[P10] Beware of machine-learning based scoring functions – On the danger of developing black boxes.

Joffrey Gabel, Jérémy Desaphy and Didier Rognan

Laboratoire d'Innovation Thérapeutique, UMR 7200 CNRS-Université de Strasbourg, F-67400 Illkirch

Training machine learning algorithms with protein-ligand descriptors has recently gained considerable attention to predict absolute binding free energies from atomic coordinates. Starting from a series of recent reports stating the superiority of this approach over empirical scoring functions, we could indeed reproduced the claimed superiority of Random Forest and Support vector machine scoring functions to predict experimental binding constants from protein-ligand X-ray structures of the PDBbind dataset. Strikingly, these scoring functions, trained on simple protein-ligand element-element distance counts, are absolutely unable to enrich virtual screening hit lists in true actives upon docking experiments of 10 reference DUD-E datasets; a feature however verified for an a priori less accurate empirical scoring function (Surflex-Dock). By systematically varying ligand poses from true X-ray coordinates, we show that the Surflex-Dock scoring function is logically sensitive to the quality of docking poses. Conversely, machine-learning based scoring functions are totally insensitive to docking poses (up to 10 Å root-mean square deviations) and just sense atomic element counts but not ligand poses. This report does not disqualify using machine learning algorithms to design scoring functions. Protein-ligand element-element distance counts should however been banned for this usage.

Being able to predict binding free energies is necessary but not sufficient to justify the accuracy of a new scoring function. We therefore propose that two additional benchmarking tests must be systematically done when developing novel scoring functions: (i) sensitivity to docking pose accuracy, (ii) ability to enrich hit lists in true actives upon structure-based (docking, receptor-ligand pharmacophore) virtual screening of reference datasets.