JOURNAL OF CHEMICAL INFORMATION AND MODELING

ERRATUM

pubs.acs.org/jcim

Correction to "A Machine Learning-Based Method To Improve Docking Scoring Functions and Its Application to Drug Repurposing"

[*Journal of Chemical Information and Modeling* **2011**, *51*, 408–419. DOI: 10.1021/ci100369f].

Sarah L. Kinnings,[†] Nina Liu,[‡] Peter J. Tonge,[‡] Richard M. Jackson,[†] Lei Xie,[*,§,∥] and Philip E. Bourne[*,§]

[†]Institute of Molecular and Cellular Biology and Astbury Centre for Structural Molecular Biology, University of Leeds, Leeds, LS2 9JT, United Kingdom

[‡]Institute of Chemical Biology & Drug Discovery, Department of Chemistry, Stony Brook University, Stony Brook, New York 11794, United States

[§]Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California San Diego, La Jolla, California 92093, United States

[∥]Department of Computer Science, Hunter College, The City University of New York, New York, New York 10065, United States

This Erratum is to clarify that eHiTS-Energy was used in the published manuscript. eHiTS-Energy is the scoring function provided by the eHiTS protein−ligand docking software used to predict binding affinity. There is another scoring function provided by the eHiTS software, called eHiTS-Score, which has been trained specifically for distinguishing active compounds from inactive ones. To avoid confusion in this regard, several points need to be clarified.

(1) The focus of the paper was to demonstrate that the nonlinear combination of energy terms trained by a support vector machine (SVM) could outperform the conventional linear combination of energy terms for estimating the binding affinity during protein−ligand docking. The aim was to apply the improved binding affinity prediction to the reverse docking problem for drug repurposing. Our strategy for drug repurposing was different from conventional virtual screening. A set of putative active compounds for the target of interest, all of which were active against a different target and usually structural analogs, was identified through ligand binding site comparison. Thus the primary concern of protein−ligand docking was to rank these compounds by their binding affinities. For this purpose, eHiTS-Energy, which is the binding affinity predicted by eHiTS and weighted combination of individual energy terms that were used as feature vectors to train our models, was used throughout the text.

(2) Throughout the text, 'eHiTS scoring function' specifically means eHiTS-Energy instead of eHiTS-Score. All of the results, discussion, and conclusions are based on eHiTS-Energy.

(3) Although eHiTS-Energy is not trained to be target specific, the eHiTS software has implemented and provided family specific eHiTS-Scores. eHiTS-Score was trained on PDB complexes. The receptors were classified as families based on their binding site composition and geometry. A score weight set was used for each family. There is good overlap between geometric and biological families, although the correspondence is not one-to-one.

(4) In eHiTS energy terms, the family term reflects to what extent the predicted binding mode matches the binding patterns recognized in the geometric family, and the depth term is the sum of the depth values of all of the atoms within the pocket. Both terms are therefore pose dependent. Because the compounds ranked in our study were usually structural analogs, we expected that their binding poses were not significantly different. To reduce the chance of overfitting, these two terms were ignored in our SVM models. As shown in the following table, Table 1, inclusion of these two terms has marginal impact on the performance of both the regression and classification SVM models.

Table 1. Comparison of the Impact of Feature Selection on the Classification and Regression Models

feature selection	F-score of classification model	correlation coefficient of regression model[a]
original feature vectors used in the paper	0.232	0.637
original feature vectors + depth term	0.227	0.647
original feature vectors + family term	0.240	0.642
original feature vectors + both depth and family terms	0.225	0.647

[a] The correlation coefficient of the regression model is based on an optimal model instead of a cross-validated model.

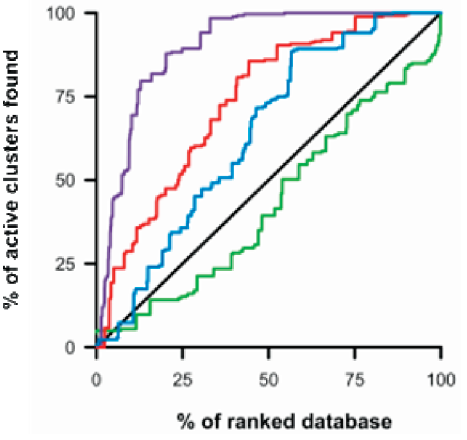

Figure 4

(5) Figure 4 is updated to include eHiTS-Score (blue line). While eHiTS-Score can be seen to outperform eHiTS-Energy (green line), it does not perform as well as either the SVM regression (red line) or the SVM classification (purple line) models in the paper.

(6) Figure 6 has been updated to include eHiTS-Score (blue line). eHiTS-Score can be seen to significantly outperform eHiTS-Energy (green line) across the vast majority of the targets. eHiTS-Score also outperforms the SVM

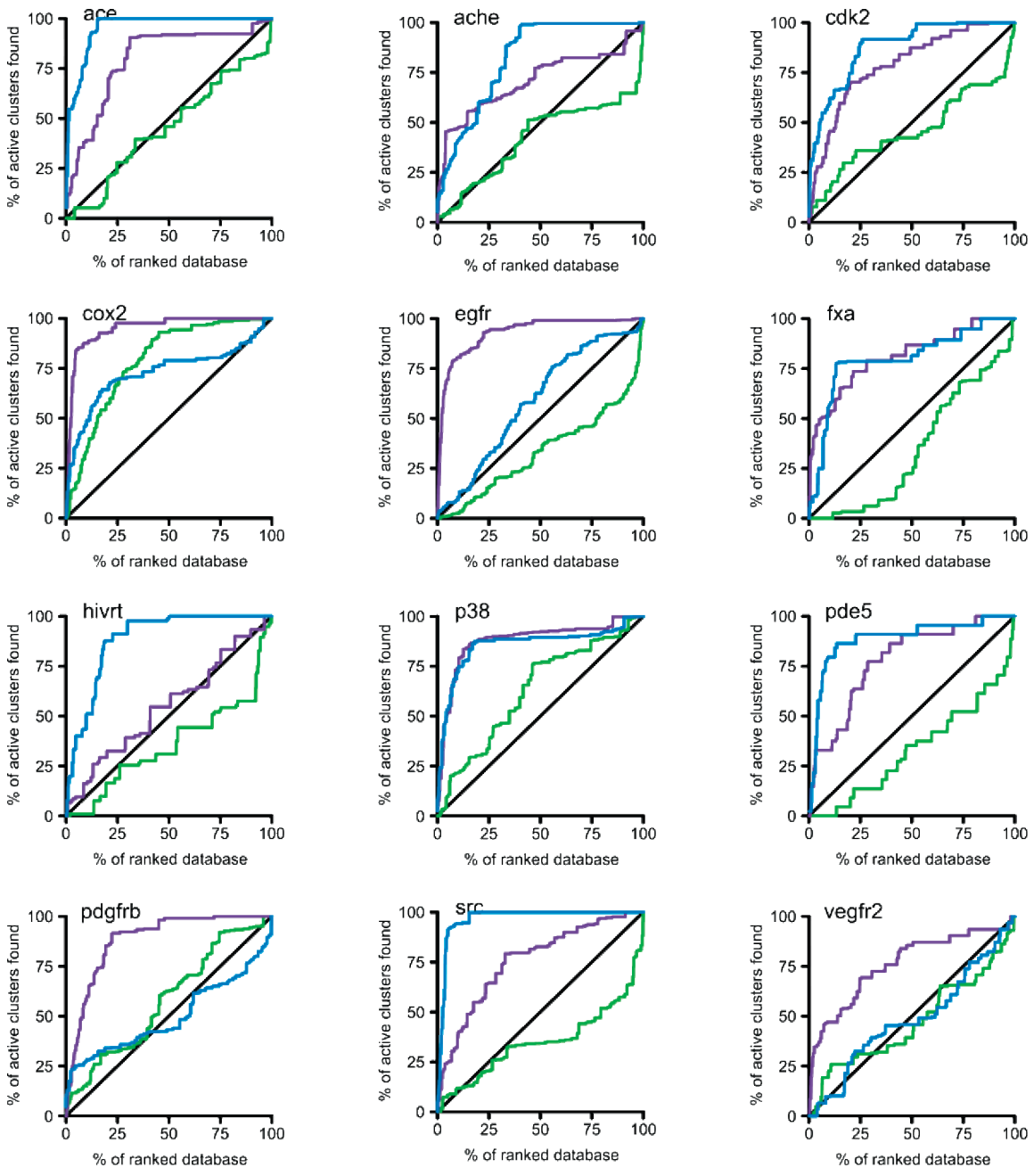

Figure 6

classification model (purple line) in 50% of the targets. However, the SVM classification model is able to outperform eHiTS-Score for the targets cox2, egfr, pdgfrb, and vegfr2. The performance of the SVM classification model and the eHiTS-Score is comparable for ache, fxa, and p38. As discussed in the paper, the variability of the performance of the SVM models among different target families may be due to the chemical space coverage of active compounds. The origin of the performance variability observed for eHiTS-Score is unclear, since insufficient information is provided by the eHiTS software. Note that the enrichment curve of the SVM model is based on a five-fold cross validation instead of an optimal model.

(7) There are three major conclusions in the published paper: (1) The nonlinear combination of energy terms performs better than the linear regression of energy terms when predicting binding affinity; (2) a novel multiplanar SVM model may outperform conventional SVM algorithms when training high-throughput screening data; and (3) phosphodiesterase inhibitors may be repurposed as lead compounds for the direct inhibition of InhA in the treatment of tuberculosis. These conclusions still hold and are not affected by this Erratum. In addition to the original conclusions, this Erratum suggests that: (1) The performance of the SVM classification model is comparable to that of eHiTS-Score; and (2) eHiTS-Score is superior to eHiTS-Energy in distinguishing active compounds from inactive ones.

dx.doi.org/10.1021/ci2001346 |*J. Chem. Inf. Model.* 2011, 51, 1195–1197

AUTHOR INFORMATION

Corresponding Author
*E-mail: lxie@ucsd.edu (L.X.); pbourne@ucsd.edu (P.E.B.).

DOI: 10.1021/ci2001346
Published on Web 04/28/2011