Journal of Chemical Information and Modeling

Article

pubs.acs.org/jcim

Computation of Binding Energies Including Their Enthalpy and Entropy Components for Protein–Ligand Complexes Using Support Vector Machines

Chaitanya A. K. Koppisetty,[†,‡] Martin Frank,[†] Graham J. L. Kemp,[*,‡] and Per-Georg Nyholm[*,†,§]

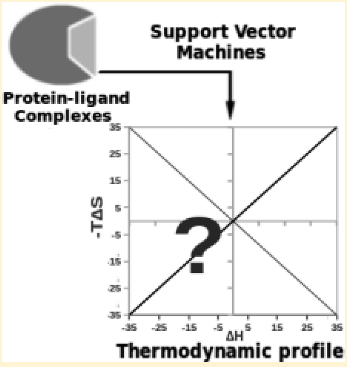
[†]Biognos AB, Generatorsgatan 1, 417 05 Gothenburg, Sweden

[‡]Department of Computer Science and Engineering, Chalmers University of Technology, 412 96 Gothenburg, Sweden

[§]Institute of Biomedicine, Department of Medical Biochemistry, University of Gothenburg, 413 90 Gothenburg, Sweden

Supporting Information

ABSTRACT: Computing binding energies of protein–ligand complexes including their enthalpy and entropy terms by means of computational methods is an appealing approach for selecting initial hits and for further optimization in early stages of drug discovery. Despite the importance, computational predictions of thermodynamic components have evaded attention and reasonable solutions. In this study, support vector machines are used for developing scoring functions to compute binding energies and their enthalpy and entropy components of protein–ligand complexes. The binding energies computed from our newly derived scoring functions have better Pearson's correlation coefficients with experimental data than previously reported scoring functions in benchmarks for protein–ligand complexes from the PDBBind database. The protein–ligand complexes with binding energies dominated by enthalpy or entropy term could be qualitatively classified by the newly derived scoring functions with high accuracy. Furthermore, it is found that the inclusion of comprehensive descriptors based on ligand properties in the scoring functions improved the accuracy of classification as well as the prediction of binding energies including their thermodynamic components. The prediction of binding energies including the enthalpy and entropy components using the support vector machine based scoring functions should be of value in the drug discovery process.



INTRODUCTION

For protein–ligand interactions, it is known that enthalpy (ΔH) and entropy terms ($T\Delta S$) constitute the thermodynamic components of the binding energy (ΔG). Information about these components is crucial for an understanding of protein–ligand binding and thus for rational drug design.[1,2] Isothermal titration calorimetry (ITC) is an established experimental method from which one can accurately measure ΔG, ΔH, and $T\Delta S$ values.[3] However, high throughput measurements using ITC are currently not feasible since substantial quantities of samples are needed for these experiments, and often these are not available. Therefore there is a need for accurate computational methods to estimate the thermodynamic components of protein–ligand binding. Enthalpy reflects the strength of the protein–ligand interactions primarily because of hydrogen bonding and van der Waals interactions. Favorable enthalpy requires complementary placement of hydrogen bond acceptor and donor groups at the binding interface. Entropic contributions could arise from the solute (protein and the ligand), as well as from the solvent (usually water). The ligand and protein often have decreased conformational freedom upon binding thus resulting in loss of entropy. On the other hand, hydrophobic interactions at the binding interface could release water from the binding site and thus increase the solvent entropy. Depending on the nature of the protein–ligand interactions, the solvent might lose or gain entropy.[4−6] In the process of rational drug design, molecules with favorable enthalpy and entropy contributions to binding energy (ΔG_{H+} in Figure 1) are usually regarded as good hits. Analysis of ITC data shows that binding energies for synthetic inhibitors often have similar contributions from enthalpy and entropy whereas biological protein–ligand interactions often have binding energy dominated by enthalpy.[4] However, within a series of congeneric inhibitors with increasing size a phenomenon called 'enthalpy–entropy compensation', in which a strong enthalpy component is accompanied by an unfavorable entropy component, has been observed in several cases.[7−10]

Ligands with binding energy dominated by specific interactions defining the enthalpy are susceptible to loss in binding due to mutations in the target protein. Drugs with entropy-dominated binding energies (ΔG_{S+} and ΔG_{S-}) could have dominating hydrophobic interactions and lack hydrogen bonds that confer specificity. In these cases conformational restraints and lack of adaptation could also cause sensitivity to mutations in the target protein. Sensitivity to mutations in the target protein causes drug resistance, which is a problem, for example, in antiviral therapies.[11] The phenomenon of drug

Received: May 29, 2013

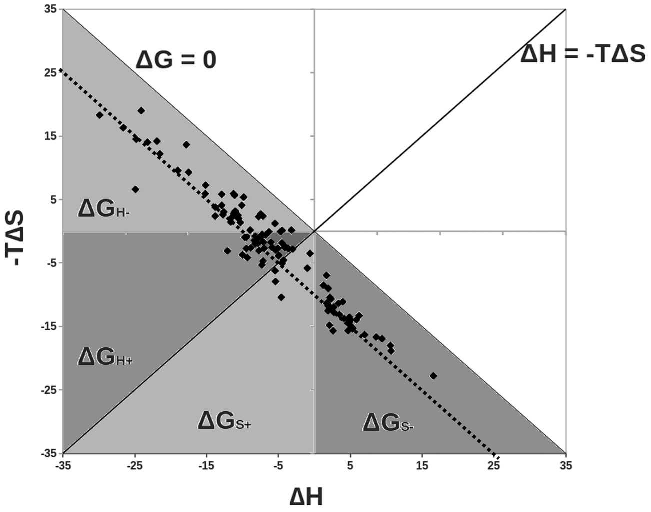


Figure 1. Four possible cases of protein−ligand binding energies with varying ΔH and $T\Delta S$ (units kcal mol^{-1}). Cases ΔG_{H-} and ΔG_{H+} represent binding energy dominated by enthalpy with unfavorable and favorable entropy components, respectively. Cases ΔG_{S+} and ΔG_{S-} represent binding energies dominated by entropy with favorable and unfavorable enthalpy components, respectively. The dots represent the enthalpy and entropy components for 120 protein−ligand complexes in the training data set used in this study. 39 belong to the ΔG_{H-} case, 36 belong to the ΔG_{H+} case, 7 belong to the ΔG_{S+} case and 38 belong to the ΔG_{S-} case. The dotted line represents a ΔG value of −10 kcal mol^{-1}. A ΔG value of −10 kcal mol^{-1} can be obtained from all four cases of enthalpy and entropy contributions.

resistance in HIV-1 protease inhibitors has been carefully studied experimentally analyzing the thermodynamic features of binding for various inhibitors.[12] Compounds with enthalpy-dominated binding (cases ΔG_{H-} and ΔG_{H+}) may be selected for further optimization of binding since optimizing the entropic term is relatively easier than enthalpic optimization of binding energy.[13] Hence, in the early stages of drug discovery it may be advantageous to avoid inhibitors with a dominant entropy term and instead focus on inhibitors with enthalpy-dominated binding energies. Thus methods that could qualitatively identify cases of protein−ligand binding with dominant enthalpy or entropy terms using descriptors derived from the three-dimensional structures should be of significant interest in computational drug design.

Empirical scoring functions for predicting ΔG values of protein−ligand complexes have gained popularity owing to their simplicity and computational effectiveness. A majority of the empirical scoring functions estimate ΔG as a weighted sum of free energy components. The weights are usually obtained by partial least-squares regression fitting to a training data set of protein−ligand complexes.[14−17] Studies show that some of the underlying terms constituting ΔG reflect the enthalpy and entropy changes upon binding.[4,18−20] Despite the development of scoring functions to predict binding energies,[21−26] very few attempts have been made to predict the thermodynamic components underlying the binding affinities.[20,27−29] In addition to estimating ΔG, most of the existing scoring functions provide little or no information (either qualitative or quantitative) about ΔH and $T\Delta S$ values. It is possible that the same values of binding energies could have varying proportions of enthalpy and entropy. For example, a ΔG value of −10 kcal mol^{-1} could be obtained from all four cases of enthalpy and entropy contributions (Figure 1). Thus, inaccurate computation of the thermodynamic components could still yield accurate ΔG values. Hence, to obtain the correct thermodynamic profile of binding energy, it is necessary that both enthalpy and entropy terms are accurately predicted by the scoring functions.

With increasing availability of data from crystallographic and calorimetric experiments, it is feasible to use state-of-the-art machine learning methods such as support vector machines for predicting ΔG and its thermodynamic components. Support vector machines (SVMs) are kernel based approximators that can learn a variety of characteristics of the training data. SVM-based methods are applicable to classification and regression tasks.[30] They have been used for predicting beta-sheets, protein function classification, central nervous system permeability to drug molecules, pharmaceutical quantitative structure−activity relationships and protein−protein interactions,[31−36] but only a few studies using SVMs to estimate binding energies for protein ligand complexes have been reported[37−42] and none have been reported so far for estimating the enthalpy and entropy components of binding energies of protein−ligand complexes. Support vector machine regression, also known as support vector regression (SVR) differs from linear regression mainly in the kernel trick and the procedure in which the weights for the individual feature functions are derived. Unlike linear regression, SVR is robust and various nonlinear kernel functions such as polynomial, normalized polynomial and radial basis functions of the features can be explored.

The aim of this study is to compute ΔG including ΔH and $T\Delta S$ values of protein−ligand complexes using SVR prediction models trained with “off-the-shelf” structure-based descriptors obtained from three-dimensional structures of the complexes. Support vector regression scoring functions (SVR-SFs) are developed exploring the high dimensional linear and nonlinear functions of the descriptor values for estimating ΔG, ΔH, and $T\Delta S$ values. The protein−ligand descriptors (Table 1) are calculated using the methods in Schrödinger suite (Schrödinger LLC, New York) and Autodock[14] for protein−ligand complexes with crystallographic structural data and binding data from calorimetric experiments. On the basis of the descriptor groups (listed in Table 2), SVR-SFs are trained to compute the ΔG, ΔH, and $T\Delta S$ values. Furthermore, using the results from SVR-SFs, qualitative classification between cases of binding energies with dominant enthalpy terms or dominant entropy terms is attempted. To our knowledge, this is the first report of SVM-based methods for computationally characterizing the thermodynamic components of protein−ligand binding energies. Moreover, the SVR-SFs reported in this paper highlights the importance of ligand-based descriptors in computations of binding energies and their thermodynamic components.

■ MATERIALS AND METHODS

Data Set Collection. The protein−ligand complexes were obtained from PDBCAL[43] and SCORPIO[4] databases of protein−ligand complexes for which X-ray crystal structures and thermodynamic data are available. Initially, protein−ligand complexes with ambiguous thermodynamic data were excluded. A collection of 120 protein−ligand complexes was used for training and validating parameters of the SVR-SFs. Out of the 120 protein−ligand complexes in the training data set, 88 protein−ligand complexes have high crystallographic resolution (0 ≤ 2.0 Å) while the remaining 28 protein−ligand complexes are of medium crystallographic resolution (2 ≤ 2.5 Å) and 4

Table 1. Descriptors Calculated Using Various Methods

descriptor values	method
van der Waal's interaction energy, electrostatic interaction energy, solvent reaction field energy and cavitation energy	LIAISION
OPLS molecular mechanics energies, solvation model for polar solvation, nonpolar solvation term (composed of the nonpolar solvent accessible surface area and van der Waals interactions)	Prime MM-GBSA
valence interaction energy, van der Waals interaction energy, electrostatic interaction energy, solvation interaction energy, constraint interaction energy	EMBRACE
van der Waals interaction energy, Coulomb interaction energy, lipophilic term derived from hydrophobic grid potential, hydrophobic interactions, hydrogen-bonding term, metal binding term, rewards and penalties for various features, penalty for freezing rotatable bonds, site polar interactions in the active site, rewards for polar but non hydrogen-bonding atoms in a hydrophobic region.	GLIDE
van der Waals interaction energy, electrostatic interaction energy, desolvation potential, hydrogen-bonding term, number of rotatable bonds	Autodock
number of acceptor groups, acidic hydrogens, amide hydrogens, charged acceptor groups, charged donor groups, divalent oxygen atoms, donor groups, neutral acceptor groups, neutral amines, neutral donor groups, reactive groups, rings, aromatic rings, aliphatic rings, rotatable bonds, atoms, heteroaromatic rings, chiral centers, nonconjugated amine groups, amidine and guanidine groups, carboxylic acid groups, nonconjugated amide groups, nonhindered rotatable bonds, reactive functional groups	ligand-based descriptors (QIKPROP and LIGPARSE)
computed dipole moment of the molecule, square of the dipole moment divided by the molecular volume, index of cohesive interaction in solids and globularity descriptor	
total solvent accessible surface area (SASA) using a probe with a 1.4 Å radius, hydrophobic component of the SASA, hydrophilic component of the SASA, carbon and attached hydrogen component of the SASA, weakly polar component of the SASA	
predicted polarizability in cubic angstroms, hexadecane/gas partition coefficient, octanol/gas partition coefficient, water/gas partition coefficient, octanol/water partition coefficient, aqueous solubility	

Table 2. Grouping of Descriptors That Constitute Terms for the Scoring Functions

descriptor groups[a] for scoring functions	constituting terms
LIA and LIA*	Features from LIAISON
EMB and EMB*	Features from EMBRACE
MMGBSA and MMGBSA*	Features from Prime-MMGBSA
AD and AD*	Features from Autodock
GLI and GLI*	Features from GLIDE
GAD and GAD*	Features from GLIDE and Autodock
ALL and ALL*	Features from LIA, EMB, MMGBSA, AD and GLI
LBD	Only ligand-based descriptors

[a] Groups with * have the ligand-based descriptors in addition to the original descriptors.

protein−ligand complexes have low crystallographic resolution (2.5 ≤ 3.5 Å). The biological assemblies of protein−ligand complexes were downloaded from the Protein Data Bank.[44] The biological assemblies provide the most realistic arrangement of the protein−ligand complexes especially when multiple chains or subunits are present. In the biological assemblies of the protein−ligand complexes, all the protein subunits were retained, while only the most representative ligand chain was retained. The crystallographic water molecules were removed from the protein−ligand complexes and the amino acid residues were protonated using the protein preparation protocol in Schrödinger's suite. The crystallographic water molecules were removed mainly because it is difficult to identify the water molecules that are important for binding of a particular ligand. Also, it is not clear if all the protein−ligand complexes in the data set contain water molecules that are required for binding of the ligands. Removing the crystallographic water molecules from the protein−ligand complexes normalizes the training data set with respect to imbalances in the presence of water molecules. The missing atoms were added with the Prime module of Schrödinger's suite. Finally, the protein−ligand complexes were subjected to a restrained minimization with a maximum displacement of 0.3 Å using the Impref protein refinement procedure in Schrödinger's suite.

Descriptor Calculation. Descriptor values were generated using Autodock4.2.2.1[14] and Schrödinger's suite for molecular modeling. In the Schrödinger's suite, protein−ligand interaction descriptors are generated using LIAISON, EMBRACE, prime-MMGBSA, and GLIDE modules while the ligand-based descriptors are generated using QIKPROP and LIGPARSE modules. The descriptors from the Schrödinger's suite were generated after minimizing the structures with truncated Newton algorithm for maximum 1000 steps in the OPLS 2005 force field. During the optimization, residues within 4 Å from the ligand were flexible while the others were rigid. In Autodock, the ligands were minimized in presence of protein for a maximum 1000 steps using the local search options prior to the descriptor generation. The descriptor generation procedure was repeated with 500 and 1500 steps in the minimization options using both Autodock and Schrödinger's suite. The performance of SVR-SFs was compared with the descriptor sets generated by varying the minimization steps. Thus, sensitivity of the predictive models to minor changes in the coordinates of the protein−ligand complexes was studied. Details of the generated descriptors and the methods used are listed in Table 1. The descriptors were grouped based on the method of generation and combination of groups. The support vector regression scoring functions (SVR-SFs) are named according to the descriptor groups (listed in Table 2). Within each descriptor group, prior to training the SVR-SFs, principal component analysis (PCA) was performed reducing the number of descriptors but still preserving the variance in the data. The data was processed and analyzed with the Konstanz information miner (KNIME)[45] interface for data mining. On the basis of the experimental ITC data, the protein−ligand complexes in the training data set were assigned as cases having binding energy that is dominated by enthalpy (ΔG_{H}) or dominated by entropy (ΔG_{S}). Thus ΔG_{H} comprises cases $\Delta G_{\mathrm{H}-}$ and $\Delta G_{\mathrm{H}+}$, while ΔG_{S} comprises cases $\Delta G_{\mathrm{S}-}$ and $\Delta G_{\mathrm{S}+}$ illustrated in Figure 1. The case assignment of protein−ligand complexes to ΔG_{H} and ΔG_{S} was later used to measure the qualitative accuracies of the SVR-SFs in retrieving the corresponding cases.

Training and Parameter Optimization of SVR-SFs. WEKA[46] software for data mining was used for training, validation and evaluation of the SVR-SFs. The choice of kernel, kernel parameters and hyper-parameters significantly influence the performance of SVMs. Since the choice of kernel depends on the data set features and the noise, kernels with linear, polynomial, normalized polynomial and radial-basis functions were systematically explored. For each combination of kernel and kernel parameters, the hyper-parameters were chosen using cross-validation and grid search of hyper-parameters. 2-fold cross-validation method was used for training the SVR-SFs mainly for optimizing the kernel and hyper-parameters. In a 2-fold cross-validation method, the data was arbitrarily partitioned into two sets of equal size. The data-model was then trained with chosen parameters on the first set and validated on the second, then trained on the second set and validated on the first. The procedure was thus repeated on various combinations of the kernel and hyper-parameters and the combination that yielded the best correlation with target values in the training data were finally chosen.

Two alternative methods were explored for estimating the protein−ligand binding energies using SVR-SFs:

$$\Delta G_{\mathrm{estimated}} = \sum_{i=1}^{n} g_i \Phi(\mathrm{descriptor}_i) \tag{1}$$

$$\Delta G_{\mathrm{estimated}} = \Delta H_{\mathrm{estimated}} - T\Delta S_{\mathrm{estimated}} \tag{2}$$

where

$$\Delta H_{\mathrm{estimated}} = \sum_{i=1}^{n} h_i \Phi(\mathrm{descriptor}_i) \text{ and } T\Delta S_{\mathrm{estimated}} = \sum_{i=1}^{n} s_i \Phi(\mathrm{descriptor}_i)$$

g_i, h_i, and s_i are the weights of kernel functions. $\Phi(\mathrm{descriptor}_i)$ is the kernel function that could be linear, polynomial, normalized polynomial, radial-basis function or any other user defined kernel function of a descriptor. Following eqs 1 and 2, the descriptors groups that are listed Table 2 are used to develop SVR based scoring functions (SVR-SF) for estimating ΔG. The weights of the kernel functions g_i, h_i, and s_i are obtained by fitting the parameters with an improved version of the sequential minimal optimization algorithm[47,48] for training support vector machines.

Pearson's correlation coefficient (R_{p}) is the measure of linear correlation between two variables, here between the predicted and experimental ΔH, $T\Delta S$, and ΔG values. R_{p} values range between +1 and −1 inclusive. An R_{p} value of +1 implies that the predicted values exhibit a direct relationship with the experimental values that could be perfectly described by a linear equation while an R_{p} value of −1 implies that the predicted values exhibit an inverse relationship with the experimental values that could be described by a linear equation. R_{p} values close to zero indicate no correlation between the experimental and predicted values from the SVR-SFs. Thus, SVR-SFs with R_{p} values close to +1 or −1 could be considered as more accurate while R_{p} values close to zero indicate SVR-SFs with poor performance. R_{p} values have been used earlier in the literature as a measure of the performance of various scoring functions for subsets of protein−ligand complexes in the PDBBind database.[29,49,50] Hence, for the SVR-SFs, the training and validation performances are reported using R_{p} and standard deviation of error (SD) values. For evaluating the performance of qualitative classification of the protein−ligand binding according to domination of enthalpy or entropy, calculation of accuracy, recall, and specificity values were performed. Details on the performance metrics are provided in the Supporting Information.

Validation of ΔH, $T\Delta S$, and ΔG Predictions. To test the ΔH, $T\Delta S$, and ΔG predictions, an internal validation within the training data set was performed using 87 high resolution protein−ligand complexes as a training set and the remaining 32 complexes with medium and low crystallographic resolutions as a test data set. Additionally, the ΔH, $T\Delta S$ and ΔG predictions were tested on 25 protein−ligand complexes (Supporting Information) from the data set of Tang and Marshall et al.[29] (hereby referred to as the TM data set). These 25 protein−ligand complexes are not present in the current training data set. The internal validation and testing with protein−ligand complexes from the TM data set is performed to check the reliability of the predictive models in estimating the ΔH and $T\Delta S$ values. In addition, the performance of the developed SVR-SFs following eq-1 and eq-2 for calculating ΔG values were evaluated with four data sets of protein−ligand complexes with known ΔG/pK_{d} values from experiments. Initially, for evaluating the performance of the SVR-SFs, protein−ligand complexes from the PDBBind[51] version 2007 refined data set were used as an external evaluation data set. Furthermore, in order to compare the R_{p} values of the SVR-SFs with the R_{p} values of previously reported scoring functions, the protein−ligand complexes from the PDBBind version 2002, PDBBind version 2007 core, and PDBBind version 2009 refined data sets were used.[29,49] The protein−ligand complexes for performance evaluation were preprocessed and descriptor values were generated in the same manner as described earlier for the protein−ligand complexes in the training data set.

■ RESULTS AND DISCUSSION

Training the SVR-SFs. Initially, a unique and common outlier was observed for all the support vector regression scoring functions (SVR-SFs) after the training with 120 protein−ligand complexes in the data set. All the SVR-SFs based on the descriptor groups (Table 2) for estimating ΔG following eq 1 (see the materials and methods section) underestimated the ΔG value for 1MK5. The training data set consists of protein−ligand complexes whose experimental binding energies are densely distributed in the range from −15 to −3 kcal mol^{-1} and only one complex, PDB 1MK5, has an experimental ΔG value of −18 kcal mol^{-1}. A reason for the outlier might be the sparse occurrence of protein−ligand complexes with experimental ΔG values around −18 kcal mol^{-1} in the training data. However, it should be noted that the outlier is a streptavidin−biotin complex whose binding energy estimation is a challenge for many scoring functions and the current SVR-SFs are not exceptions. Considering the difficulty in modeling the binding energy during training, the outlier was later excluded from the data set and the training of SVR-SFs was performed with 119 protein−ligand complexes. The SVR-SFs showed similar correlation coefficients and standard error values using descriptor calculations performed with 500, 1000, and 1500 minimization steps for the training data set. The minimization procedure refines any close contacts and steric clashes that might be a result of crystallographic refinement and the preprocessing of the protein−ligand complexes. Thus the minimization is considered as an essential step prior to generation of the descriptors. Training of SVR-SFs with 87 protein−ligand complexes with high resolution (resolution ≤2.0 Å) structures had lower accuracy in estimating the ΔG values when compared with the training data set of 119 protein−ligand complexes with varying crystallographic resolutions.

Computation of Δ*G* Using SVR-SFs (Eq 1). The SVR-SFs have diverse combinations of kernel parameters because, for each SVR-SF, the parameters were optimized to yield the best R_p for the training data set. For the SVR-SFs trained to estimate ΔG using eq 1, the radial-basis function (RBF) kernel and the normalized polynomial (N-POLY) kernels had better performances when compared with linear and polynomial kernels. The SVR-SFs EMB, EMB*, GAD, GAD*, AD, AD*, ALL, and ALL* have optimal performances with RBF kernels while the remaining SVR-SFs had optimal performance with the N-POLY kernel. The SVR-SFs with RBF kernels have gamma values varying from 0.05 to 0.1 while SVR-SFs with normalized polynomial kernel have the polynomial degree varying from 2 to 5. Further details on the kernel parameters for the SVR-SFs are provided in the Supporting Information. The SVR-SFs EMB and LBD are observed to have the lowest R_p value for the protein−ligand complexes in the training data set (Table 3). The SVR-SFs ALL*, MMGBSA*, GLI*, and AD* have R_p values greater than 0.9 with experimental ΔG values for the protein−ligand complexes in the training data set (Table 3). The scoring functions showed significant improvement in R_p values for the training data set when the ligand-based descriptors were included. Moreover, the SVR-SFs with ligand-based descriptors included have lower standard deviations for the estimated binding energy values than the corresponding SVR-SFs that lack the ligand-based descriptor values.

For protein−ligand complexes in the training data set, the SVR-SFs including ligand-based descriptors show improved R_p values with minor differences in estimating binding energies when compared with the corresponding SVR-SFs excluding the ligand-based descriptor values. The SVR-SF with ligand-based descriptors alone has a training R_p value of 0.55, which is better when compared with the EMB SVR-SF that has descriptors from the EMBRACE methodology. However, the SVR-SF EMB* with the ligand-based descriptors has better training accuracy than EMB and LBD. Despite the decent R_p value of the LBD scoring function, the standard deviation of the predicted values is relatively high when compared with the other SVR-SFs. It should be noted that optimization of each SVR-SF's parameters through cross-validation within the training data set has facilitated achieving such high performance values. Thus the true performance could be better assessed by measuring the performances of the SVR-SF's with external evaluation data sets.

Table 3. Training Statistics of the Developed SVR-SFs for Estimating Δ*G* Following Eq 1[a]

SVR-SF	R_p[b]	SD[b]
LIA	0.75	1.07
LIA*	0.89	0.94
EMB	0.32	0.74
EMB*	0.69	1.30
MMGBSA	0.72	1.16
MMGBSA*	0.93	0.77
GAD	0.73	0.57
GAD*	0.83	1.00
GLI	0.74	1.10
GLI*	0.95	0.67
AD	0.83	1.20
AD*	0.92	0.83
ALL	0.77	0.83
ALL*	0.94	0.73
LBD	0.55	2.41

[a] R_p is Pearson's correlation coefficient between the experimental values of ΔG and the estimated values from the SVR-SFs following eq 1. SD is the standard deviation of error. [b] The training data set has 119 protein−ligand complexes.

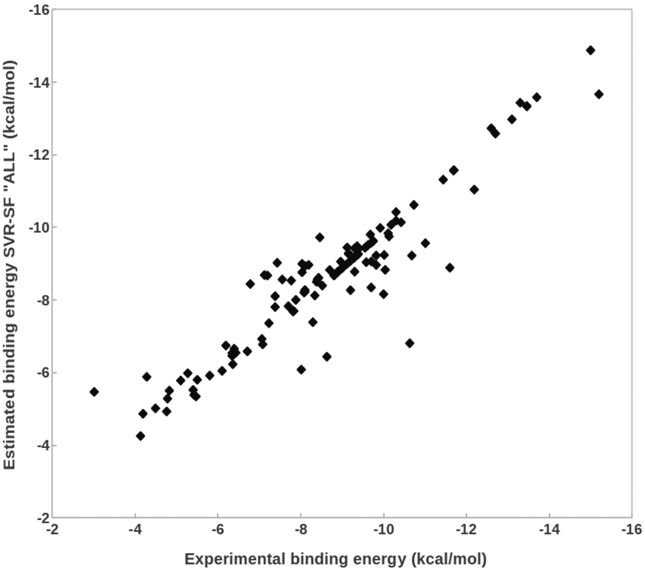


Figure 2. Scatter plot of support vector regression scoring function ALL* for estimating binding energy following eq 1 for protein−ligand complexes in the training data set. The ALL* SVR-SF has an R_p value of 0.94 and SD of 0.73 kcal mol^{-1} in the training data set.

All the SVR-SFs have significant differences in the R_p values for the training and the validation data sets (Tables 3 and 4). The internal validation with 87 high resolution protein−ligand complexes as training data set shows that the standard deviations are minimal for most of the predictive models in estimating the ΔG values. The final SVR-SFs trained with the data set of 119 protein−ligand complexes could predict the ΔG values for the protein ligand complexes in the TM data set with significant accuracy. The LIA*, GLI*, and AD* SVR-SFs had the highest correlations in predicting the ΔG values for the protein−ligand complexes in the TM data set. The SD values

Table 4. Performance Statistics of SVR-SFs in Estimating ΔG Following Eq 1[a]

SVR-SF	R_p[b]	SD[b]	R_p[c]	SD[c]	R_p[d]	SD[d]
LIA	**0.76**	1.13	0.52	1.92	0.50	2.01
LIA*	**0.90**	1.13	**0.61**	1.81	**0.65**	1.70
EMB	0.40	2.76	0.23	2.30	0.20	2.50
EMB*	**0.89**	1.60	0.53	2.00	**0.60**	1.81
MMGBSA	**0.75**	1.89	0.55	1.87	**0.60**	1.72
MMGBSA*	**0.86**	1.12	**0.63**	1.79	**0.67**	1.22
GAD	**0.73**	1.32	0.56	1.00	0.55	1.85
GAD*	**0.85**	1.21	**0.61**	1.86	**0.68**	1.33
GLI	**0.73**	1.30	0.57	1.91	0.50	1.92
GLI*	**0.95**	0.85	**0.64**	1.66	**0.67**	1.60
AD	**0.76**	1.41	0.54	1.89	0.55	1.89
AD*	**0.95**	0.88	**0.61**	1.85	**0.64**	1.53
ALL	**0.76**	1.32	0.58	1.57	0.56	1.60
ALL*	**0.88**	1.25	**0.63**	1.73	**0.86**	1.10
LBD	0.59	1.95	0.46	2.30	0.52	2.10

[a] R_p values ≥0.6 are shown in bold. [b] The external evaluation was performed with the protein−ligand complexes from TM data set. [c] The external evaluation was performed with 1300 protein−ligand complexes from the PDBBind version 2007 refined data set. [d] The internal evaluation was performed by splitting the original training data set of 119 protein−ligand complexes into 87 high resolution structures as training data set and remaining structures as a test data set.

for these SVR-SFs are less than 1.5 kcal mol^{-1} with GLI* having the lowest SD of 0.85 kcal mol^{-1} and highest R_p value of 0.95 for the protein−ligand complexes in the TM data set (Figure 3). For the PDBBind version 2007 refined data set, the SVR-SFs based on LIA*, MMGBSA*, GAD*, GLI*, AD*, and ALL* descriptor sets have the best performance. The SVR-SF EMB has the least R_p value of 0.23 while GLI* has the best R_p value of 0.64. The other SVR-SFs have performances with R_p ranging from 0.52 to 0.46 with significant differences in the performance values (Table 4). Further analysis of the estimated ΔG values from GLI* SVR-SF for the PDBBind version 2007 refined data set shows that the majority of visually noticeable outliers have the binding energy underestimated and the corresponding experimental ΔG values are mainly in the range of −15 to −18 kcal mol^{-1} (Figure 4). The absence of training samples in the range of −15 to −18 kcal mol^{-1} might be only a partial explanation because data points exist where the ΔG values are estimated accurately despite absence of corresponding ΔG values in the training data set. Out of the seven identified outliers (Figure 4), three are transferase-peptide complexes while the remaining four outliers belong to diverse families of protein−ligand complexes. Excluding the visual outliers from the test data set, the GLI* scoring function has an R_p value of 0.68.

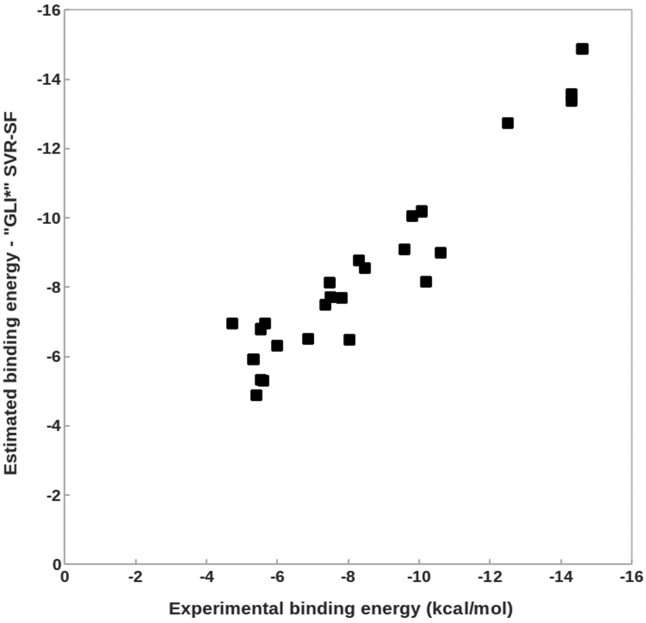


Figure 3. Scatter plot of support vector regression scoring function GLI* for estimating binding energies for the protein−ligand complexes from the TM data set. The GLI* SVR-SF has an R_p value of 0.95 for this data set in estimating ΔG values following eq 1.

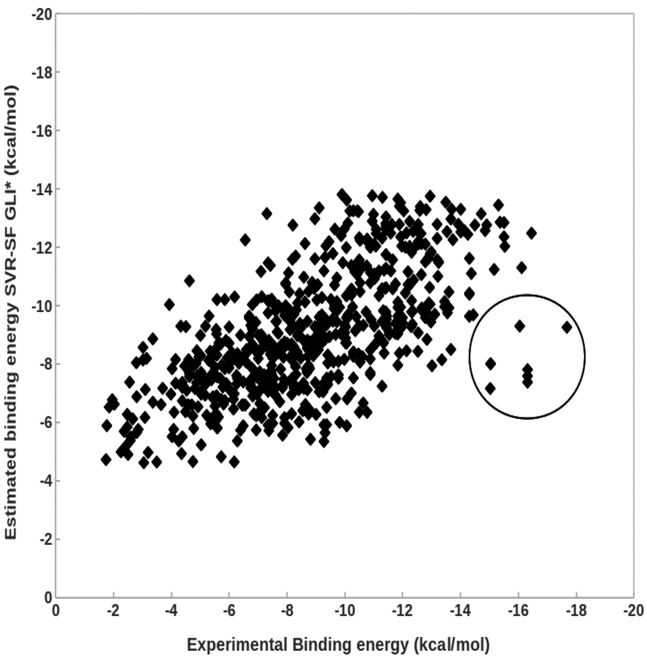


Figure 4. Scatter plot of ΔG estimates following eq-1 using SVR-SF constructed using the "GLI*" descriptor set for protein−ligand complexes from PDBBind v2007 refined data set. The SVR-SF "GLI*" has an R_p of 0.64 and SD of 1.66 kcal mol^{-1}. Excluding the circled outliers, the scoring function has an R_p of 0.68. Out of the seven identified outliers, three are transferase-peptide complexes while the rest four belong to diverse families of protein−ligand complexes.

It is observed in the internal and external validation studies that the SVR-SFs without the ligand-based descriptors had lower accuracy when compared with the SVR-SFs with ligand-based descriptors (Table 4). The inclusion of ligand-based descriptors seems to have increased the accuracy in predicting the ΔG values and lower standard deviation values. However, the SVR-SF LBD alone has an R_p value ranging from 0.46 to 0.59 for the protein−ligand complexes in the validation data sets (Table 4) and relatively high standard deviations in estimating ΔG values. Thus the SVR-SF with only ligand-based descriptors is less accurate than majority of the SVR-SFs except EMB. Thus comparing the SVR-SFs with and without ligand-based terms together with the performance of LBD alone, it could be concluded that the ligand-based descriptors when used with the other terms have a significant effect on the accuracy of the ΔG values following eq 1.

Computation of ΔH and $T\Delta S$ Values Using SVR-SFs (Eq 2). With the available thermodynamic data from calorimetric experiments, SVR-SFs were trained to estimate the ΔH and $T\Delta S$ values, which were then used to compute ΔG values. The optimization results of the SVR-SF parameters suggest that the RBF and N-POLY kernels had better performance than the linear and polynomial kernels. In estimating both ΔH and $T\Delta S$, the majority of the SVR-SFs had the best performances with the RBF kernel. The SVR-SFs for estimating ΔH were trained with the RBF kernel with gamma varying from 0.05 to 0.1. For estimating ΔH values, the SVR-SFs EMB, EMB*, MMGBSA, MMGBSA*, and LBD were trained with N-POLY kernel with exponent varying from 3 to 6. In estimating the $T\Delta S$ values, the SVR-SFs LIA, LIA*, and LBD have N-POLY kernels with the exponent varying from 2 to 6 while the remaining SVR-SFs have the RBF kernel with the gamma value varying from 0.05 to 0.1. Also for estimating $T\Delta S$, the SVR-SFs resulted in optimized performance with the RBF and N-POLY kernels with the majority having the RBF kernel. Complete details of the SVM kernels with the final optimized parameter values are provided in the Supporting Information. For the protein–ligand complexes in the training data set, the SVR-SF ALL* with RBF kernel have the best accuracy in estimating the thermodynamic components of the binding energy. The ALL* SVR-SF has the best R_p values of 0.93 and 0.94 for estimating ΔH and $T\Delta S$ respectively (Figure 5 and Table 5). The SVR-SFs apart from EMB had an R_p value of at least 0.6 for predicting ΔH and $T\Delta S$ values of the protein–ligand complexes in the training data set. For protein–ligand complexes in the training data set, the ΔG values computed from the predictions of ΔH and $T\Delta S$ values had very low standard error values (Table 5). With the SVR-SFs excluding ligand-based descriptor features, the predictions of the $T\Delta S$ values have slightly higher R_p values than the predictions of ΔH values. However, the SVR-SFs including ligand-based descriptor features had similar R_p values in estimating both ΔH and $T\Delta S$ values. The LBD SVR-SF has better correlation with the experimental $T\Delta S$ values than the ΔH values. The inclusion of ligand based descriptors significantly increased the accuracies in estimating both ΔH and $T\Delta S$ values. On comparing the ALL* and ALL SVR-SFs, the ALL* scoring function has better R_p values in estimating both ΔH and $T\Delta S$ values and lower standard errors. As mentioned earlier, the optimization of the SVR parameters might have been a primary reason behind the excellent correlations of the SVR-SFs with the experimental values of enthalpy and entropy for the protein–ligand complexes in the training data set.

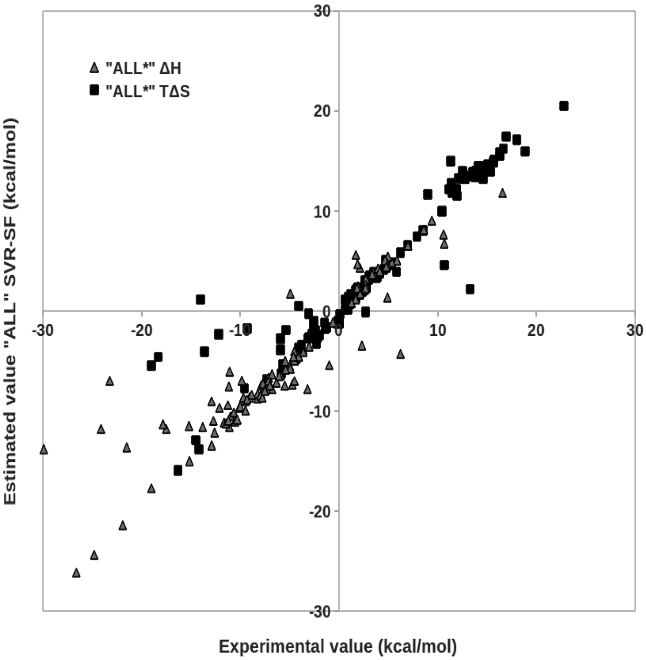


Figure 5. Scatter plot of ΔH and $T\Delta S$ estimates using SVR-SF ALL* with the experimental binding energies for protein–ligand in the training data set. The SVR-SF ALL* has an R_p of 0.93 and 0.94 in estimating ΔH and $T\Delta S$ values for the protein–ligand complexes training data set.

Table 5. Performance Statistics of SVR-SFs in Computing of ΔG Following Eq 2, ΔH, and $T\Delta S$ Values for Protein–Ligand Complexes in the Training Data Set[a]

SVR-SF	R_p (ΔH)	R_p ($T\Delta S$)	R_p (ΔG)	SD (ΔG)
LIA	0.66	0.69	0.55	1.47
LIA*	0.90	0.91	0.56	1.10
EMB	0.41	0.42	0.20	1.99
EMB*	0.90	0.90	0.55	1.18
MMGBSA	0.60	0.67	0.59	1.49
MMGBSA*	0.92	0.92	0.57	1.10
GAD	0.71	0.74	0.67	1.27
GAD*	0.92	0.93	0.62	1.07
GLI	0.65	0.68	0.66	1.27
GLI*	0.92	0.92	0.66	1.14
AD	0.65	0.68	0.46	0.67
AD*	0.91	0.92	0.57	1.14
ALL	0.81	0.84	0.68	1.50
ALL*	0.93	0.94	0.89	0.94
LBD	0.55	0.79	0.48	2.26

[a] R_p is the Pearson's correlation coefficient and SD is the standard deviation of error.

The reliability of the SVR-SFs in estimating the ΔH and $T\Delta S$ values was established through internal validation within the training data set and external validation with the protein–ligand complexes from the TM data set that have experimental values for ΔH and $T\Delta S$ (Tables 6 and 7). In the internal validation, the ALL scoring function had the best correlation with the experimental ΔH values, while the LBD SVR-SF had a very good correlation with the $T\Delta S$ values (Table 6). This supports the general assumption that the protein–ligand interaction descriptors could significantly capture the enthalpic contributions while the ligand based descriptors effectively capture the entropic contributions. For the protein–ligand complexes in the internal validation data set, the ALL* SVR-SF had very good correlations with experimental values of both enthalpy and entropy components. However, the best correlations with the binding energy are seen for LIA* and AD* SVR-SFs. Thus the internal validation shows that significant accuracies in estimating the ΔH and $T\Delta S$ values could be obtained with the SVR-SFs with a set of protein–ligand complexes within the training data set.

The estimations of ΔH and $T\Delta S$ values by the SVR-SFs in this study were further tested with the protein–ligand complexes from TM data set with experimental values of ΔH and $T\Delta S$ components of binding energy (Table 7 and Figure 6). The external validation results further supports the observation that the "LBD" SVR-SF has relatively better

Table 6. Performance Statistics of SVR-SFs in Computing of ΔG Following Eq 2 for Protein Ligand Complexes in the Internal Validation Data Set[a]

	ΔH	$T\Delta S$	ΔG (eq 2)	
SVR-SF	R_p	R_p	R_p	SD
LIA	0.59	0.40	0.58	2.01
LIA*	0.58	**0.70**	**0.70**	1.65
EMB	0.51	0.20	0.44	2.55
EMB*	0.55	0.56	0.59	1.56
MMGBSA	**0.62**	0.48	0.55	1.96
MMGBSA*	**0.60**	**0.65**	**0.61**	1.87
GAD	**0.60**	0.55	**0.60**	1.89
GAD*	0.51	**0.60**	**0.62**	1.80
GLI	0.58	0.45	0.59	1.88
GLI*	0.59	**0.72**	**0.61**	1.81
AD	0.60	0.40	0.55	1.85
AD*	0.59	**0.69**	**0.70**	1.74
ALL	**0.64**	0.50	0.50	1.90
ALL*	**0.60**	**0.72**	**0.64**	1.70
LBD	0.48	**0.69**	0.59	2.55

[a] Internal validation was performed with a training data set of 87 high resolution structures and the predictive models were tested on 32 medium and low resolution protein–ligand complexes. R_p values ≥0.6 are shown in bold.

Table 7. Performance Statistics of SVR-SFs in Computing of ΔG Following Eq 2 for the Protein–Ligand Complexes in External Test Data Sets[a]

	ΔH^b	$T\Delta S^b$	ΔG^b		ΔG^c	
SVR-SF	$R_p{}^b$	$R_p{}^b$	$R_p{}^b$	SD^b	$R_p{}^c$	SD^c
LIA	0.34	0.36	0.62	2.35	0.49	2.20
LIA*	**0.80**	**0.81**	**0.93**	1.08	0.56	1.72
EMB	0.3	0.3	0.4	2.60	0.14	2.22
EMB*	**0.7**	**0.74**	**0.85**	1.58	0.55	1.86
MMGBSA	0.28	0.5	0.65	2.28	0.53	2.10
MMGBSA*	**0.83**	**0.85**	**0.93**	1.08	0.57	1.75
GAD	0.56	0.67	**0.74**	2.00	0.59	1.75
GAD*	**0.8**	**0.93**	**0.94**	1.06	**0.62**	1.76
GLI	0.49	**0.64**	**0.71**	2.10	0.57	1.80
GLI*	**0.81**	**0.82**	**0.88**	1.26	**0.61**	1.80
AD	0.36	0.34	0.48	2.50	0.43	2.10
AD*	**0.77**	**0.79**	**0.91**	0.99	0.57	1.74
ALL	**0.63**	**0.75**	**0.71**	2.11	0.59	1.80
ALL*	**0.82**	**0.85**	**0.92**	1.17	**0.62**	1.70
LBD	**0.57**	**0.74**	**0.60**	1.51	0.49	2.55

[a] R_p values ≥0.6 are shown in bold. [b] The external evaluation was performed with the protein–ligand complexes from TM data set. [c] The evaluation was performed on data set with total 1300 protein ligand complexes from the PDBBind version 2007 refined data set. Only the ΔG values for these complexes from experiments are available.

accuracy in capturing the entropic contributions when compared with the SVR-SFs that lack the ligand based descriptors (Table 7). For the TM data set, the best accuracies in estimating the ΔG values were observed with the LIA*, MMGBSA*, GAD*, AD*, and ALL* SVR-SFs with R_p values greater than 0.9. The ΔG estimations following the eq 2 are tested for the protein–ligand complexes from the PDBBind version 2007 refined data set that have only the experimental ΔG values. For the protein–ligand complexes in the PDBBind version 2007 refined data set, the computed ΔG values using ALL* and GAD* SVR-SFs have the best R_p values of 0.62. In addition, the GLI* SVR-SF has an R_p value of 0.61 for the computed ΔG values for the protein–ligand complexes in the PDBBind version 2007 refined data set. As described earlier in computation of ΔG values following eq 1, the inclusion of the ligand-based descriptor features has also increased the accuracy in computation of ΔG values following eq 2 for the protein–ligand complexes in the evaluation data set. Thus it could be concluded that the ligand-based descriptors play significant role in improving the accuracies in calculation of the ΔH and $T\Delta S$ values for the protein–ligand complexes using the SVR-SFs. Further evaluation of ΔH and $T\Delta S$ prediction models would require the availability of more experimental data (e.g., from calorimetric studies).

Comparing the internal and external validations of SVR-SFs in estimating the ΔG values, it is observed that both schemes (eqs 1 and 2) are equally efficient. In some cases it is observed that the estimation of ΔG following eq 1 is more accurate when compared with eq 2. It should be noted that eq 1 involves optimizing one scoring function and the ΔG values are used as the target function for optimizing the parameters of SVR-SFs. On the other hand, eq 2 involves computing ΔG through two scoring functions that estimate ΔH and $T\Delta S$ values independently. Thus SVR-SFs estimating ΔG following eq 2 are optimized to predict the ΔH and $T\Delta S$ values and their results are used to compute the ΔG values.

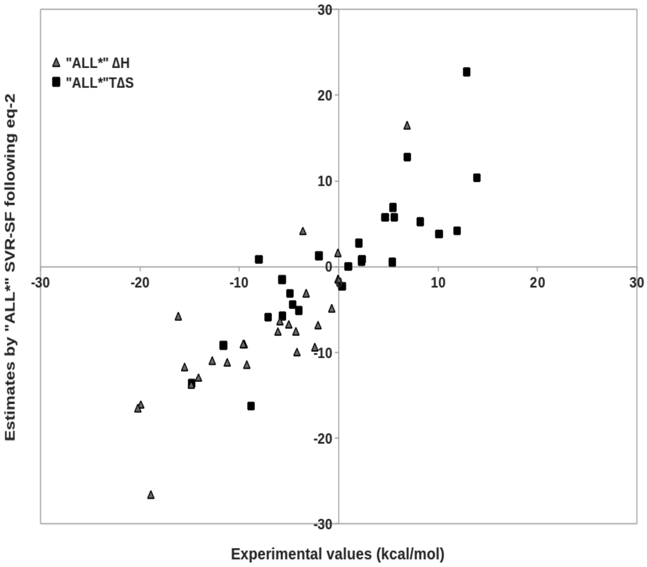


Figure 6. Scatter plot of ΔH and $T\Delta S$ estimates using SVR-SF ALL* with the experimental binding energies for the protein–ligand complexes from TM data set. The SVR-SF ALL* has an R_p of 0.82 and 0.85 in estimating ΔH and $T\Delta S$ values, respectively, and R_p of 0.92 for the computed ΔG values from ΔH and $T\Delta S$ estimates.

Qualitative Identification of Enthalpy or Entropy Domination in Binding Energy. In the drug discovery process, lead optimization involves improving the binding affinity and specificity of chosen lead compounds to a particular target protein. An important aspect in lead optimization process is the information about qualitative dominance of enthalpy or entropy in binding energy, ΔG_H and ΔG_S respectively. Knowledge of ΔG_H and ΔG_S for a protein–ligand binding is helpful in strategic selection of initial hits and further optimization of the binding energy to the target protein. For

Table 8. Comparison of R_p Values of the SVR-SFs with Benchmarks of Previously Reported Scoring Functions in Literature[29,49,50] for Protein−Ligand Complexes in PDBBind Version 2002 Data Set, PDBBind Version 2007 Core, and PDBBind Version 2009 Refined Data Sets[a]

scoring function	R_p PDBBind version 2007 core	R_p PDBBind version 2002	R_p PDBBind version 2009 refined
SVR-SF LIA	0.539 (0.514)	0.496 (0.485)	0.552 (0.519)
SVR-SF LIA*	0.564 (0.566)	0.585 (0.518)	**0.612** (0.591)
SVR-SF EMB	0.213 (0.201)	0.213 (0.200)	0.251 (0.225)
SVR-SF EMB*	0.451 (0.480)	0.512 (0.514)	0.540 (0.542)
SVR-SF MMGBSA	0.560 (0.572)	0.539 (0.548)	0.582 (0.558)
SVR-SF MMGBSA*	0.585 (0.569)	0.535 (0.536)	**0.620 (0.602)**
SVR-SF GAD	0.578 **(0.620)**	0.548 (0.589)	0.590 **(0.612)**
SVR-SF GAD*	0.590 **(0.646)**	0.576 **(0.602)**	**0.612 (0.603)**
SVR-SF GLI	0.585 **(0.603)**	0.543 (0.573)	**0.615** (0.592)
SVR-SF GLI*	0.585 **(0.654)**	0.610 (0.601)	**0.630 (0.600)**
SVR-SF AD	0.556 (0.592)	0.502 (0.402)	0.531 (0.552)
SVR-SF AD*	0.570 (0.492)	0.585(0.522)	0.601 (0.566)
SVR-SF ALL*	**0.600 (0.667)**	**0.610 (0.612)**	**0.630 (0.652)**
SVR-SF ALL	0.595 (0.639)	0.555 (0.593)	0.581 **(0.607)**
PHOENIX	NA **(0.616)**	NA (0.524)	NA (0.575)
SFCscore::met	NA	0.585	NA
X-Score::HPScore	NA	0.514	0.571
X-Score::HMScore	**0.644**	0.566	0.563
X-Score::HSScore	NA	0.506	0.565
DrugScore::Pair	NA	0.473	NA
DrugScoreCSD	0.569	NA	NA
DrugScore::Surf	NA	0.463	NA
DrugScore::Pair/Surf	NA	0.476	NA
Sybyl::D-Score	0.392	0.322	NA
Sybyl::PMF-Score	0.268	0.147	NA
Sybyl::G-Score	NA	0.443	NA
Sybyl::ChemScore	0.555	0.499	NA
Sybyl::F-Score	0.216	0.141	NA
DS::PLP1	0.545	NA	NA
Cerius2::LigScore	NA	0.406	NA
Cerius2::PLP1	NA	0.458	NA
Cerius2::PLP2	NA	0.455	NA
Cerius2::PMF	NA	0.253	NA
Cerius2::LUDI1	NA	0.334	NA
Cerius2::LUDI2	NA	0.379	NA
Cerius2::LUDI3	NA	0.331	NA
GOLD::GoldScore	0.295	0.285	NA
GOLD::GoldScore_opt	NA	0.365	NA
GOLD::ChemScore	0.441	0.423	NA
GOLD::ChemScore_opt	NA	0.449	NA
GOLD::ASP	0.534	NA	NA
GlideScore	0.457	NA	NA
HINT	NA	0.33	NA
DS::Jain	0.316	NA	NA

[a] The values in parentheses are the R_p values for scoring functions in computation of ΔG values following eq 2. The remaining are the R_p values in computation of ΔG values following eq 1. NA means data not available. R_p values ≥0.6 are shown in bold.

this reason, using the estimated ΔH and $T\Delta S$ values from SVR-SFs, the dominance of the thermodynamic terms was qualitatively analyzed and compared with the existing experimental data. The accuracy, recall and specificity values (Supporting Information) for GAD*, GLI*, ALL*, and ALL were calculated for the protein−ligand complexes in the training data set and TM data set. Recall and specificity values are measures of "true ΔG_H" prediction rate and "true ΔG_s" prediction rates respectively. The GAD*, GLI*, and ALL* SVR-SFs have higher overall accuracy when compared with the ALL SVR-SF in discriminating the dominance of enthalpy or entropy in the binding energy for the protein−ligand complexes in the training data set. The recall and specificity values for the training data set further indicate that all the true ΔG_H cases were identified accurately by GAD* SVR-SF, while all the true ΔG_S cases were identified accurately by GLI* and ALL* SVR-SFs. Out of the 25 protein ligand complexes from the TM data set, 9 have a binding profile of ΔG_H, while the remaining 16 have a binding profile of ΔG_S. The ALL* SVR-SF exhibited an accuracy of 85% in discriminating between ΔG_H and ΔG_S cases of the TM data set. With the ALL* SVR-SF the true ΔG_H cases could be identified with a precision of 100%

while the ΔG_S cases were identified with a recall value of 75%. In the TM data set, the protein–ligand complex with PDB 1UAE has an experimental ΔG value of −6 kcal mol^{-1}, ΔH and $T\Delta S$ values of 6.88 and 12.87 kcal mol^{-1}, respectively. This protein–ligand interaction has unfavorable enthalpy and favorable entropy resulting in entropy dominated binding with ΔG_{s-} binding profile. The ALL* SVR-SF could successfully predict the ΔG_{s-} binding profile in addition to the computed binding energy of −6.2 kcal mol^{-1} from the ΔH and $T\Delta S$ estimates. The performance metrics of the SVR-SFs with and without the ligand-based descriptors shows that the ligand-based descriptor features have significant importance in identifying both ΔG_H and ΔG_S cases as supported by comparing the accuracy, recall and specificity values.

Performance Comparisons of SVR-SFs with Benchmarks of Existing Scoring Functions. The SVR-SFs were tested in estimating the ΔG values for the protein–ligand complexes in PDBBind version 2002, PDBBind version 2007 core, and PDBBind version 2009 refined data sets. The performance of several scoring functions with these data sets has been previously reported in the literature. In the PDBBind version 2002 data set, the descriptors were successfully generated for 799 protein–ligand complexes. In the PDBBind version 2007 core data set the descriptors were successfully generated for 195 protein–ligand complexes and in the PDBBind version 2009 refined data set the descriptors were successfully generated for 1625 protein–ligand complexes. Some of the SVR-SFs using eqs 1 and 2 have better correlation values in estimating ΔG values than most of the previously reported scoring functions for the PDBBind version 2002 and PDBBind version 2007 core and PDBBind version 2009 refined data sets. Among the SVR-SFs, the ALL* SVR-SF has the best R_p value of 0.667 and 0.612 in estimating the ΔG values (Table 8) following eq 2 for protein–ligand complexes in the PDBBind version 2007 core and PDBBind version 2002 data sets, respectively. The X-Score::HMScore[23] has a correlation coefficient of 0.644 and 0.566 for protein–ligand complexes in the PDBBind version 2007 core and PDBBind version 2002 data sets, respectively. The ALL* and GLI* SVR-SFs have excellent correlations with the experimental binding energies for the protein–ligand complexes from the PDBBind version 2009 refined data set. For the protein–ligand complexes in the PDBBind version 2009 refined data set, the LIA*, MMGBSA*, GAD*, GLI, GLI*, AD*, and ALL* SVR-SFs have correlation values better than the PHOENIX and X-score scoring functions. The X-Score::HMScore estimates the ΔG values following eq 1, which has the general disadvantage of lack of information about ΔH values and $T\Delta S$ values when compared to ΔG computation following eq 2. PHOENIX scoring function[29] estimates ΔG values based on prediction of thermodynamic components and has R_p values of 0.616, 0.524, and 0.575 for protein–ligand complexes in the PDBBind version 2007 core, PDBBind version 2002, and PDBBind version 2009 refined data sets respectively. An advantage of the currently reported SVR-SFs is that the majority of the descriptors are generated using Autodock and Schödinger's suite of computational chemistry programs that makes data collection and management less complex. For the protein–ligand complexes in the PDBBind version 2007 core, version 2002, and version 2009 refined data sets, the SVR-SF based on ALL* descriptor data set following eq 2 has the best R_p of 0.667, 0.612, and 0.652 respectively in estimating the ΔG values, which are the highest R_p values in predicting the ΔG values when compared with the rest of the scoring functions including PHOENIX. The SVR-SFs developed following eq 1 share the basic methodology with many of previously reported scoring functions by which the ΔG values are directly estimated as a function of the descriptors. The SVR-SFs developed following eq 2 are completely different from the other existing scoring functions (except the PHOENIX scoring function) where the ΔG values are calculated from ΔH values and $T\Delta S$ values that are estimated as a function of the descriptors. However, the major differences between the SVR-SFs from all the existing scoring functions are the choice of kernel/function and the method of fitting the weights to each of the contributing functions. Moreover, the variants of the current SVR-SFs with included ligand-based features have extensive ligand-based descriptors in addition to the intermolecular interaction descriptors. This is apparently a reason why some of the SVR-SFs reported here outperform existing scoring functions in benchmarks of estimating ΔG values for protein–ligand complexes from PDBBind version 2002, PDBBind version 2007 core, and PDBBind version 2009 refined data sets. It can be concluded in general that computation of ΔG following eq 2 is more useful since it provides valuable information about the thermodynamic components of binding energy.

■ CONCLUSIONS

Support vector machines can be used for computing the binding energies and their thermodynamic components of protein–ligand complexes. Binding energy (ΔG) estimates obtained using support vector regression scoring functions (SVR-SFs) show good correlations with experimental data and some of the SVR-SFs perform better than existing scoring functions in benchmarks based on protein–ligand complexes taken from the PDBBind version 2002, PDBBind version 2007 core, and PDBBind version 2009 refined data sets. In addition to estimating the ΔG values, the SVR-SFs can also estimate the thermodynamic components ΔH and $T\Delta S$ with significant accuracy. SVR-SFs with protein–ligand interaction based descriptors have significant correlations in estimating both ΔH and $T\Delta S$ values, while the inclusion of ligand-based descriptors improves the correlations mainly in estimating the $T\Delta S$ values. It is concluded that ligand-based descriptors are crucial for accurate computational estimation of ΔG, ΔH, and $T\Delta S$ values. The qualitative identification of ΔH or $T\Delta S$ dominance in binding energies can be accurately identified by the SVR-SFs. Depending on the extent of details, qualitative and quantitative results could be used for identifying the thermodynamic profiles of binding as well as their absolute values. The qualitative classification using SVR-SFs could be used to postfilter virtual screening results based on their predicted thermodynamic profile, while the quantitative computational prediction models could be used for precise estimation of enthalpy and entropy components during lead optimization. Binding energies estimated using SVR-SFs following eq 1 are seen to be slightly more accurate than binding energies estimated following eq 2. However, the computation of ΔG values from the ΔH and $T\Delta S$ estimates (eq 2) is more useful than computing ΔG values directly from the descriptors (eq 1) because of the advantage that the information about thermodynamic components is obtained. These results suggest that support vector machines could be of great value in computational drug design.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

The list of PDB codes in the training and TM datasets, details of the performance metrics and the parameters of SVR-SFs. This information is available free of charge via the Internet at http://pubs.acs.org

■ AUTHOR INFORMATION

Corresponding Authors

*E-mail: kemp@chalmers.se. Telephone: +46-(0)31 772 54 11. Fax: +46-(0)31 772 36 63.

*E-mail: per-georg.nyholm@biognos.se. Telephone: +46(0)31 65 62 40. Fax: +46(0)31 23 66 44.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

Financial support and resources from Biognos AB are gratefully acknowledged. The authors thank Dr. Chrianjib Battacharya, Department of computer science and automation, Indian Institute of Science, Bangalore, India for discussions on support vector machines.

■ REFERENCES

(1) Garbett, N. C.; Chaires, J. B. Thermodynamic studies for drug design and screening. *Expert Opin. Drug Discovery* **2012**, *7*, 299−314.

(2) Ferenczy, G. G.; Keseru, G. M. Thermodynamics guided lead discovery and optimization. *Drug Discovery Today* **2010**, *15*, 919−932.

(3) Jelesarov, I.; Bosshard, H. R. Isothermal titration calorimetry and differential scanning calorimetry as complementary tools to investigate the energetics of biomolecular recognition. *J. Mol. Recognit.* **1999**, *12*, 3−18.

(4) Olsson, T. S.; Williams, M. A.; Pitt, W. R.; Ladbury, J. E. The thermodynamics of protein−ligand interaction and solvation: insights for ligand design. *J. Mol. Biol.* **2008**, *384*, 1002−1017.

(5) Genheden, S.; Mikulskis, P.; Hu, L.; Kongsted, J.; Soderhjelm, P.; Ryde, U. Accurate predictions of nonpolar solvation free energies require explicit consideration of binding-site hydration. *J. Am. Chem. Soc.* **2011**, *133*, 13081−13092.

(6) Beuming, T.; Che, Y.; Abel, R.; Kim, B.; Shanmugasundaram, V.; Sherman, W. Thermodynamic analysis of water molecules at the surface of proteins and applications to binding site prediction and characterization. *Proteins* **2012**, *80*, 871−883.

(7) Lafont, V.; Armstrong, A. A.; Ohtaka, H.; Kiso, Y.; Amzel, L. M.; Freire, E. Compensating enthalpic and entropic changes hinder binding affinity optimization. *Chem. Biol. Drug Des.* **2007**, *69*, 413−422.

(8) Exner, O. Entropy−enthalpy compensation and anticompensation: solvation and ligand binding. *Chem. Commun.* **2000**, 1655−1656.

(9) Gallicchio, E.; Kubo, M. M.; Levy, R. M. Entropy−enthalpy compensation in solvation and ligand binding revisited. *J. Am. Chem. Soc.* **1998**, *120*, 4526−4527.

(10) Gilli, P.; Ferretti, V.; Gilli, G.; Borea, P. A. Enthalpy−entropy compensation in drug-receptor binding. *J. Phys. Chem.* **1994**, *98*, 1515−1518.

(11) Locarnini, S.; Bowden, S. Drug resistance in antiviral therapy. *Clin. Liver Dis.* **2010**, *14*, 439−459.

(12) King, N. M.; Prabu-Jeyabalan, M.; Bandaranayake, R. M.; Nalam, M. N.; Nalivaika, E. A.; Ozen, A.; Haliloglu, T.; Yilmaz, N. K.; Schiffer, C. A. Extreme entropy−enthalpy compensation in a drug-resistant variant of HIV-1 protease. *ACS Chem. Biol.* **2012**, *7*, 1536−1546.

(13) Freire, E. Do enthalpy and entropy distinguish first in class from best in class? *Drug Discovery Today* **2008**, *13*, 869−874.

(14) Morris, G. M.; Huey, R.; Lindstrom, W.; Sanner, M. F.; Belew, R. K.; Goodsell, D. S.; Olson, A. J. AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. *J. Comput. Chem.* **2009**, *30*, 2785−2791.

(15) Verdonk, M. L.; Cole, J. C.; Hartshorn, M. J.; Murray, C. W.; Taylor, R. D. Improved protein−ligand docking using GOLD. *Proteins* **2003**, *52*, 609−623.

(16) Laederach, A.; Reilly, P. J. Specific empirical free energy function for automated docking of carbohydrates to proteins. *J. Comput. Chem.* **2003**, *24*, 1748−1757.

(17) Head, R. D.; Smythe, M. L.; Oprea, T. I.; Waller, C. L.; Green, S. M.; Marshall, G. R. VALIDATE: A new method for the receptor-based prediction of binding affinities of novel ligands. *J. Am. Chem. Soc.* **1996**, *118*, 3959−3969.

(18) Lee, B. Enthalpy−entropy compensation in the thermodynamics of hydrophobicity. *Biophys. Chem.* **1994**, *51*, 271−277 discussion 277−278.

(19) Urquiza, M.; Guevara, T.; Rodriguez, C.; Melo-Cardenas, J.; Vanegas, M.; Patarroyo, M. E. Decreasing the configurational entropy and the hydrophobicity of EBV-derived peptide 11389 increased its antigenicity, immunogenicity and its ability of inducing IL-6. *Amino Acids* **2012**, *42*, 2165−2175.

(20) Ruvinsky, A. M. Role of binding entropy in the refinement of protein-ligand docking predictions: Analysis based on the use of 11 scoring functions. *J. Comput. Chem.* **2007**, *28*, 1364−1372.

(21) Eldridge, M. D.; Murray, C. W.; Auton, T. R.; Paolini, G. V.; Mee, R. P. Empirical scoring functions 0.1. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. *J. Comput.-Aided Mol. Des.* **1997**, *11*, 425−445.

(22) Cheney, D. L.; Mason, J. S. Semiempirical scoring functions for ligand binding based on molecular mechanics and continuum model calculations. *Abstr. Pap. Am. Chem. Soc.* **2000**, *219*, U600−U600.

(23) Wang, R. X.; Lai, L. H.; Wang, S. M. Further development and validation of empirical scoring functions for structure-based binding affinity prediction. *J. Comput.-Aided Mol. Des.* **2002**, *16*, 11−26.

(24) Puvanendrampillai, D.; Marsden, P. M.; Mitchell, J. B. O.; Glen, R. C. Comparative evaluation of five scoring functions for accurate prediction of protein-ligand binding energy. *Abstr. Pap. Am. Chem. Soc.* **2004**, *227*, U1018−U1018.

(25) Zhong, S. J.; MacKerell, A. D. Novel scoring functions for in silico database screening: Binding response and pose-based scaling. *Abstr. Pap. Am. Chem. Soc.* **2005**, *230*, U1307−U1307.

(26) Rahaman, O.; Estrada, T. P.; Doren, D. J.; Taufer, M.; Brooks, C. L.; Armen, R. S. Evaluation of several two-step scoring functions based on linear interaction energy, effective ligand size, and empirical pair potentials for prediction of protein−ligand binding geometry and free energy. *J. Chem. Inf. Model.* **2011**, *51*, 2047−2065.

(27) Huang, S. Y.; Zou, X. Inclusion of solvation and entropy in the knowledge-based scoring function for protein-ligand interactions. *J. Chem. Inf. Model.* **2010**, *50*, 262−273.

(28) Ruvinsky, A. M. What is the statistical-thermodynamic cost of binding entropy in protein−ligand docking and virtual screening? Analysis based on the use of 11 scoring functions. *J. Biomol. Struct. Dyn.* **2007**, *24*, 767−768.

(29) Tang, Y. T.; Marshall, G. R. PHOENIX: A scoring function for affinity prediction derived using high-resolution crystal structures and calorimetry measurements. *J. Chem. Inf. Model.* **2011**, *51*, 214−228.

(30) Cortes, C.; Vapnik, V. Support-vector networks. *Mach. Learn.* **1995**, *20*, 273−297.

(31) Zavaljevski, N.; Stevens, F. J.; Reifman, J. Support vector machines with selective kernel scaling for protein classification and identification of key amino acid positions. *Bioinformatics* **2002**, *18*, 689−696.

(32) Cai, Y. D.; Liu, X. J.; Xu, X. B.; Chou, K. C. Support vector machines for prediction of protein domain structural class. *J. Theor. Biol.* **2003**, *221*, 115−120.

(33) Cai, Y. D.; Lin, S. L. Support vector machines for predicting rRNA-, RNA-, and DNA-binding proteins from amino acid sequence. *Biochim. Biophys. Acta* **2003**, *1648*, 127−133.

(34) Doniger, S.; Hofmann, T.; Yeh, J. Predicting CNS permeability of drug molecules: Comparison of neural network and support vector machine algorithms. *J. Comput. Biol.* **2002**, *9*, 849−864.

(35) Burbidge, R.; Trotter, M.; Buxton, B.; Holden, S. Drug design by machine learning: Support vector machines for pharmaceutical data analysis. *Comput. Chem.* **2001**, *26*, 5−14.

(36) Bock, J. R.; Gough, D. A. Predicting protein−protein interactions from primary structure. *Bioinformatics* **2001**, *17*, 455−460.

(37) Kinnings, S. L.; Liu, N. N.; Tonge, P. J.; Jackson, R. M.; Xie, L.; Bourne, P. E. A machine learning-based method to improve docking scoring functions and its application to drug repurposing. *J. Chem. Inf. Model.* **2011**, *51*, 408−419.

(38) Li, L.; Wang, B.; Meroueh, S. O. Support vector regression scoring of receptor−ligand complexes for rank-ordering and virtual screening of chemical libraries. *J. Chem. Inf. Model.* **2011**, *51*, 2132−2138.

(39) Li, L. W.; Khanna, M.; Jo, I. H.; Wang, F.; Ashpole, N. M.; Hudmon, A.; Meroueh, S. O. Target-specific support vector machine scoring in structure-based virtual screening: Computational validation, in vitro testing in kinases, and effects on lung cancer cell proliferation. *J. Chem. Inf. Model.* **2011**, *51*, 755−759.

(40) Fukunishi, Y. Structure-based drug screening and ligand-based drug screening with machine learning. *Comb. Chem. High Throughput Screening* **2009**, *12*, 397−408.

(41) Hecht, D.; Fogel, G. B. Computational intelligence methods for docking scores. *Curr. Comput.-Aided Drug Des.* **2009**, *5*, 56−68.

(42) Bock, J. R.; Gough, D. A. A new method to estimate ligand−receptor energetics. *Mol. Cell. Proteomics* **2002**, *1*, 904−910.

(43) Li, L. W.; Dantzer, J. J.; Nowacki, J.; O'Callaghan, B. J.; Meroueh, S. O. PDBcal: A comprehensive dataset for receptor-ligand interactions with three-dimensional structures and binding thermodynamics from isothermal titration calorimetry. *Chem. Biol. Drug Des.* **2008**, *71*, 529−532.

(44) Berman, H.; Henrick, K.; Nakamura, H. Announcing the worldwide Protein Data Bank. *Nat. Struct. Biol.* **2003**, *10*, 980−980.

(45) Berthold, M. R.; Cebron, N.; Dill, F.; Gabriel, T. R.; Kotter, T.; Meinl, T.; Ohl, P.; Sieb, C.; Thiel, K.; Wiswedel, B. KNIME: The Konstanz Information Miner. *Stud. Class Data Anal.* **2008**, 319−326.

(46) Hall, M. F., E.; Holmes, G.; Pfahringer, B.; Reutemann, P.; Witten, I. H. The WEKA data mining software: An update. *SIGKDD Explor.* **2009**, *11*, 5−10.

(47) Platt, J. C. Fast training of support vector machines using sequential minimal optimization In *Advances in Kernel Methods*; MIT Press: Cambridge, MA, U.S.A., 1999; pp 185−208.

(48) Keerthi, S. S.; Shevade, S. K.; Bhattacharyya, C.; Murthy, K. R. K. Improvements to Platt's SMO algorithm for SVM classifier design. *Neural Comput.* **2001**, *13*, 637−649.

(49) Wang, R. X.; Lu, Y. P.; Fang, X. L.; Wang, S. M. An extensive test of 14 scoring functions using the PDBbind refined set of 800 protein−ligand complexes. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 2114−2125.

(50) Wang, R. X.; Lu, Y. P.; Wang, S. M. Comparative evaluation of 11 scoring functions for molecular docking. *J. Med. Chem.* **2003**, *46*, 2287−2303.

(51) Wang, R. X.; Fang, X. L.; Lu, Y. P.; Yang, C. Y.; Wang, S. M. The PDBbind database: Methodologies and updates. *J. Med. Chem.* **2005**, *48*, 4111−4119.