JOURNAL OF CHEMICAL INFORMATION AND MODELING

EDITORIAL

pubs.acs.org/jcim

Comments on "Leave-Cluster-Out Cross-Validation Is Appropriate for Scoring Functions Derived from Diverse Protein Data Sets": Significance for the Validation of Scoring Functions

Recently, Kramer and Gedeck published an article[1] in this journal which referred extensively to our previous work on the scoring function RF-Score.[2] This machine learning-based scoring function is designed to predict the binding affinities of protein−ligand complexes and thus can also be used to rescore poses as generated by in silico docking techniques (however this first version of RF-Score is not suitable to guide pose generation). Whereas other scoring functions assume a particular mathematical relationship between the atomic-level description of the protein−ligand complex and various theory inspired contributions to binding free energy, we used random forest machine learning instead to implicitly infer this relationship in an entirely data-driven manner. Such a relationship typically takes the form of a sum of physicochemical contributions to binding in the case of empirical scoring functions or a reverse Boltzmann methodology in the case of knowledge-based scoring functions. Our unconstrained approach was likely to result in performance improvement, as it is well-known[3] that the strong assumption of a predetermined functional form for a scoring function introduces an error in addition to that inherent in all feasible methodologies for high-throughput binding affinity estimation (such inherent sources of error include the coarse description of protein flexibility and implicit treatment of solvent). Indeed, scoring functions using highly flexible machine learning for nonlinear regression are already showing[2,4,5] substantial performance improvements in comparison to established scoring functions. RF-Score, which is the only stand-alone open source scoring function we are aware of, is available to all[6] without charge on a Creative Commons license,[7] requiring only the free statistical software suite R,[8] a C language compiler, and the PDBbind[9,10] complexes to be used for training (step-by-step instructions are included in the Supporting Information in our article).[2] Our intention on releasing the RF-Score software was to facilitate the use, analysis, and future development of machine learning-based scoring functions.

Kramer and Gedeck explained that their interest in RF-Score was due to the following facts: It outperformed all other standard scoring functions on the 2007 PDBbind core set benchmark introduced by Cheng et al.;[11] it represents an entirely new class of scoring functions; and it was calibrated on the largest set of complexes used to date for this purpose. When validating RF-Score, we followed the careful validation protocol used by Cheng et al. on a widely representative set of scoring functions, using the 2007 PDBbind core data set as the test set (a total of 195 complexes). The training set, used to calibrate RF-Score, consists of the 1105 protein−ligand complexes belonging to the PDBbind refined set that are not part of the core (test) set. It is worth noting that this type of refined set partitions is common in the literature (see for instance models I and V in a recent study[4] by Breneman and co-workers in addition to X-Score::HMScore v1.3 in Cheng et al. and RF-Score in our study). A consequence of the experimental design is that the proportion of complexes in the test set with training complexes belonging to the same sequence-derived cluster (loosely speaking, protein family) will be higher than that occurring in a random split of the refined set.

Kramer and Gedeck's contribution is centered on a discussion of suitable validation methodologies for generic scoring functions, such as RF-Score and also those analyzed by Cheng et al. using the PDBbind benchmark. The thesis of these authors appears to be that "it is necessary to employ . . . leave-cluster-out cross-validation".[1] This is a cross-validation procedure, leave-cluster-out cross-validation (LCOCV), in which for each cluster, the data set is partitioned in such a way that the model on which any given instance is tested has been trained on data that excludes all members of the same cluster, and thereafter model performance is averaged across the considered training/test partitions. Much of the thrust of their work is to try to deconvolute the part of the performance of RF-Score that comes from identifying features common to a sequence-derived cluster from that part that is learned from other training complexes. This is an interesting hypothetical exercise suggesting that a substantial part of RF-Score's predictive power comes from cluster-specific information, although it is clear from the results that this model also exploits the information contained in other instances. Indeed, their Table 2 shows that in 50/62 cases, RF-Score either gets the within-cluster order of the three test binding affinities exactly correct (30 times) or else just flips the median affinity rank with one of the others.

Remarkably, RF-Score was the only scoring function used in this work. The lack of any performance results on other functions means that Kramer and Gedeck's study in itself tells us nothing about how RF-Score compares with other methods for binding affinity prediction. We agree that LCOCV is a more exacting test, against which one would expect a scoring function probably to generate a higher (worse) error between measured and predicted binding affinity, as quantified by root-mean-squared error (RMSE) or an almost identical standard deviation and by a lower (worse) Pearson correlation coefficient R. Table 7 in that study shows that the deterioration in RMSE on moving from PDBbind core validation to LCOCV is actually insignificant, from 1.58 to 1.60 log units, though R drops significantly from 0.77 to 0.46. However, there is no reason to believe that RF-Score should do relatively better or worse than any other scoring function on LCOCV versus PDBbind core validation. In other words, alternative scoring functions could perform worse on LCOCV if only these have been tested. In fact, when we compared the best performing scoring function in the Cheng et al. study (X-Score::HMScore) against RF-Score using exactly the same training and test sets, it was observed that RF-Score obtained a substantially better correlation (R = 0.78 versus 0.65)

Published: May 18, 2011

dx.doi.org/10.1021/ci200057e | J. Chem. Inf. Model. 2011, 51, 1739–1741

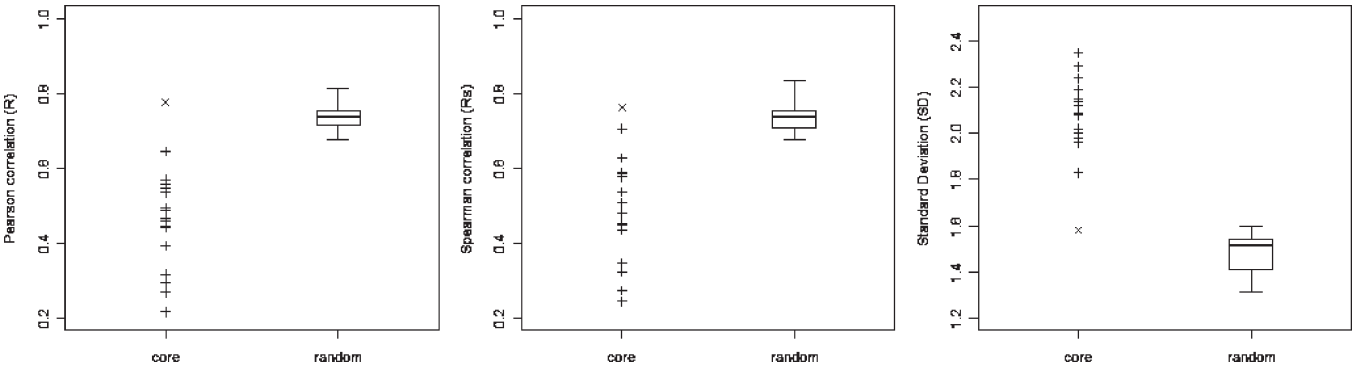

Figure 1. RF-Score performance against that of 16 established scoring functions on the pre-existing PDBbind benchmark (full details in the original publication).[2] On the left of each plot, the "core" column shows the performance of each tested scoring function on the same independent test set (PDBbind 2007 core set). The "×" sign marks the performance of RF-Score, which is substantially better than that of the rest of functions marked by × signs in each of the three performance measures (Pearson correlation coefficient, R, Spearman rank-correlation, R_S, and standard deviation between predicted and measured binding affinity in log units, SD). On the right of each plot, the "random" column shows a boxplot summarizing the performance of RF-Score using 25 randomly generated training/test partitions with the same sizes as the core partition (1105/195). The experiment demonstrates that there is a minor difference in RF-Score performance between PDBbind core and more realistic random partitions.

and standard deviation (SD = 1.58 versus 1.83 log units) than X-Score::HMScore. In order to investigate whether this training/test split was particularly advantageous for both scoring functions, we also generated 25 additional RF-Score models based on random nonoverlapping partitions and observed a median correlation coefficient of R = 0.74 (a difference of +0.04, i.e., RF-Score does somewhat better than the median on this metric) and a median SD = 1.51 (a difference of +0.07, i.e., RF-Score does somewhat worse than the median on this metric). These experiments demonstrate that there is a minor difference in RF-Score performance between PDBbind core and more realistic random partitions (see last paragraph of Section 4 and Appendix A4 in our article[2] for full details of these experiments, which are here summarized in Figure 1). It is unclear why Kramer and Gedeck did not mention either of these rigorous validations, which are entirely relevant to the aim of their study.

Most importantly, we consider that LCOCV is ultimately of little practical value. Indeed, the ultimate goal of a validation strategy is to simulate with sufficient accuracy the difficulties that one would encounter when applying a methodology in a real-world scenario. So, in the context of the proposed validation, this question translates to whether the protein target against which we want to screen shows high sequence similarity to any of the high-quality structures deposited in the PDB (please note that the PDBbind refined set is the result of a nonredundant sampling of the entire PDB). Overington et al.[12] investigated this question and concluded that over 92% (300) of all existing drug targets had sufficient sequence similarity to indicate that they share a fold with known proteins in the PDB. Another example is in the 1741 PDBbind 2009 refined set complexes used by Kramer and Gedeck's study,[1] where as many as 82% (1420) of the complexes contain proteins with BLAST sequence similarity above 90% to proteins of other complexes in the set. Therefore, since a target protein that does not have high sequence similarity to any other protein in a diverse and large training set constitutes an uncommon scenario, LCOCV will not be appropriate in most cases of interest, and certainly it is not necessary. LCOCV would, however, be suitable for estimating the performance that a generic scoring function would achieve on a truly new target protein that does not belong to a cluster represented by any of the proteins in the training set. From a practical standpoint, one always will be able to calculate the sequence similarity between the target protein and those in the training set to establish which pre-existing validation is more relevant for a particular case.

As extensively discussed in our paper, we believe that the quality of RF-Score is due to the circumvention of error-prone modeling assumptions and that the component from prior familiarity with other members of the same cluster is at a level typical in computational drug design. Any learning of cluster membership is indirect rather than explicit. There is nothing wrong with learning residue-level cluster membership as a byproduct of learning how the binding strength is linked to the atomic-level structure of the complex. In fact, such a correlation between sequence and atomic-level properties is to be expected, as the spatial arrangement of atoms in a binding site is a consequence of the characteristic folding of that particular sequence of residues. Therefore, removing from the training set those complexes with proteins that belong to the same cluster as the target protein means that we are depriving the scoring function from the most relevant data for calibration without good reason. Since the only published comparison of which we are aware was our own, and it is clear that RF-Score performed very well in that study, we encourage the community to carry out comparisons of RF-Score against alternative scoring functions. We and others[4,11] believe that a fair comparison requires a common benchmark for all tested scoring functions. Borrowing the words[4] of Breneman and co-workers on a recent study presenting a support vector regression-based scoring function: "Rigorous quantitative comparison between two empirical scoring functions requires that not only the test set but also the training set be identical. This is because the choice of training sets can have a great impact on the performance of empirical scoring functions." Furthermore, as we did for the validation of RF-Score, we recommend using a pre-existing benchmark where other scoring functions had previously been tested, so as to ensure the optimal application of such functions by their authors as well as to avoid the danger of constructing a benchmark complementary to the presented scoring function. With large data sets and performance results for the PDBbind benchmark being publicly and freely available, we cannot think of a valid reason for not including a comparison of scoring functions using the same

training and test sets in addition to the performance of the version preferred by the authors.

To sum up, we consider that Kramer and Gedeck's proposed validation protocol is a valuable contribution that would be useful for predicting the performance of a scoring function on a minority of target proteins but would not be a reliable performance indicator for most cases of interest. Regarding RF-Score itself, we have argued here why there is nothing in the composition of the training or test sets that explains why it performs so well compared to other scoring functions, since the only real difference between RF-Score and standard scoring functions is that the former makes absolutely no assumption about the relationship between structure and binding measurements. Finally, we fully agree with these authors that the expertise that chemoinformatics professionals have on the use of machine learning to build quantitative structure−activity relationship (QSAR) models is now fully transferable to scoring function development and that such transfer would certainly be valuable, whether this entails using widely tested machine learning-based regression methodologies or studying the applicability of a rich body of work on QSAR model validation techniques.[13−17] In our view, the same applies to computationally minded physical chemists and structural biologists with expertise in this area.

AUTHOR INFORMATION

Corresponding Author

*E-mail: pedro.ballester@ebi.ac.uk; jbom@st-andrews.ac.uk.

ACKNOWLEDGMENT

P.J.B. would like to thank Dr. John Overington from EMBL-EBI for helpful discussions and the Medical Research Council for a Methodology Research Fellowship. We acknowledge funding from the Biotechnology and Biological Sciences Research Council (grant BB/G000247/1) and the Scottish Universities Life Sciences Alliance (SULSA).

NOTE ADDED AFTER ASAP PUBLICATION

This paper was published on the Web on May 18, 2011. The format of the paper has been changed, but the scientific content has not. The corrected version was reposted on July 15, 2011.

REFERENCES

(1) Kramer, C.; Gedeck, P. Leave-Cluster-Out Cross-Validation Is Appropriate for Scoring Functions Derived from Diverse Protein Data Sets. *J. Chem. Inf. Model.* **2010**, *50*, 1961–1969.

(2) Ballester, P. J.; Mitchell, J. B. O. A machine learning approach to predicting protein-ligand binding affinity with applications to molecular docking. *Bioinformatics* **2010**, *26*, 1169–1175.

(3) Baum, B.; Muley, L.; Smolinski, M; Heine, A.; Hangauer, D.; Klebe, G. Non-additivity of functional group contributions in protein-ligand binding: a comprehensive study by crystallography and isothermal titration calorimetry. *J. Mol. Biol.* **2010**, *397*, 1042–1054.

(4) Das, S.; Krein, M. P.; Breneman, C. M. Binding Affinity Prediction with Property-Encoded Shape Distribution Signatures. *J. Chem. Inf. Model.* **2010**, *50*, 298–308.

(5) Kinnings, S. L.; Liu, 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.

(6) *RF-Score*; University of St. Andrews: Scotland, U.K.; http://chemistry.st-andrews.ac.uk/staff/jbom/group/RF-Score.html. Accessed April 4, 2011).

(7) *Creative Commons*; Creative Commons corporation: Mountain View, CA, U.S.A.; http://creativecommons.org/. Accessed April 4, 2011.

(8) *The R Project for Statistical Computing*; The Institute for Statistics and Mathematics: Wien, Austria; http://www.r-project.org/. Accessed April 4, 2011.

(9) Wang, R.; Fang, X.; Lu, Y.; Yang, C.-Y.; Wang, S. The PDBbind Database: Methodologies and updates. *J. Med. Chem.* **2005**, *48*, 4111–4119.

(10) *PDBbind-CN Database*; Shanghai Institute of Organic Chemistry: Shanghai, China; www.pdbbind.sioc.ac.cn. (temporarily at http://www.sioc-ccbg.ac.cn/pdbbind/). Accessed April 4, 2011.

(11) Cheng, T.; Li, X.; Li, Y.; Liu, Z.; Wang, R. Comparative Assessment of Scoring Functions on a Diverse Test Set. *J. Chem. Inf. Model.* **2009**, *49*, 1079–1093.

(12) Overington, J. P.; Al-Lazikani, B.; Hopkins, A. L. How many drug targets are there?. *Nat. Rev. Drug Discovery* **2006**, *5*, 993–996.

(13) Golbraikh, A.; Shen, M.; Xiao, Z.; Xiao, Y.; Lee, K.; Tropsha, A. Rational selection of training and test sets for the development of validated QSAR models. *J. Comput.-Aided Mol. Des* **2003**, *17*, 241–253.

(14) Baumann, K. Chance Correlation in Variable Subset Regression: Influence of the Objective Function, the Selection Mechanism, and Ensemble Averaging. *QSAR Comb. Sci.* **2005**, *24*, 1033–1046.

(15) Gramatica, P. Principles of QSAR models validation: internal and external. *QSAR Comb. Sci.* **2007**, *26*, 694–701.

(16) Consonni, V.; Ballabio, D.; Todeschini, R. Comments on the definition of the Q2 parameter for QSAR validation. *J. Chem. Inf. Model.* **2009**, *49*, 1669–1678.

(17) Todeschini, R.; Ballabio, D.; Consonni, V.; Manganaro, A.; Mauri, A. Canonical Measure of Correlation (CMC) and Canonical Measure of Distance (CMD) between sets of data. Part 1. Theory and simple chemometric applications. *Anal. Chim. Acta* **2009**, *648*, 45–51.

Pedro J. Ballester*,† and John B. O. Mitchell*,‡

† European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, United Kingdom

‡ Biomedical Sciences Research Complex and EaStCHEM School of Chemistry, University of St. Andrews, North Haugh, St. Andrews, Fife KY16 9ST, United Kingdom