# Predicting bioactivity of compound-drug target protein pairs using support vector regression models reflecting ligand efficiency

**Nobuyoshi Sugaya**[1]
sugaya@pharmadesign.co.jp

[1] Drug Discovery Department, Research & Development Division, PharmaDesign, Inc.,
2-19-8, Hatchobori, Chuo-ku, Tokyo 104-0032, Japan

**Keywords**: Support vector regression, G-protein coupled receptor, Protein kinase, Ion channel, ligand efficiency

Predicting bioactivity of compounds to drug target proteins using machine learning methods is one of the most intensively studied area in drug discovery and development. Although many previous machine learning studies have succeeded in predicting novel ligand-protein interactions with high performance, all of the previous studies to date have been heavily dependent on the simple use of raw bioactivity data of ligand potencies measured by IC50, EC50, Ki, and Kd deposited in databases. In our previous study [1], we have showed that, using support vector machines, binary classification models based on training data reflecting one of the representatives of ligand efficiency, Binding Efficiency Index (BEI) [2] can offer better performance in classifying active and inactive compound-protein pairs than models based on training data reflecting IC50 or Ki. In this study, we report that this result holds also when support vector regression (SVR) models are applied to bioactivity data.

Utilizing bioactivity data measured by IC50 in GPCRSARfari ver. 2, KinaseSARfari ver. 4, and ChEMBL 14 databases [3], we retrieved bioactivity data associated with G protein-coupled receptors, protein kinases, and ion channels and created four types of training data; IC50-based, pIC50-based, BEI-based, and Surface Efficiency Index (SEI)-based. Values of pIC50, BEI, and SEI were transferred from observed values of IC50 in the databases. The number of instances in the training data is shown in Table 1. To represent compound-protein pairs in the training data, three kinds of compound descriptors (MACCS, 2D descriptors in MOE, and OpenBabel FP2) and single protein descriptor (frequencies of dimmers of amino acid in protein sequence) were used. From GPCRSARfari ver. 3, KinaseSARfari ver. 5.01, and ChEMBL 15 databases, we collected newly added bioactivity data and used the data as validation data for evaluating the performance of the constructed SVR models. Objective comparisons of the performance of the SVR models showed that their prediction capabilities follow an order of SEI > BEI > pIC50 > IC50 as a whole. This result is independent of compound descriptors used and drug target protein families. The superiority of ligand efficiency-based SVR models may be partially attributed to distinct distribution patterns of pIC50s, BEIs, and SEIs, showing narrower range of BEIs than pIC50s and SEIs than BEIs.

[1] Sugaya, N., Training based on ligand efficiency improves prediction of bioactivities of ligands and drug target proteins in a machine learning approach, *Journal of Chemical Information and Modeling*, in press.

[2] Abad-Zapatero, C. and Metz, J. T., Ligand efficiency indices as guideposts for drug discovery, *Drug Discovery Today*, 10:464-469, 2005.

[3] Gaulton, A., *et al.*, ChEMBL: a large-scale bioactivity database for drug discovery, *Nucleic Acids Research*, 40:D1100-D1107, 2012.