**A Special Statement about the PDBbind core set and the CASF benchmark**

In the past, we have received a number of queries about the status of the PDBbind core set. We also noticed that there are some confusions in literature regarding the naming convention of the CASF benchmark developed by our group. Here, we would like to make a formal statement about the PDBbind core set and the CASF benchmark in a hope to answer those queries and also clarify the confusion.

Our group has a long-standing interesting in scoring function development. The PDBbind database is a notable outcome along the path (Liu et al., *Acc. Chem. Res.* 2017, 50, 302-309). The PDBbind database is now updated on an annual basis, and each release of PDBbind is named after the release year, such as PDBbind v.2016, PDBbind v.2017, and so on. The PDBbind database collects experimentally measured binding affinity data for four types of molecular complexes, i.e. protein-ligand complexes, nucleic acid-ligand complexes, protein-protein complexes, protein-nucleic acid complexes. Among them, we have named the collection of protein-ligand complexes as the "general set". We put a focus on this data set because it is most relevant to drug design and discovery studies. Apparently, not every entry in the general set is suitable for calibrating or validating docking/scoring methods due to misc problems in 3D structure, binding data, and other aspects. Therefore, we have selected the relatively "healthy" entries from the general set to compile the so-called "refined set". The refined set serves as a generally acceptable data set for docking/scoring studies. Other researchers may apply the refined set directly to their studies, or use the refined set as the starting point to compile data sets with their own focus. Both the general set and the refined set are updated with the PDBbind database on an annual basis. They should be correctly cited as, for example, "the PDBbind general set v.2016", "the PDBbind refined set v.2017", and so on.

As another part of our efforts, we have established the CASF benchmark (Comparative Assessment of Scoring Functions), which aims at providing an objective platform for assessing scoring functions. The first published work was CASF-2007 (Cheng et al., *J. Chem. Inf. Model.* 2009, 49, 1079−1093). Another major update, i.e. CASF-2013, was published a few years later (Li et al., *J. Chem. Inf. Model.* 2014, 54, 1700−1716; *J. Chem. Inf. Model.* 2014, 54, 1717−1736). The CASF benchmark employs a high-quality set of protein-ligand complexes as the primary test set. This data set is selected from the PDBbind refined set through a systematic, non-redundant sampling procedure, which is named as the PDBbind "core set" by us. Accordingly, each public release of the CASF benchmark is named after the version of the PDBbind database from which the test set is selected. For example, the test set in CASF-2007 was compiled based on PDBbind v.2007, the test set in CASF-2013 was compiled based on PDBbind v.2013, and so on. It is not a good idea to name each CASF benchmark by its publish year. It is because we cannot predict when our paper will be published in prior when we prepare the manuscript.

It is important to point out that unlike the PDBbind database, the PDBbind core set is not updated on an annual basis. As implied above, the PDBbind core set is a component of the CASF benchmark rather than the PDBbind database. The CASF benchmark is not updated on an annual basis due to the following reasons:

* A HUGE amount of efforts is needed to finish each CASF update. The CASF benchmark is more than a simple data set. For instead, it consists of a whole set of evaluation methods, the test set, as well as a large panel of standard scoring functions to be tested as demonstration. A lot of material needs to be prepared, and a lot of computation needs to be conducted for each CASF update.
* Even if it were doable, in our opinion, there is no need to update CASF so frequently. Our current plan is to update the CASF benchmark every three years. In fact, we have finished CASF-2016, and are preparing a manuscript regarding it. We hope that this paper can be published in the year of 2018.

As mentioned above, the last published version of the PDBbind core set is v.2013. This data set was not updated with PDBbind v.2014 and v.2015, so there is no PDBbind core set v.2014 and v.2015. For historical reasons, the PDBbind core set used to be included in the downloadable data package in some previous releases of PDBbind. To avoid further confusion, we have removed the core set from the data packages of recent releases of PDBbind (e.g. PDBbind v.2014, v.2015, v.2016, and v.2017). If needed, the user can obtain the information of the PDBbind core set in the data package of the corresponding CASF benchmark (e.g. CASF-2007 and CASF-2013), which is also downloadable from the PDBbind-CN web site.

In conclusion, the take-home message is:

* The CASF benchmark should not be referred to as the "PDBbind benchmark". There are such wrong naming conventions in literature, and now you know what the correct one is.
* Data package of the CASF benchmark can be downloaded from the PDBbind-CN web site under the "CASF" tab (http://www.pdbbind-cn.org/casf.asp). At this point, we do not think it is necessary to set up two separate web sites to host PDBbind and CASF, respectively.
* Currently, the latest public release of the CASF benchmark is CASF-2013. There will be CASF-2016 soon.

-----------------------------------------------------------------------------------------------------------------------------------------------------

By Prof. Renxiao Wang, Mar 9th, 2018