	0.513	0.501	0.550	0.433	0.397	0.608	0.578	0.424	0.631	0.602	0.459	0.608
AchE															
0.5	-	0.670	-	-	0.673	-	-	0.438	-	-	0.676	-	-	***0.697***	-
1.0	0.626	0.629	0.454	0.632	0.637	0.476	0.404	0.474	0.506	0.660	0.587	0.405	**0.665**	0.601	0.402
1.5	0.632	0.422	0.460	0.623	0.490	0.488	0.414	0.423	0.518	0.658	0.696	0.443	0.659	0.595	0.385
2.0	0.587	0.459	0.454	0.583	0.317	0.484	0.366	0.494	**0.537**	0.634	0.491	0.445	0.644	0.500	0.359
2.5	0.603	0.393	0.493	0.648	0.481	0.471	0.373	0.364	0.495	0.608	0.345	0.456	0.601	0.364	0.359
3.0	0.606	0.429	0.468	0.637	0.365	0.472	0.383	0.335	0.526	0.621	0.208	0.431	0.654	0.269	0.397
BZR															
0.5	-	0.186	-	-	0.166	-	-	0.088	-	-	0.183	-	-	0.187	-
1.0	0.177	0.142	0.202	0.190	0.142	0.198	0.095	0.114	0.205	0.203	0.125	0.198	0.197	0.165	0.192
1.5	0.184	0.078	0.197	0.181	0.205	0.202	0.084	0.045	0.180	**0.209**	0.195	0.196	0.192	0.092	0.203
2.0	0.171	0.130	0.214	0.194	**0.208**	0.193	0.092	0.156	0.193	0.184	0.055	0.199	0.191	0.150	0.198
2.5	0.189	0.116	0.189	0.188	0.033	0.186	0.111	0.125	0.188	0.172	0.114	0.183	0.184	0.074	0.185
3.0	0.155	0.076	0.203	0.166	0.060	0.193	0.071	0.118	0.188	0.193	0.102	***0.217***	0.151	0.095	0.209
COX2															
0.5	-	0.327	-	-	0.334	-	-	0.200	-	-	**0.351**	-	-	0.336	-
1.0	0.343	0.322	0.347	0.346	0.341	0.347	0.176	0.241	0.355	0.365	0.342	0.353	0.366	0.282	0.341
1.5	0.326	0.243	0.353	0.348	0.224	0.357	0.166	0.260	0.372	0.344	0.248	0.364	0.363	0.266	0.370
2.0	0.308	0.303	0.338	0.331	0.257	0.335	0.188	0.183	0.341	0.334	0.216	0.357	0.342	0.269	0.339
2.5	0.303	0.164	0.349	0.299	0.251	0.339	0.176	0.182	0.346	0.312	0.225	0.338	0.323	0.228	0.374
3.0	0.323	0.249	0.343	0.323	0.188	0.318	0.156	0.139	0.367	0.348	0.198	0.318	***0.382***	0.172	**0.375**
DHFR															
0.5	-	0.548	-	-	0.545	-	-	0.421	-	-	**0.567**	-	-	0.548	-
1.0	0.540	0.513	0.603	0.539	0.548	0.604	0.397	0.485	0.567	0.533	0.524	0.597	0.538	0.514	0.600
1.5	0.534	0.532	0.606	0.535	0.527	0.601	0.392	0.486	0.566	0.529	0.560	0.590	0.537	0.504	0.630
2.0	0.517	0.439	0.610	0.531	0.519	0.612	0.390	0.455	0.555	0.510	0.430	0.601	0.532	0.475	0.604
2.5	0.518	0.396	0.621	0.515	0.371	0.622	0.381	0.410	0.540	0.524	0.497	0.598	0.538	0.453	0.620
3.0	0.536	0.425	***0.652***	0.530	0.463	0.612	0.375	0.351	0.528	0.532	0.454	0.589	**0.562**	0.515	0.613

[a] Grid Spacing; [b] Oxygen Distribution; [c] Hydrogen Distribution; [d] Solvent Free Energy Distribution

Table 4: Best test set predictive accuracy statistics (r^2) for the pIC_{50} data sets compared to CoMFA and best literature model.

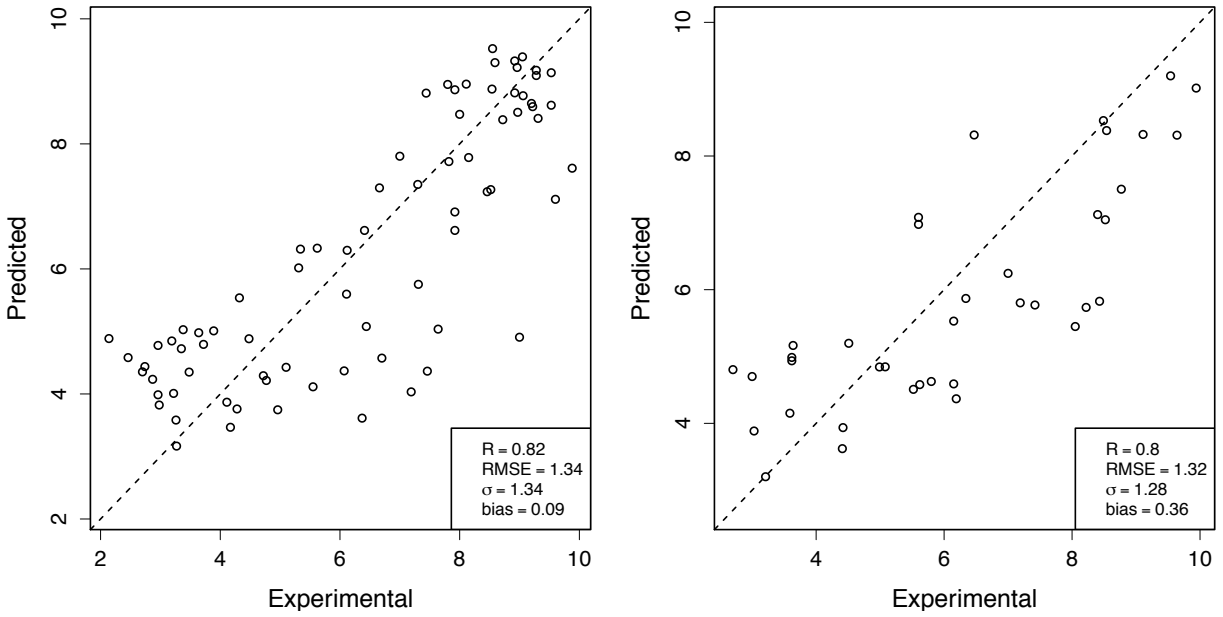
	q^2	Grid Spacing Å	RMSE[a]	Descriptor
ACE				
CoMFA	0.490	2.0	1.520	-
CoMSIA Basic	0.520	2.0	1.460	-
PLS	0.638	2.5	1.325	$g_{C-}(r)$
GA-PLS	0.615	0.5	1.366	$g_{C+}(r)$
RF	0.636	1.0	1.304	$g_{C-}(r)$
AchE				
CoMFA	0.470	2.0	0.937	-
PLS	0.665	1.0	0.791	$g_{C+}(r)$
GA-PLS	0.697	0.5	0.761	$g_{C+}(r)$
RF	0.537	2.0	0.918	SFED
BZR				
CoMFA	0.000	2.0	0.960	-
2.5D	0.200	2.0	0.861	-
PLS	0.209	1.5	0.878	$g_{C-}(r)$
GA-PLS	0.208	2.0	0.848	$g_H(r)$
RF	0.217	3.0	0.863	$g_{C-}(r)$
COX2				
CoMFA	0.290	2.0	1.233	-
CoMSIA Extra	0.370	2.0	1.164	-
PLS	0.382	3.0	1.159	$g_{C+}(r)$
GA-PLS	0.351	0.5	1.211	$g_{C-}(r)$
RF	0.375	3.0	1.252	$g_{C+}(r)$
DHFR				
CoMFA	0.590	2.0	0.886	-
HQSAR	0.630	2.0	0.837	-
PLS	0.562	3.0	0.913	$g_{C+}(r)$
GA-PLS	0.567	0.5	0.913	$g_{C-}(r)$
RF	0.652	3.0	0.837	$g_O(r)$

[a] For literature results this has been recalculated from the standard error of prediction (s) reported by Sutherland *et al.*[1] as: RMSE = $\sqrt{((s^2)(N-1/N)}$.

ACE Dataset. The ACE dataset comprises pIC50 data for 114 inhibitors of angiotensin converting enzyme separated into a training dataset of 76 and a test dataset of 38 molecules. The pIC50 values range between 2.1 - 9.9. Inspection of the data in Tables 2 and 3 show that the CARMa models are relatively insensitive to the choice of 3D RISM field or grid-spacing for this dataset. The most accurate predictions were obtained using either PLS or RF

regression on $g_{C-}(r)$ variables. For the external test set, the RF model has a slightly smaller error ($RMSE = 1.304$) than the PLS model ($RMSE = 1.325$), but both methods report $R^2 = 0.64$ (2 decimal places). The correlation between experimental and predicted pIC50 data for the PLS model is illustrated Figure 6. By comparison, the most accurate predictions reported by Sutherland et al. were less accurate: CoMSIA ($R^2 = 0.520, RMSE = 1.46$) and CoMFA ($R^2 = 0.490, RMSE = 1.52$).

Using a GA to select input variables for the PLS method leads to a high q^2 for cross-validation, which is not surprising given that the GA fitness function was $RMSE$ for 3-fold cross-validation, but these models do not generalise as well as the PLS or RF models; the best GA-PLS prediction of the test set is $R^2 = 0.615$ and $RMSE = 1.366$.

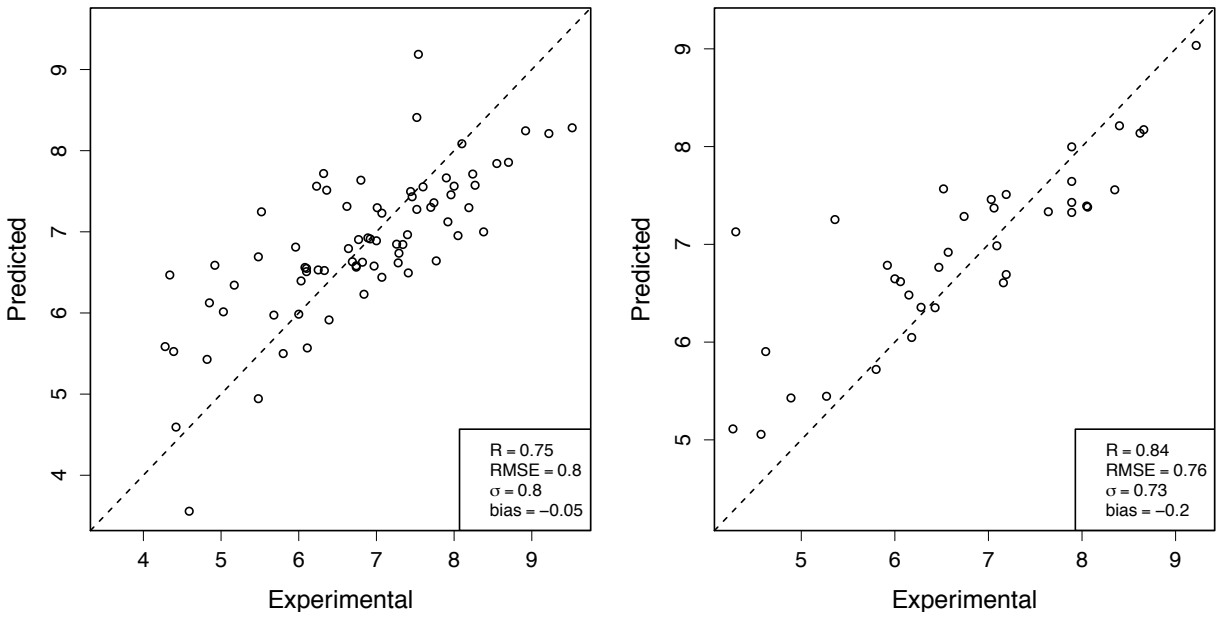


(a) Training Set (LOO-CV) (b) Test set

Figure 6: ACE correlation graphs of leave-one-out cross-validation (LOO-CV) (a) and test set (b) predictive accuracy for the CARMa PLS model using the $g_{C-}(r)$ probe atom distribution descriptor at 2.5 Å grid spacing.

AchE Dataset. The pIC50 values for the 111 acetylcholinesterase inhibitors in the AchE dataset range from 4.3 - 9.5. Sutherland et al. found CoMFA to be more accurate than other QSAR methods for modelling this dataset ($R^2 = 0.47$ and $RMSE = 0.937$). Tables 3 and

4 show that an improvement in accuracy can be made by replacing CoMFA's electric/steric fields with $g_{C+}(r)$ variables giving $R^2 = 0.665$ and $RMSE = 0.791$ for PLS regression. Using a GA to select input variables for PLS further improves the accuracy for most 3DRISM fields and grid-spacings. The best CARMa model was obtained with GA-PLS regression on $g_{C+}(r)$ variables giving $R^2 = 0.697$ and $RMSE = 0.761$ (Table 4). The correlation diagrams for this model are presented in Figure 7.



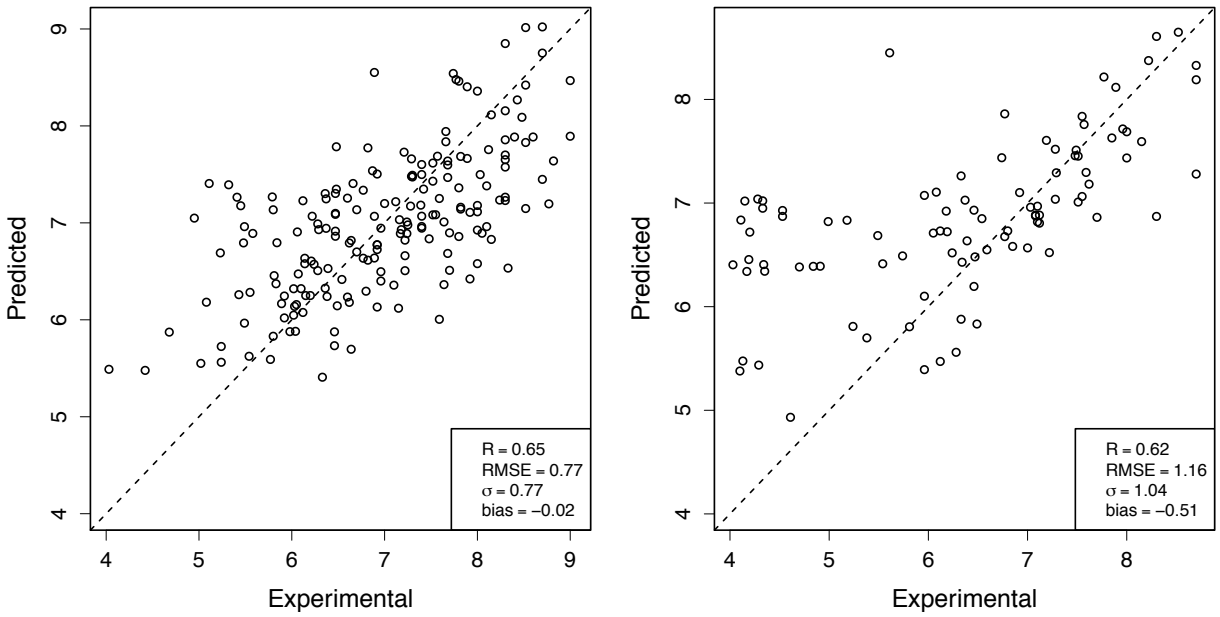
(a) Training Set (LOO-CV) (b) Test Set

Figure 7: AchE correlation graphs of leave-one-out cross-validation (LOO-CV) (a) and test set (b) predictive accuracy for the CARMa GA-PLS model using the $g_{C+}(r)$ probe atom distribution descriptor at 0.5 Å grid spacing.

BZR and COX2 datasets The BZR and COX2 data have previously proven to be almost impossible to model accurately using QSAR methods. Sutherland et al. reported $R^2 = 0$ and $R^2 = 0.29$ for CoMFA predictions of the BZR and COX2 test sets, respectively. The best results were $R^2 = 0.200$ and $RMSE = 0.861$ for a "2.5D" QSAR model of the BZR data and $R^2 = 0.370$ and $RMSE = 1.164$ for a CoMSIA Extra model of the COX2 data; both of these models were considered to be too poor to be particularly useful. As would be expected, the CARMa method is also not able to produce very accurate models for these

datasets, but in both cases it improves on the CoMFA results and matches or improves upon the other predictions. For the BZR dataset, a CARMa model using $g_{C-}(r)$ variables and RF regression gives $R^2 = 0.217$ and $RMSE = 0.863$, while for the COX2 dataset a PLS model trained on $g_{C+}(r)$ variables gives $R^2 = 0.217$ and $RMSE = 1.159$.

For the COX2 dataset, part of the reason for the poor test set prediction is that the training and test sets cover different ranges of property space. The correlation diagram for the PLS model on $g_{C+}(r)$ variables is given in Figure 8a. There are only three compounds with pIC_{50} values below 5 in the training set, whereas in the test set there are 19 compounds fitting this criteria. Figure 8b shows that compounds with pIC_{50} values above 5 are relatively well predicted, with the exception of one or two outliers, but the 19 compounds with pIC_{50} values below 5 have all been overestimated.

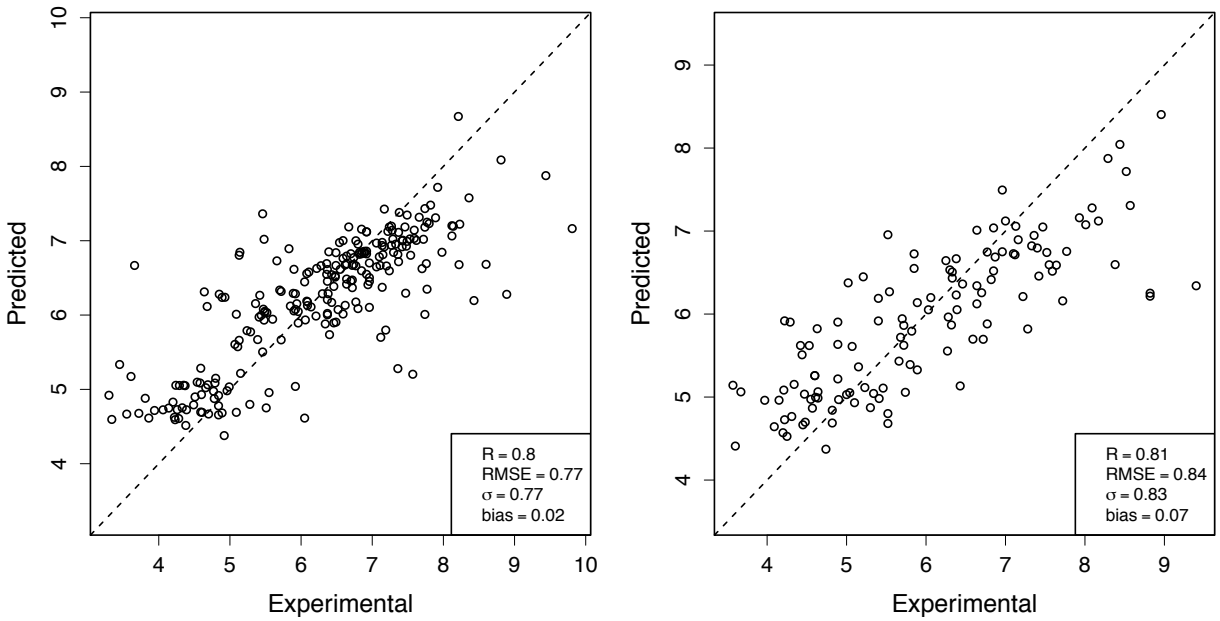


(a) Training Set (LOO-CV) (b) Test Set

Figure 8: COX2 correlation graphs of leave-one-out cross-validation (LOO-CV) (a) and test set (b) predictive accuracy for the CARMa PLS model using the $g_{C+}(r)$ probe atom distribution descriptor at 3.0 Å grid spacing.

DHFR Dataset. A CoMFA model of the DHFR data has previously been reported to give a $R^2 = 0.590$ and $RMSE = 0.886$, while the HQSAR produces an improved result

$R^2 = 0.630$ and $RMSE = 0.837$. The best CARMa model is found using the RF method and $g_O(r)$ variables at 3.0 Å grid-spacing, which has $R^2 = 0.652$ and $RMSE = 0.837$. In Table 4, CARMa is shown to improve R^2 in comparison to CoMFA by 6.2% when the RF method is used with the $g_O(r)$ descriptor. In fact, the RF method produces the best result for all five descriptors tested here. The poorest results are obtained from the SFED descriptors as shown in Table 3. The PLS and GA-PLS methods produce results comparable to the literature when used with $g_O(r)$ and $g_H(r)$ descriptors, but improved results when used with the probe atom descriptors, $g_{C-}(r)$ and $g_{C+}(r)$. Figure 9a shows the correlation diagram for cross-validation of the DHFR training data using the best RF model. It is apparent that the models do not make very accurate predictions for molecules with pIC_{50} values above 8, which may partly be because this region of property space is under-represented in the training dataset.



(a) Training Set (LOO-CV) (b) Test Set

Figure 9: DHFR correlation graphs of leave-one-out cross-validation (LOO-CV) (a) and test set (b) predictive accuracy for the CARMa RF model using the $g_O(r)$ distribution descriptor at 3.0 Å grid spacing.

Discussion

The predictive accuracy of CARMa models using various parameters and descriptors have been examined using the steroid dataset defined by Cramer *et al* in 1988 and the five largest data sets reported by Sutherland *et al* in 2004.[1,9] The physiochemical properties of compounds were encoded using 3D-RISM calculations for application in field-based QSAR. The 3D-RISM calculations provided solvent density distribution functions ($g_O(r)$, $g_H(r)$, $g_{C-}(r)$, $g_{C+}(r)$) and SFED distribution functions. The models were implemented using PLS, GA-PLS and RF regression.

Steroids

For the steroid dataset, only PLS regression was used as the dataset was considered to be too small to train reliable GA/PLS or RF models; using PLS with three latent variables also permits a direct comparison with the CoMFA results of Cramer et al.[7] From the CV results, it emerges that the solvent density distribution functions ($g_O(r)$,$g_H(r)$, $g_{C+}(r)$, $g_{C-}(r)$) all perform better than the SFED distribution function. With the exception of SFED, the CARMa and CARMa(electrolyte) descriptors all give more accurate models than CoMFA, Table 1. We find that there is no preferred grid spacing, which suggests that the coarse grids contain much of the same information as the finer grids.

pIC_{50} Data Sets

Several trends are evident on comparison of the predictive accuracy of the CARMa,CoMFA and 3D-QSAR models reported in Tables 3 and 4,

(1) the predictions made by CARMa and CARMa(electrolyte) are more accurate than CoMFA for most of the 5 data sets regardless of which regression method is used. The exceptions are the PLS and GA/PLS (but not RF) models of the DHRF dataset.

(2) the probe atom distribution descriptors consistently produce accurate predictions of

the protein-ligand binding assay data. The best predictive model derived from the probe atom descriptors performs substantially better than the best literature model for the ACE, AchE, BZR and COX2 data sets. In the case of DHFR the best model derived from the probe atom descriptors is of comparable accuracy to the best literature model (HQSAR).

(3) the SFED descriptor tends to perform poorly in comparison to the solvent density descriptors ($g_O(r)$, $g_H(r)$, $g_{C+}(r)$, $g_{C-}(r)$). This observation is interesting because both previous 3D-QSAR studies using 1D or 3D RISM have employed SFED.[20,21]

(4) the $g_O(r)$ and $g_H(r)$ models tend to generate similar results. This is perhaps because oxygen and hydrogen atoms are covalently bonded in water molecules, resulting in similar information being captured in their distribution functions.

Overview

Since no clear consensus is reached with respect to the optimal choices of grid-spacing or regression method for CARMa, it is instructive to compare the CARMa and CoMFA results for the parameters used in the CoMFA model (PLS regression with 3 latent variables and a 2 Å grid spacing). As shown in Table 5, more of the variance in the test set is explained by models trained on CARMa solvent density distribution functions than either the CoMFA or CARMa SFED variables. The rank order of the accuracy of the CARMa variables is $g_{C+}(r) > g_{C-}(r) > g_H(r) > g_O(r) >> SFED$.

Table 5: Comparison of test set predictive accuracy statistics (r^2) for the pIC_{50} data sets modelled using PLS regression and a 2 Å grid spacing. The final row gives the mean of the r^2 values for all five datasets

Dataset	CoMFA	$g_{C+}(r)$	$g_{C-}(r)$	$g_H(r)$	$g_O(r)$	SFED
ACE	0.49	0.623	0.596	0.572	0.519	0.45
AcHE	0.47	0.644	0.634	0.583	0.587	0.366
BZR	0.00	0.191	0.184	0.194	0.171	0.092
COX2	0.29	0.342	0.334	0.331	0.308	0.188
DHFR	0.59	0.532	0.510	0.531	0.517	0.39
$\overline{(r^2)}$	0.368	0.466	0.452	0.442	0.420	0.297

Conclusions

We have proposed an extension of the CARMa methodology introduced by Güssregen et al.[21] in which charged carbon probe atoms commonly used in CoMFA are inserted in the 3DRISM solvent model to capture specific molecular interactions. Extensive benchmarking over datasets for six different protein-ligand systems demonstrates that the original CARMa method performs better than CoMFA in most cases. Using solvent density distribution functions ($g_O(r)$,$g_H(r)$) gives consistently more accurate predictions than using solvation free energy density distributions. When the CARMa models are developed using density distribution functions for C+/C- probe atoms in place of those for water, there is a small but consistent increase in prediction accuracy; the $g_{C+}(r)$ and $g_{C-}(r)$ variables give the most accurate results for 5 of the 6 datasets. Although the 3DRISM equations should be solved on a relatively fine grid to ensure physical accuracy (grid spacing $\approx$ 0.5 Å), in most cases converting to a coarser grid (1 Å to 3 Å spacing) for use in CARMa doesn't significantly reduce prediction accuracy, but does simplify the statistical modelling procedure. Using a 2 Å grid spacing and PLS regression, the CARMa solvent density distribution functions give consistently more accurate predictions than CoMFA (Table 5).

There is clearly scope for future work. From one side, the CARMa models would be expected to benefit from improvements in standard QSAR procedures or 3DRISM theory. Better algorithms for molecular alignment may reduce statistical noise and improve prediction accuracy. More advanced machine learning techniques may be better suited to solve the underdetermined regression problem posed by CARMa (and CoMFA). Open problems in 3DRISM theory include the design of bridge functionals, free energy functionals, and more efficient and robust algorithms for solving the RISM equations.[28] From another side, it may be possible to choose solvents or probe atoms that capture molecular interactions more completely or that more closely mimic biological environments. However, none of these ideas for future work should limit the use of CARMa now. Since the CARMa variables can be computed at minimal computational expense using existing software (e.g. AMBER),[61] the

method can already be easily implemented and used in drug discovery.

Acknowledgement

We thank Ian Wall, Lucia Fusani and Alvaro Cortes from GSK (Stevenage, UK) for useful discussions. S.M.A and D.S.P. are grateful for use of the EPSRC funded ARCHIE-WeSt High Performance Computer (www.archie-west.ac.uk, EPSRC Grant No. EP/K000586/1). D.S.P. thanks the University of Strathclyde for support through its Strategic Appointment and Investment Scheme.

Supporting Information Available

Summative statistics for training, cross-validation and testing of each CARMa model.

This material is available free of charge via the Internet at `http://pubs.acs.org/`.

References

(1) Sutherland, J. J.; O'Brien, L. A.; Weaver, D. F. A Comparison of Methods for Modeling Quantitative Structure- Activity Relationships. *J. Med. Chem* **2004**, *47*, 5541–5554.

(2) Hall, L. H.; Kier, L. B. The Molecular Connectivity Chi Indexes and Kappa Shape Indexes in Structure-Property Modeling. *Rev. Comput. Chem., Volume 2* **2007**, 367–422.

(3) Stone, M.; Jonathan, P. Statistical Thinking and Technique for QSAR and Related Studies. Part I: General theory. *J. Chemom.* **1993**, *7*, 455–475.

(4) Oprea, T. I.; Waller, C. L. Theoretical and Practical Aspects of Three-Dimensional Quantitative Structure-Activity Relationships. *Rev. Comput. Chem., Volume 11* **2007**, 127–182.

(5) Hopfinger, A.; Wang, S.; Tokarski, J. S.; Jin, B.; Albuquerque, M.; Madhav, P. J.;

Duraiswami, C. Construction of 3D-QSAR models using the 4D-QSAR Analysis Formalism. *J. Am. Chem. Soc.* **1997**, *119*, 10509–10524.

(6) Andrade, C. H.; Pasqualoto, K. F.; Ferreira, E. I.; Hopfinger, A. J. 4D-QSAR: Perspectives in Drug Design. *Molecules* **2010**, *15*, 3281–3294.

(7) Cramer, R. D.; Patterson, D. E.; Bunce, J. D. Comparative Molecular Field Analysis (CoMFA). 1. Effect of Shape on Binding of Steroids to Carrier Proteins. *J. Am. Chem. Soc.* **1988**, *110*, 5959–5967.

(8) Clark, M.; Cramer, R. D.; Jones, D. M.; Patterson, D. E.; Simeroth, P. E. Comparative Molecular Field Analysis (CoMFA). 2. Toward its use with 3D-Structural Databases. *Tetrahedron Comput. Methodol.* **1990**, *3*, 47–59.

(9) Coats, E. A. The CoMFA Steroids as a Benchmark Dataset for Development of 3D QSAR Methods. *Perspect. Drug Discov.* **1998**, *12*, 199–214.

(10) Klebe, G.; Abraham, U.; Mietzner, T. Molecular Similarity Indices in a Comparative Analysis (CoMSIA) of Drug Molecules to Correlate and Predict their Biological Activity. *J. Med. Chem* **1994**, *37*, 4130–4146.

(11) Klebe, G.; Abraham, U. Comparative Molecular Similarity Index Analysis (CoMSIA) to Study Hydrogen-Bonding Properties and to Score Combinatorial Libraries. *J. Comput. Aided Mol. Des.* **1999**, *13*, 1–10.

(12) Riniker, S. Molecular Dynamics Fingerprints (MDFP): Machine Learning from MD Data To Predict Free-Energy Differences. *J. Chem. Inf. Model.* **2017**, *57*, 726–741.

(13) Ratkova, E. L.; Chuev, G. N.; Sergiievskyi, V. P.; Fedorov, M. V. An Accurate Prediction of Hydration Free Energies by Combination of Molecular Integral Equations Theory with Structural Descriptors. *J. Phys. Chem. B* **2010**, *114*, 12068–12079.

(14) Palmer, D. S.; Frolov, A. I.; Ratkova, E. L.; Fedorov, M. V. Towards a universal method to calculate hydration free energies: a 3D reference interaction site model with partial molar volume correction. *J. Phys. Cond. Matt.* **2010**, *22*, 492101.

(15) Palmer, D. S.; Frolov, A. I.; Ratkova, E. L.; Fedorov, M. V. Toward a Universal Model To Calculate the Solvation Thermodynamics of Druglike Molecules: The Importance of New Experimental Databases. *Mol. Pharmaceutics* **2011**, *8*, 1423–1429.

(16) Chandler, D.; Andersen, H. C. Optimized Cluster Expansions for Classical Fluids. 2. Theory of Molecular Liquids. *J. Chem. Phys.* **1972**, *57*, 1930–1937.

(17) Beglov, D.; Roux, B. Numerical-Solution Of The Hypernetted-Chain Equation For A Solute Of Arbitrary Geometry In 3 Dimensions. *J. Chem. Phys.* **1995**, *103*, 360–364.

(18) Beglov, D.; Roux, B. Solvation of complex molecules in a polar liquid: An integral equation theory. *J. Chem. Phys.* **1996**, *104*, 8678–8689.

(19) Beglov, D.; Roux, B. An Integral Equation to Describe the Solvation of Polar Molecules in liquid water. *J. Phys. Chem.* **1997**, *101*, 7821–7826.

(20) Palmer, D. S.; Mišin, M.; Fedorov, M. V.; Llinas, A. Fast and General Method To Predict the Physicochemical Properties of Druglike Molecules Using the Integral Equation Theory of Molecular Liquids. *Mol. Pharmaceutics* **2015**, *12*, 3420–3432.

(21) Gussregen, S.; Matter, H.; Hessler, G.; Lionta, E.; Heil, J.; Kast, S. M. Thermodynamic Characterization of Hydration Sites from Integral Equation-Derived Free Energy Densities: Application to Protein Binding Sites and Ligand Series. *J. Chem. Inf. Model.* **2017**, *57*, 1652âĂŞ1666.

(22) Du, Q. H.; Beglov, D.; Roux, B. Solvation free energy of polar and nonpolar molecules in water: An extended interaction site integral equation theory in three dimensions. *J. Phys. Chem. B* **2000**, *104*, 796–805.

(23) Kovalenko, A.; Hirata, F. Three-dimensional density profiles of water in contact with a solute of arbitrary shape: A RISM approach. *Chem. Phys. Lett.* **1998**, *290*, 237–244.

(24) Hirata, F., Ed. *Molecular Theory of Solvation*; Kluwer Academic Publishers, Dordrecht, Netherlands, 2003.

(25) Luchko, T.; Gusarov, S.; Roe, D. R.; Simmerling, C.; Case, D. A.; Tuszynski, J.; Kovalenko, A. Three-Dimensional Molecular Theory of Solvation Coupled with Molecular Dynamics in Amber. *J. Chem. Theory Comput.* **2010**, *6*, 607–624.

(26) Kovalenko, A.; Hirata, F. Potential of Mean Force between Two Molecular Ions in a Polar Molecular Solvent: A Study by the Three-Dimensional Reference Interaction Site Model. *J. Phys. Chem. B* **1999**, *103*, 7942–7957.

(27) Kast, S. M.; Kloss, T. Closed-Form Expressions of the Chemical Potential for Integral Equation Closures with Certain Bridge Functions. *J. Chem. Phys.* **2008**, *129*, 236101.

(28) Ratkova, E. L.; Palmer, D. S.; Fedorov, M. V. Solvation Thermodynamics of Organic Molecules by the Molecular Integral Equation Theory: Approaching Chemical Accuracy. *Chem. Rev.* **2015**, *115*, 6312–6356, PMID: 26073187.

(29) Palmer, D. S.; Sergiievskyi, V. P.; Jensen, F.; Fedorov, M. V. Accurate Calculations of the Hydration Free Energies of Druglike Molecules using the Reference Interaction Site Model. *J. Chem. Phys.* **2010**, *133*, 044104.

(30) Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. Development and Testing of a General Amber Force Field. *J. Comput. Chem.* **2004**, *25*, 1157–1174.

(31) Kovalenko, A.; Hirata, F. Potentials of mean force of simple ions in ambient aqueous solution. I. Three-dimensional reference interaction site model approach. *J. Chem. Phys.* **2000**, *112*, 10391–10402.

(32) Chandler, D.; Singh, Y.; Richardson, D. M. Excess Electrons In Simple Fluids .1. General Equilibrium-Theory For Classical Hard-Sphere Solvents. *J. Chem. Phys.* **1984**, *81*, 1975–1982.

(33) Misin, M.; Fedorov, M. V.; Palmer, D. S. Accurate Hydration Free Energies at a Wide Range of Temperatures from 3D-RISM. *J. Chem. Phys.* **2015**, *142*, 091105.

(34) Friedman, J.; Hastie, T.; Tibshirani, R. Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **2010**, *33*, 1.

(35) Conrads, M.; Nordin, P.; Banzhaf, W. *Lecture Notes in Computer Science*; Springer Berlin Heidelberg, 1998; pp 113–129.

(36) Bies, R. R.; Muldoon, M. F.; Pollock, B. G.; Manuck, S.; Smith, G.; Sale, M. E. A Genetic Algorithm-Based, Hybrid Machine Learning Approach to Model Selection. *J. Pharmacokinet. Pharmacodyn.* **2006**, *33*, 195–221.

(37) Schmitt, L. M. Theory of Genetic Algorithms II: Models for Genetic Operators over the String-Tensor Representation of Populations and Convergence to Global Optima for Arbitrary Fitness Function under Scaling. *Theor. Comput. Sci.* **2004**, *310*, 181–231.

(38) Breiman, L. Random forests. *Machine learning* **2001**, *45*, 5–32.

(39) DePriest, S. A.; Mayer, D.; Naylor, C. B.; Marshall, G. R. 3D-QSAR of Angiotensin-Converting Enzyme and Thermolysin Inhibitors: A Comparison of CoMFA Models Based on Deduced and Experimentally Determined Active Site Geometries. *J. Am. Chem. Soc.* **1993**, *115*, 5372–5384.

(40) Golbraikh, A. Validation of Protein-Based Alignment in 3D Quantitative Structure–Activity Relationships with CoMFA models. *Eur. J. Med. Chem.* **2000**, *35*, 123–136.

(41) Maddalena, D. J.; Johnston, G. A. R. Prediction of Receptor Properties and Binding Affinity of Ligands to Benzodiazepine/GABAA Receptors Using Artificial Neural Networks. *J. Med. Chem* **1995**, *38*, 715–724.

(42) Chavatte, P.; Yous, S.; Marot, C.; Baurin, N.; Lesieur, D. Three-Dimensional Quantitative Structure-Activity Relationships of Cyclo-oxygenase-2 (COX-2) Inhibitors: A Comparative Molecular Field Analysis. *J. Med. Chem* **2001**, *44*, 3223–3230.

(43) Mattioni, B. E.; Jurs, P. C. Prediction of Dihydrofolate Reductase Inhibition and Selectivity using Computational Neural Networks and Linear Discriminant Analysis. *J. Mol. Graph. Model.* **2003**, *21*, 391–419.

(44) Lajiness, M.; Johnson, M.; Maggiora, G. Implementing Drug Screening Programs using Molecular Similarity Methods. *Prog. Clin. Biol. Res.* **1989**, *291*, 173.

(45) Hassan, M.; Bielawski, J. P.; Hempel, J. C.; Waldman, M. Optimization and Visualization of Molecular Diversity of Combinatorial Libraries. *Molec. Divers.* **1996**, *2*, 64–74.

(46) Case, D. A.; Berryman, J. T.; Betz, R. M.; Cerutti, D. S.; Cheatham, T. E.; III, T. A. D.; Duke, R. E.; Giese, T. J.; Gohlke, H.; Goetz, A. W.; Homeyer, N.; Izadi, S.; Janowski, P.; Kaus, J.; Kovalenko, A.; Lee, T. S.; LeGrand, S.; Li, P.; Luchko, T.; Luo, R.; Madej, B.; Merz, K. M.; Monard, G.; Needham, P.; Nguyen, H.; Nguyen, H. T.; Omelyan, I.; Onufriev, A.; Roe, D. R.; Roitberg, A.; Salomon-Ferrer, R.; Simmerling, C. L.; Smith, W.; Swails, J.; Walker, R. C.; Wang, J.; Wolf, R. M.; Wu, X.; York, D. M.; Kollman, P. A. AMBER 2015. 2015; `http://ambermd.org/.`, Accessed on 1st January 2018.

(47) Kovalenko, A.; Ten-No, S.; Hirata, F. Solution of Three-Dimensional Reference Interaction Site Model and Hypernetted Chain Equations for Simple Point Charge Water by

Modified Method of Direct Inversion in Iterative Subspace. *J. Comput. Chem.* **1999**, *20*, 928–936.

(48) Perkyns, J. S.; Pettitt, B. M. A Dielectrically Consistent Interaction Site Theory For Solvent Electrolyte Mixtures. *Chem. Phys. Lett.* **1992**, *190*, 626–630.

(49) Lue, L.; Blankschtein, D. Liquid-State Theory of Hydrocarbon Water-Systems: Application to Methane, Ethane, and Propane. *J. Phys. Chem.* **1992**, *96*, 8582–8594.

(50) Berendsen, H. J. C.; Grigera, J. R.; Straatsma, T. P. The Missing Term in Effective Pair Potentials. *J. Phys. Chem.* **1987**, *91*, 6269–6271.

(51) Hirata, F.; Rossky, P. J. An Extended Rism Equation For Molecular Polar Fluids. *Chem. Phys. Lett.* **1981**, *83*, 329–334.

(52) Lee, P. H.; Maggiora, G. M. Solvation Thermodynamics Of Polar-Molecules In Aqueous-Solution By The Xrism Method. *J. Phys. Chem.* **1993**, *97*, 10175–10185.

(53) Kovalenko, A.; Hirata, F. Hydration Free Energy of Hydrophobic Solutes Studied by a Reference Interaction Site Model with a Repulsive Bridge Correction and a Thermodynamic Perturbation Method. *J. Chem. Phys.* **2000**, *113*, 2793–2805.

(54) Chuev, G.; Fedorov, M.; Crain, J. Improved estimates for hydration free energy obtained by the reference interaction site model. *Chem. Phys. Lett.* **2007**, *448*, 198–202.

(55) Allen, M. P., Tildesley, D. J., Eds. *Computer Simulation of Liquids*; Clarendon Press, Oxford, 1987.

(56) Wehrens, R.; Mevik, B.-H. The PLS Package: Principal Component and Partial Least Squares Regression in R. *J. Stat. Soft.* **2007**, *18*, 1–23.

(57) Team, R. C. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2014. 2014.

(58) Liaw, A.; Wiener, M. Classification and Regression by randomForest. *R news* **2002**, *2*, 18–22.

(59) Eklund, M.; Norinder, U.; Boyer, S.; Carlsson, L. Choosing Feature Selection and Learning Algorithms in QSAR. *J. Chem. Inf. Model.* **2014**, *54*, 837–843.

(60) Palmer, D. S.; O'Boyle, N. M.; Glen, R. C.; Mitchell, J. B. O. Random Forest Models to Predict Aqueous Solubility. *J. Chem. Inf. Model.* **2007**, *47*, 150–158.

(61) Case, D. A.; Cheatham, T. E.; Darden, T.; Gohlke, H.; Luo, R.; Merz, K. M.; Onufriev, A.; Simmerling, C.; Wang, B.; Woods, R. J. The Amber Biomolecular Simulation Programs. *J. Comput. Chem.* **2005**, *26*, 1668–1688.

Graphical TOC Entry