**Graphical abstract caption**: Two theoretical scoring functions can be equally apt at scoring a protein-ligand complex, but only the capacities of the scoring function which was trained in a self-consistent manner would be expected to be generalizable.