[SC7] Why adaptively-built simple models using small datasets can be sufficient for chemical modeling

J.B. Brown[1],

[1]*Kyoto University Graduate School of Medicine, Sakyo Yoshida Konoemachi, Graduate School of Medicine Building E, 606-8501, Kyoto, Japan.*

In the modern era of combinatorial chemistry and high-throughput screening assays, it is relatively easy for an organization to generate matrices of ligand-target bioactivity that contain thousands, tens of thousands, or hundreds of thousands of entries. Alternatively, this data may be obtained from public resources such as PubChem BioAssay or ChEMBL. Chemogenomic models can make use of these matrices to build models that can infer bioactivity for new ligands on existing targets (drug discovery, hit-to-lead, lead optimization applications) or for new targets of existing ligands (drug repurposing).

A common perception about structure-activity modeling is that more data yields more predictive ability, and more complex model methodologies are certain to yield more predictive models than less complex model techniques. This is an invalid perception, however. As has been shown by Kangas et al[1], Rarey et al.[2], and our group[3–5], it is possible to build an adaptive model that uses only a fraction of a bioactivity collection and can still achieve high predictive accuracy over the entire set of available data. These methods, which employ some type of selection function to iteratively pick examples with accompanying model reconstruction, are referred to as active learning methods.

Our group has investigated the ability of active learning for family-wide chemogenomic bioactivity databases, finding that 5~20% of the total data available is sufficient to achieve high prediction performance (MCC, F1, PPV) over the entire set of available data. Active learning can also be applied in the special case of a single target or single bioactivity endpoint. In this talk, I will detail the method and recent results, explain retrospective and prospective applications, and discuss potential impact.

Bibliography:

1. Kangas JD, Naik AW, Murphy RF (2014) Efficient discovery of responses of proteins to compounds using active learning. BMC Bioinformatics. doi: 10.1186/1471-2105-15-143
2. Lang T, Flachsenberg F, Von Luxburg U, Rarey M (2016) Feasibility of Active Machine Learning for Multiclass Compound Classification. J Chem Inf Model 56:12–20. doi: 10.1021/acs.jcim.5b00332
3. Reker D, Schneider P, Schneider G, Brown J (2017) Active learning for computational chemogenomics. Future Med Chem 9:381–402. doi: 10.4155/fmc-2016-0197
4. Rakers C, Najnin RA, Polash AH, et al (2018) Chemogenomic Active Learning's Domain of Applicability on Small, Sparse qHTS Matrices: A Study Using Cytochrome P450 and Nuclear Hormone Receptor Families. ChemMedChem. doi: 10.1002/cmdc.201700677
5. Rakers C, Reker D, Brown JB (2017) Small Random Forest Models for Effective Chemogenomic Active Learning. J Comput Aided Chem 8:124–142. doi: 10.2751/jcac.18.124