JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

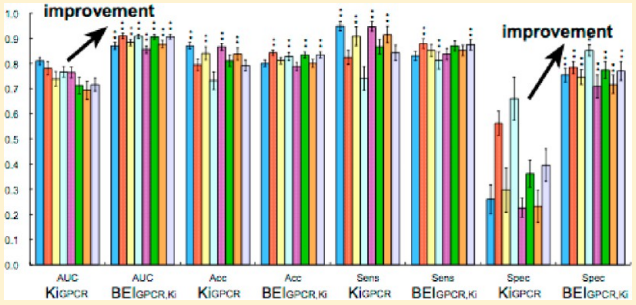
Training Based on Ligand Efficiency Improves Prediction of Bioactivities of Ligands and Drug Target Proteins in a Machine Learning Approach

Nobuyoshi Sugaya*

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

Supporting Information

ABSTRACT: Machine learning methods based on ligand–protein interaction data in bioactivity databases are one of the current strategies for efficiently finding novel lead compounds as the first step in the drug discovery process. Although 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 IC_{50}, EC_{50}, K_i, and K_d deposited in databases. ChEMBL provides us with a unique opportunity to investigate whether a machine-learning-based classifier created by reflecting ligand efficiency other than the IC_{50}, EC_{50}, K_i, and K_d values can also offer high predictive performance. Here we report that classifiers created from training data based on ligand efficiency show higher performance than those from data based on IC_{50} or K_i values. Utilizing GPCRSARfari and KinaseSARfari databases in ChEMBL, we created IC_{50}- or K_i-based training data and binding efficiency index (BEI) based training data then constructed classifiers using support vector machines (SVMs). The SVM classifiers from the BEI-based training data showed slightly higher area under curve (AUC), accuracy, sensitivity, and specificity in the cross-validation tests. Application of the classifiers to the validation data demonstrated that the AUCs and specificities of the BEI-based classifiers dramatically increased in comparison with the IC_{50}- or K_i-based classifiers. The improvement of the predictive power by the BEI-based classifiers can be attributed to (i) the more separated distributions of positives and negatives, (ii) the higher diversity of negatives in the BEI-based training data in a feature space of SVMs, and (iii) a more balanced number of positives and negatives in the BEI-based training data. These results strongly suggest that training data based on ligand efficiency as well as data based on classical IC_{50}, EC_{50}, K_d, and K_i values are important when creating a classifier using a machine learning approach based on bioactivity data.



INTRODUCTION

Now, machine learning methods that combine the properties of small molecule compounds and proteins by utilizing ligand-protein interaction databases are one of the current strategies for efficiently finding novel lead compounds to target a protein or protein family of interest in the drug discovery process. Over the past decade, significant effort has been dedicated to the development of machine-learning-based classifiers or predictors that can discriminate active and inactive pairs of drug-like small molecule ligands and their target proteins with high predictive performance. Most protein families that are popular as drug targets, such as G protein-coupled receptors (GPCRs),[1−9] protein kinases (PKs),[5,7,10] ion channels,[2,6,8,9] nuclear receptors,[6,8,9] proteases,[11,12] and other enzymes,[2,5,6,8,9,13] are inside the scope of previous machine learning studies. In particular, GPCRs have been intensively studied. The pioneer work of Bock and Gough[1] showed that SVM-based regression models based on ligand–GPCR binding data in PDSP database[14] performed well in predicting binding affinities between novel ligands and orphan GPCRs. Using ligand–GPCR interaction data in the KEGG[15] or GLIDA[16] databases, Jacob and colleagues created SVM classifiers with high discriminative power[2,3] and succeeded in predicting ligand–GPCR pairs with an accuracy of 78%.[3] Following these studies, many papers[3−9] have reported machine-learning-based classifiers with much higher accuracy (>90%). PKs are another protein family on which previous machine learning studies using ligand–target interaction data have focused. Wang et al.[5] constructed a SVM classifier from BindingDB,[17] and it was successfully used in predicting novel ligands for several pharmaceutically important drug target proteins, including PKs. Yabuuchi et al.[7] demonstrated that an SVM classifier based on binding affinity data for ligand–PK pairs in the GVK Biosciences kinase inhibitor database[18] can predict novel compounds that bind to EGFR or CDK2; the high potencies of these compounds for the target proteins were validated by in vitro experiments. Buchwald et al.[10] applied SVMs to a set of ligand–PK binding

Received: April 25, 2013

   dx.doi.org/10.1021/ci400240u | J. Chem. Inf. Model. XXXX, XXX, XXX–XXX

data reported by Fabian et al.[19] and showed that it can predict binding/nonbinding with moderate accuracy (>70%).

Most ligand−protein interaction databases or data sets used in these studies, such as PDSP,[14] GLIDA,[16] BindingDB,[17] GVK Biosciences kinase inhibitor database,[18] and Fabian et al.,[19] are a collection of bioactivity data that records ligand potency measured by IC_{50}, EC_{50}, K_i, and K_d to a target protein, assayed by in vivo or in vitro experiments. On the basis of these data, many previous studies[2−6,8,9,13] have simply used ligand−target interaction pairs registered in the databases as positive data for machine learning methods, without considering ligand potencies. There are several studies[1,7,10−12] that do take ligand potencies into consideration, and the positive data were prepared from bioactivity data that satisfied a researcher-defined threshold (for example, IC_{50}, EC_{50}, or $K_i < 1\ \mu M$ in the work of Yabuuchi et al.,[7] $K_d < 10\ \mu M$ in the work of Buchwald et al.[10]). However, the thresholds are also based on raw data of ligand potency measurements recorded in the databases. Therefore, all previous machine-learning studies appear to depend heavily on the simple use of the raw data in the bioactivity databases. In contrast, ChEMBL[20] is distinct from other bioactivity databases in that it stores not only the raw data of ligand potencies recorded as IC_{50}, EC_{50}, K_i, and K_d, etc., but also a measurement of ligand efficiency transferred from these values. This database provides us with a unique opportunity to investigate whether a machine-learning-based classifier created from training data also can have high predictive power when the training data is prepared to take into consideration a measurement of ligand efficiency other than raw IC_{50}, EC_{50}, K_i, and K_d values.

Hopkins et al.[21] introduced ligand efficiency as a metric of the binding energy of a ligand per the number of non-hydrogen atoms in the ligand interacting with a target protein. Soon thereafter, the concept was extended to encompass the potency of the ligand divided by the size of the ligand, such as molecular weight and polar surface area.[22,23] Today, several measures of ligand efficiency have been developed including the original ligand efficiency index,[21] binding efficiency index (BEI),[22,23] percentage efficiency index,[22,23] and surface efficiency index (SEI).[22,23] Ligand efficiency is considered to be more useful than ligand potency in selecting potential leads and optimizing their structures in the drug development process.[21−23] Ligand efficiency enables medicinal chemists to directly compare the activities of drug candidates with a wide range of sizes, even though the candidates inevitably increase in size during the development process.[21−23] If ligand efficiency is utilized for creating training data for machine learning methods, this characteristic of ligand efficiency will produce a classifier that is more appropriate for predicting better lead compounds from the viewpoint of potency per unit size of the compound.

The aim of this study was to construct machine-learning-based classifiers using SVMs from training data incorporating ligand efficiency, and, through objective classifier comparisons by statistical tests, to investigate whether these classifiers have better predictive performance than those constructed from data that consider ligand potency (IC_{50} and K_i). Previous versions of GPCRSARfari and KinaseSARfari in ChEMBL were used for creating the training data. BEI, which is available in ChEMBL, was adopted as a measure of ligand efficiency. We validated the predictive power of the BEI-based and the IC_{50}- or K_i-based SVM classifiers by applying them to the prediction of new data in the latest versions of GPCRSARfari and KinaseSARfari.

METHODS

Bioactivity Data in GPCRSARfari and KinaseSARfari. GPCRSARfari and KinaseSARfari are bioactivity databases of ligands and class A rhodopsin-like GPCRs or PKs and are components of ChEMBL. As of October 30, 2012, GPCRSARfari versions 2 and 3 had been released, containing 947 914 and 1 008 927 bioactivity data points, respectively. KinaseSARfari versions 4 and 5.01 had also been released, containing 417 092 and 503 041 data points, respectively. GPCRSARfari and KinaseSARfari were downloaded from the ftp site.[24,25] We utilized the bioactivity data in GPCRSARfari version 2 and KinaseSARfari version 4 as training data for our SVM-based method. New data in the latest versions were used to validate our SVM classifiers constructed from the previous versions (see the Validation Data section). The bioactivity data in GPCRSARfari and KinaseSARfari were collected from peer-reviewed scientific journals by curators and are composed of data assayed using various experimental methods and under various conditions.[20] Thus, the reliability of these data vary widely, and measurement types of the data are not standardized. To retrieve more reliable data for machine learning, we selected the data with "assay type B". B means "binding" assayed by in vitro experiments. Next, among various measurement types in the databases, we selected measurement types providing data sufficient for machine learning and then collected bioactivity data measured by the types. In total, 40 826 data points measured by IC_{50} (called "IC_{50GPCR} data" hereafter) and 75 614 data points measured by K_i (called "K_{iGPCR} data" hereafter) were retrieved from GPCRSARfari version 2. Each data set is composed of 22 882 compounds and 142 GPCRs (IC_{50GPCR} data) or 33 541 compounds and 136 GPCRs (K_{iGPCR} data). From KinaseSARfari version 4, we selected 74 204 IC_{50} data points (28 835 compounds and 374 PKs, called "IC_{50PK} data" hereafter).

Bioactivity data reflecting ligand efficiency were created by transferring IC_{50} or K_i values in the IC_{50GPCR}, K_{iGPCR}, or IC_{50PK} data sets to BEI. BEI is one of the simplest measures representing ligand efficiency and is available in ChEMBL. BEI is defined as

$$\mathrm{BEI} = \frac{\mathrm{pIC_{50}(or\ pK_i)}}{\text{molecular weight (kDa) of compound}}$$

BEI data created from the IC_{50GPCR}, K_{iGPCR}, and IC_{50PK} data were named "$BEI_{GPCR,IC_{50}}$", "BEI_{GPCR,K_i}", and "$BEI_{PK,IC_{50}}$", respectively. These data served as the source of the training data for our SVM-based method.

Compound and Protein Descriptors. We used three representative compound fingerprints, MACCS, MACCSF, and TGT, as compound descriptors. These fingerprints were calculated by the software package Molecular Operating Environment (MOE) (ver. 2011.10).[26] Physicochemical properties of the 2D structures of compounds calculated by MOE (called "MOE2D" hereafter) were also used. The MACCS, MACCSF, and TGT fingerprints were composed of 166, 166, and 1704 elements, respectively. The MOE2D descriptor has 186 elements.

We adopted two types of descriptors for representing GPCRs and PKs. One is the frequency of dimers of amino acids in a protein amino acid sequence and is composed of 400 elements (called "diAA" (frequencies of *di*mers of *a*mino *a*cid) hereafter). We used amino acid sequences of GPCRs and PKs downloaded from GPCRSARfari and KinaseSARfari ftp sites

dx.doi.org/10.1021/ci400240u | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

Table 1. Number of Positive and Negative Instances in the Training and Validation Data

	definition of positives and negatives		training data[a]		validation data[b]	
data type	positives	negatives	no. positives	no. negatives	no. positives	no. negatives
IC_{50GPCR}	$IC_{50} \leq 1$ μM	$IC_{50} \geq 10$ μM	22822	5422	1248	91
$BEI_{GPCR,IC_{50}}$	BEI ≥ 12.9	BEI ≤ 11.5	20859	10219	1037	287
K_{iGPCR}	$K_i \leq 1$ μM	$K_i \geq 10$ μM	45894	6816	2244	283
BEI_{GPCR,K_i}	BEI ≥ 15.8	BEI ≤ 13.9	34546	20888	1641	1019
IC_{50PK}	$IC_{50} \leq 1$ μM	$IC_{50} \geq 10$ μM	33236	21001	120	28
$BEI_{PK,IC_{50}}$	BEI ≥ 15.2	BEI ≤ 13.3	31402	23991	2609	1994

[a]Redundant instances were removed except for one instance among them. [b]Instances overlapped with those in the training data were removed. Redundant instances within the validation data were also removed except for one instance among them.

for calculating the diAA descriptor. These amino acid sequences focus on the transmembrane regions of GPCRs and the ATP-binding domains of PKs; other regions and domains are removed. Another descriptor is original and is based on the tertiary structures of the ligand-binding sites of GPCRs and PKs. This descriptor is composed of physicochemical properties (hydrophobicity, volume, and p*I*) of amino acids in contact with the ligands. We retrieved the crystal structures of GPCRs in PDB[27] (as of June 27, 2012) by searching for "7tm_1" Pfam domain[28] using HMMER3[29] with default parameters against all amino acid sequence entries in the PDB. For all ligand-bound structures, we extracted amino acid residues contacting the ligands using LIGPLOT.[30] These residues were mapped on multiple sequence alignment of GPCRs obtained from GPCRDB;[31] all alignment sites with at least one mapped residue were adopted, resulting in the extraction of 66 amino acid residues (Supporting Information Table S1). We assigned values for hydrophobicity, volume, and p*I* (Supporting Information Table S2) to each amino acid. If there was a gap in the alignment, 0 was assigned to the residue. The GPCR descriptor is composed of 198 (= 66 × 3) elements and is called "GPS66" (*G*PCR *p*ocket *s*ite-surrounding *66* amino acid residues) hereafter. The descriptor for PKs is based on the representative 36 amino acid residues surrounding the ATP binding sites.[32] Multiple sequence alignment of PKs was downloaded from KinaseSARfari ftp site. After 36 residues were mapped on the alignment (Supporting Information Table S3), the hydrophobicity, volume, and p*I* values were assigned. The PK descriptor is composed of 108 (= 36 × 3) elements and is called "ABS36" (*A*TP *b*inding *s*ite-surrounding *36* amino acid residues) hereafter.

We represented compound−target protein pairs in the bioactivity data by concatenating compound and protein descriptors. Eight kinds of descriptor concatenation (MACCS +GPS66(or ABS36), MACCSF+GPS66(or ABS36), TGT +GPS66(or ABS36), MOE2D+GPS66(or ABS36), MACCS +diAA, MACCSF+diAA, TGT+diAA, and MOE2D+diAA) were generated.

Training Data. We defined the positive and negative instances based on the values of IC_{50}, K_i, or BEI and created the training data from the IC_{50GPCR}, K_{iGPCR}, IC_{50PK}, $BEI_{GPCR,IC_{50}}$, BEI_{GPCR,K_i}, and $BEI_{PK,IC_{50}}$ data. For IC_{50GPCR}, K_{iGPCR}, and IC_{50PK}, we divided bioactivity data into four categories ("highly actives" (IC_{50} (or K_i) ≤ 100 nM), "actives" (100 nM < IC_{50} (or K_i) ≤ 1 μM), "weakly actives" (1 μM < IC_{50} (or K_i) < 10 μM), and "inactives" (10 μM ≤ IC_{50} (or K_i)) according to a previous study,[1] and then. we used highly actives and actives as the positive instances and inactives as the negative instances. The weakly actives category was not used in this study. This is because, when using a single cutoff threshold such as 1 or 10 μM, there is difficulty in treating bioactivity data on or near the threshold. For example, if a single cutoff 1 μM were adopted, one of two compound−protein pairs showing similar IC_{50} values (one shows IC_{50} = 0.95 μM and another shows IC_{50} = 1.05 μM) would be classified as positive but another would be as negative, in spite of the slight difference in their IC_{50} values. This situation can lead to decrease of predictive performance of the constructed SVM classifiers. To avoid this, we discarded intermediate bioactivity data and adopted two cutoff thresholds.

The plots of BEI versus pIC_{50} (or pK_i) show weak but statistically significant positive correlation (Supporting Information Figure S1). In Figure S1A, IC_{50}s of 1 and 10 μM approximately correspond to BEIs of 12.9 and 11.5, respectively. On the basis of this observation, we defined BEI ≥ 12.9 as the positives and BEI ≤ 11.5 as the negatives when creating the training data from the $BEI_{GPCR,IC_{50}}$ data. Likewise, K_i's of 1 and 10 μM approximately correspond to BEIs of 15.8 and 13.9 in Figure S1B, and IC_{50}s of 1 and 10 μM approximately correspond to BEIs of 15.2 and 13.3 in Figure S1C. Consequently, thresholds of BEI ≥ 15.8 for the positives and BEI ≤ 13.9 for the negatives were adopted for creating the training data from the BEI_{GPCR,K_i} data, and the thresholds of BEI ≥ 15.2 for the positives and BEI ≤ 13.3 for the negatives were adopted for creating the training data from the $BEI_{PK,IC_{50}}$ data.

When used in machine learning, redundant instances (that is, the definition of positive or negative (1 or 0), compound ID, and protein ID are all the same) are removed except for one instance among them. If a compound−protein pair is included in both positives and negatives, both instances of the compound−protein pair are retained. The number of positive and negative instances in the training data is shown in Table 1. All training data thus created can be obtained from Tables S4−S9 in the Supporting Information.

To circumvent the following problems, we created 1000 random training data sets. Each data set is composed of 3000 positives and 3000 negatives randomly chosen from the original training data. One problem is that it is very time-consuming to compute SVM classifiers using all compound−protein pairs in the training data. Another is that using a training data set with a different number of positives and negatives (Table 1) may influence the discriminative power of SVM classifiers calculated from the training data. Using a random training data set instead of the original training data enabled us to not only reduce computational time and avoid an unbalanced number of positives and negatives but also to conduct statistical tests of the constructed SVM classifiers. We calculated the mean values

dx.doi.org/10.1021/ci400240u | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

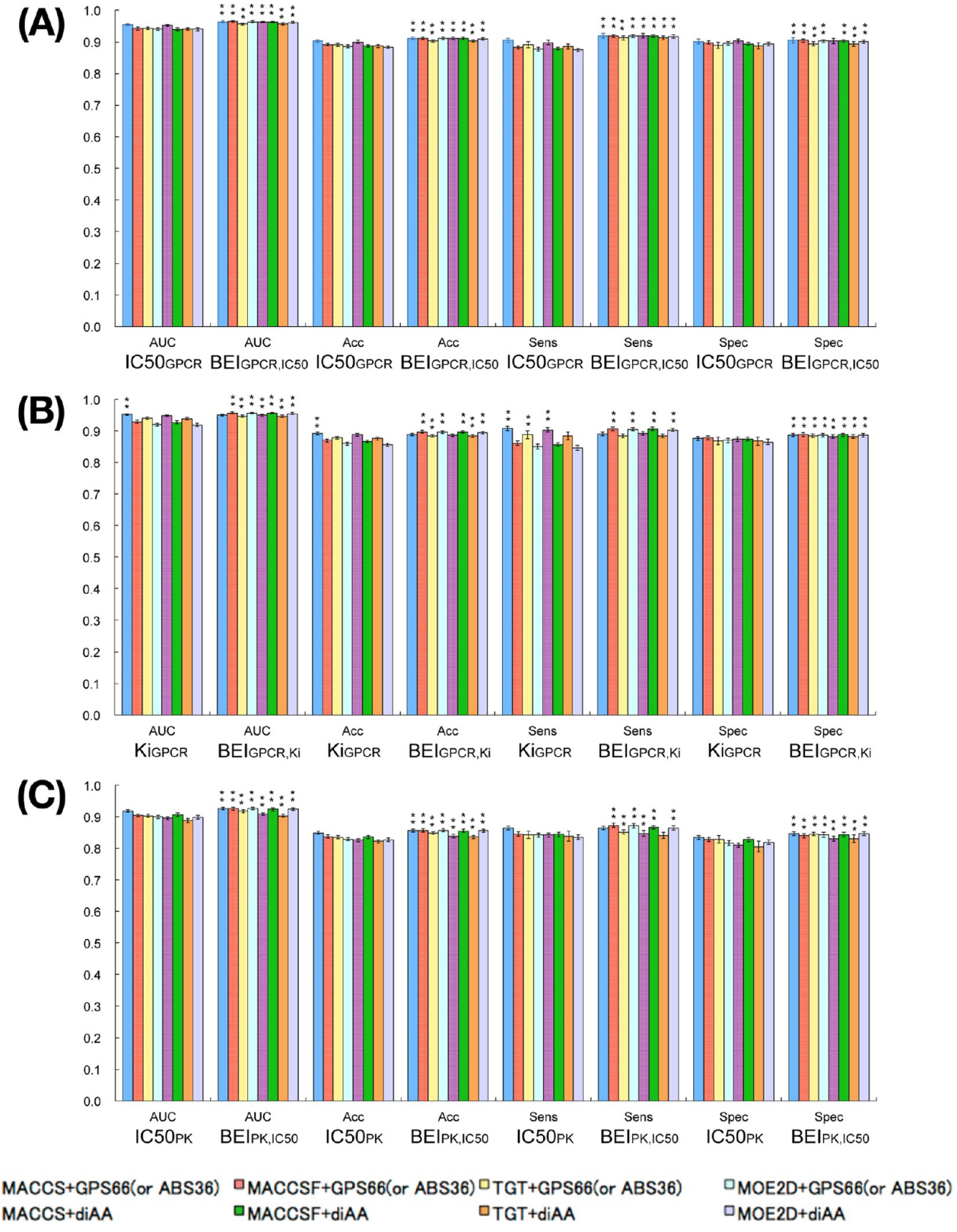


Figure 1. Cross-validation tests using the training data from GPCRSARfari version 2 and KinaseSARfari version 4. Mean values (over the 1000 random training data sets) of AUC, accuracy (Acc), sensitivity (Sens), and specificity (Spec) are shown by a bar plot. SDs are shown by error bars. Training data are from (A) IC_{50GPCR} and $BEI_{GPCR,IC_{50}}$ data, (B) K_{iGPCR} and BEI_{GPCR,K_i} data, or (C) IC_{50PK} and $BEI_{PK,IC_{50}}$ data. * or ** at the top of the bars indicate that the difference of the mean values between an IC_{50}- or K_i-based and a BEI-based classifier is statistically significant with a P value < 0.05 or <0.01, respectively, by the Wilcoxon rank sum test. For more details regarding the statistical tests, see Supporting Information Table S16.

dx.doi.org/10.1021/ci400240u | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

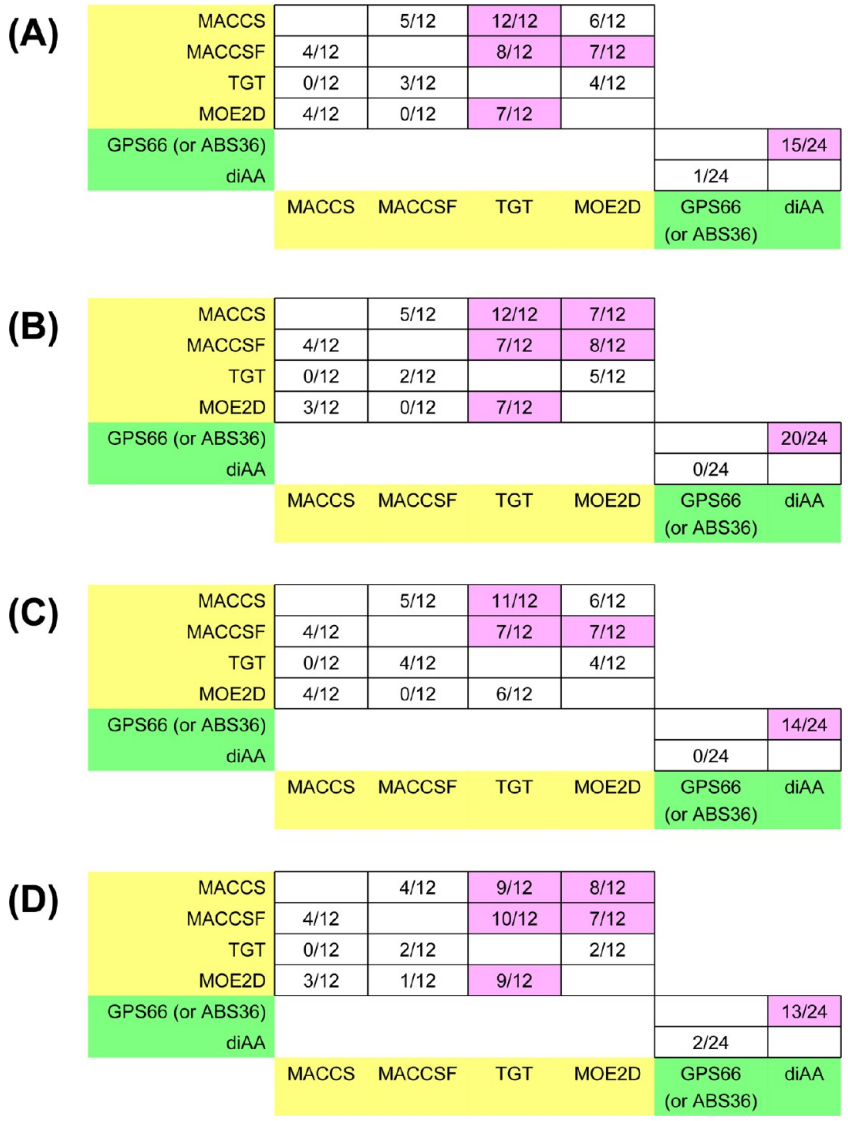


Figure 2. Comparisons of the performance of the descriptors in the cross validation tests. (A) AUC. (B) Accuracy. (C) Sensitivity. (D) Specificity. In each cell, the denominator is the maximum number of practicable comparisons of two SVM classifiers created from two training data sets in which only descriptors of interest differ. The numerator is the number of comparisons in which a SVM classifier in a row shows a higher value with statistical significance (P value < 0.05) than a classifier in a column. Cells in which a value is >0.5 are colored pink. The compound descriptors and the protein descriptors are colored yellow or green, respectively. For more details regarding the statistical tests, see Supporting Information Table S17.

and standard deviations (SDs) over the 1000 random training data sets. The SVM classifiers were evaluated using the area under a receiver operating characteristic curve (AUC), accuracy (Acc), sensitivity (Sens), and specificity (Spec) in 10-fold cross validation tests. Acc, Sens, and Spec are defined as Acc = (TP + TN)/(TP + TN + FP + FN), Sens = TP/(TP + FN), and Spec = TN/(FP + TN), where TP, TN, FP, and FN are the numbers of true positives, true negatives, false positives, and false negatives, respectively.

We used the program package Libsvm[33] and its graphics processing unit-accelerated version.[34] All elements of the descriptors in the training data were scaled in the range 0 to 1 for every element. A radial basis function kernel was used, and the parameters in the kernel function were optimized by the python script "grid.py" in the Libsvm package. AUCs were calculated by the python script "plotroc.py".[35]

Validation Data. To validate our SVM classifiers, we utilized the difference between the latest and previous (used for creating SVM classifiers) versions. GPCRSARfari version 3 and KinaseSARfari version 5.01 contained 61 013 and 85 949 new data points, respectively. From the new data, we selected those with assay type B measured by IC_{50} or K_i. As a result, 1683 IC_{50} data points (composed of 1120 compounds and 75 GPCRs) and 3503 K_i data points (1830 compounds, 92 GPCRs) were selected from GPCRSARfari version 3, and from KinaseSARfari version 5.01, 6761 IC_{50} data points (2623 compounds, 231 PKs) were selected. The positives and negatives were defined according to the thresholds described above (Table 1). Instances overlapping with the training data were removed from the validation data. Redundant instances within the validation data were also removed except for one instance among them. We calculated AUC, accuracy, sensitivity, and specificity when each SVM classifier predicted the positives and negatives in the validation data. All validation data can be obtained from Tables S10−S15 in the Supporting Information.

dx.doi.org/10.1021/ci400240u | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

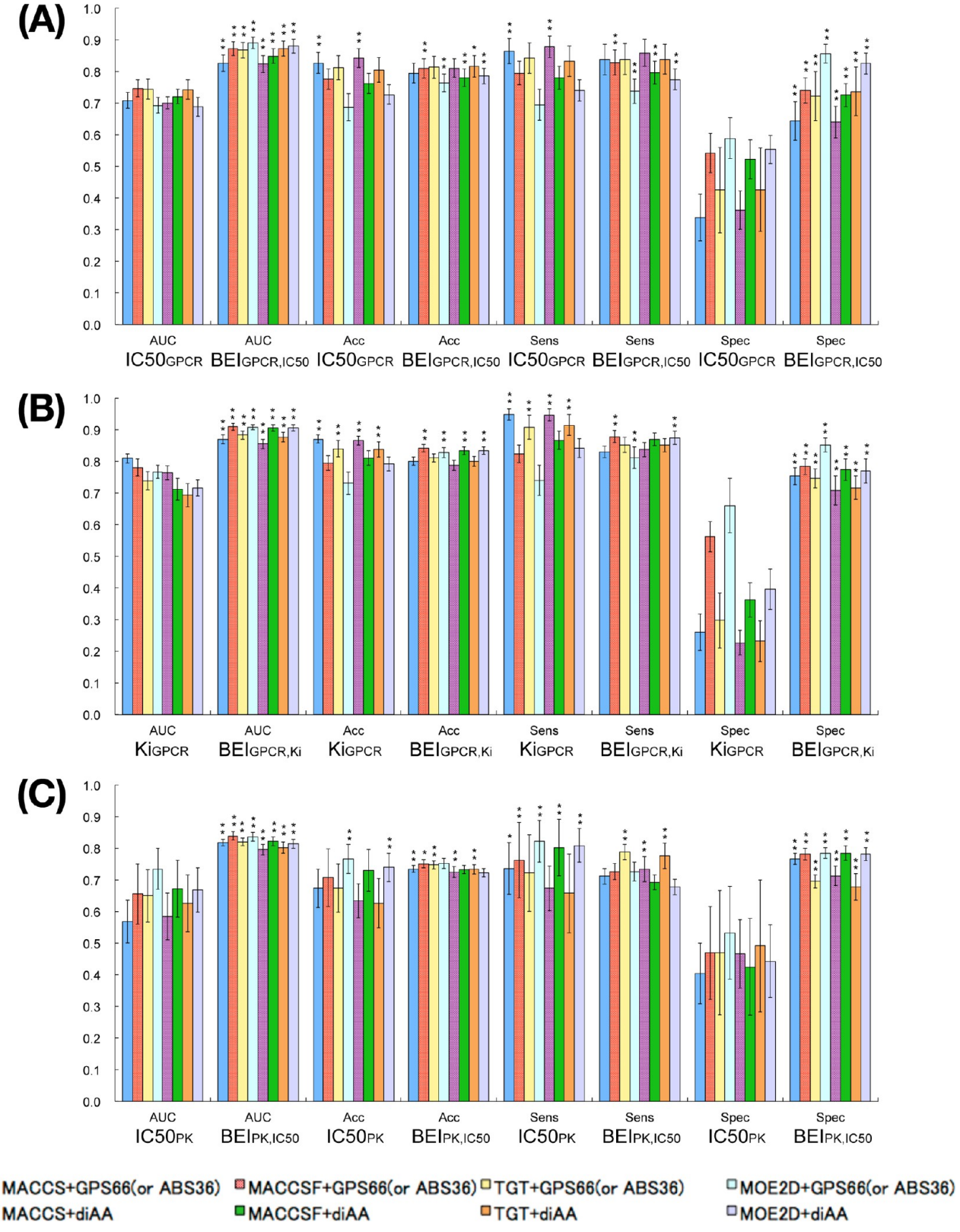


Figure 3. Prediction of the newly data in GPCRSARfari version 3 and KinaseSARfari version 5.01. Mean values (over the 1000 random training data sets) of AUC, accuracy (Acc), sensitivity (Sens), and specificity (Spec) are shown by a bar plot. SDs are shown by error bars. Validation data are from (A) IC_{50GPCR} and $BEI_{GPCR,IC_{50}}$ data, (B) K_{iGPCR} and BEI_{GPCR,K_i} data, or (C) IC_{50PK} and $BEI_{PK,IC_{50}}$ data. * or ** at the top of the bars indicate that the difference of the mean values between an IC_{50}- or K_i-based and BEI-based classifier is statistically significant with a P value < 0.05 or < 0.01, respectively, by the Wilcoxon rank sum test. For more details regarding the statistical tests, see Supporting Information Table S18.

dx.doi.org/10.1021/ci400240u | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

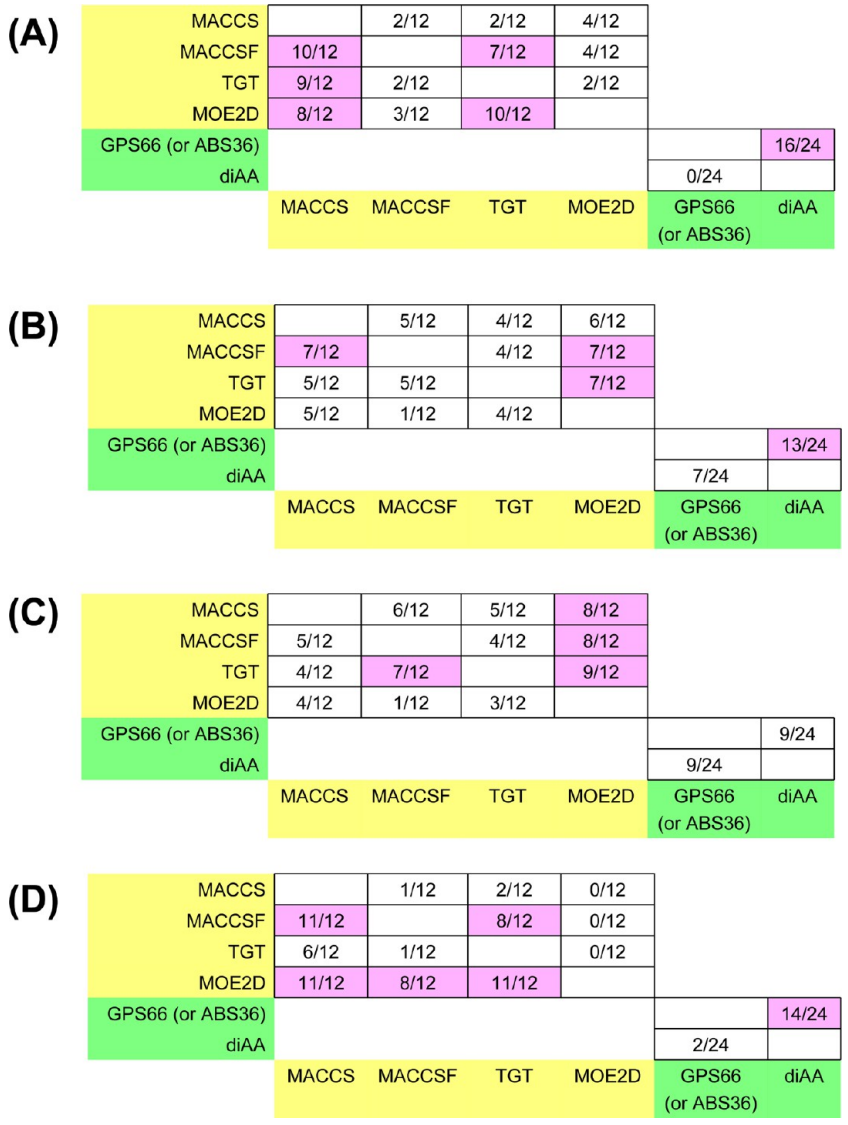


Figure 4. Comparisons of the performance of the descriptors in the prediction of the validation data. (A) AUC. (B) Accuracy. (C) Sensitivity. (D) Specificity. In each cell, the denominator is the maximum number of practicable comparisons of two SVM classifiers created from two training data sets in which only descriptors of interest differ. The numerator is the number of comparisons in which a SVM classifier in a row shows a higher value with statistical significance (P value < 0.05) than a classifier in a column. Cells in which a value is >0.5 are colored pink. The compound descriptors and the protein descriptors are colored yellow or green, respectively. For more details regarding the statistical tests, see Supporting Information Table S19.

RESULTS

Cross-validation Tests. From the bioactivity data in GPCRSARfari version 2 and KinaseSARfari version 4, we created 48 (8 kinds of descriptor concatenation × 6 kinds of data type) original training data sets. After the 1000 random training data sets were created from the original training data (see Training Data in Methods), an SVM classifier was constructed from each training set in the 1000 random data sets. We conducted 10-fold cross validation tests for the constructed classifiers, and mean values of the four statistics (AUC, accuracy, sensitivity, and specificity) over the 1000 random data sets were calculated.

Figure 1 indicates that, on the whole, both the SVM classifiers constructed from the IC_{50}- or K_i-based training data and those from the BEI-based data show high mean values of the statistics. The former classifiers display mean AUCs of 0.89−0.95, mean accuracies of 0.82−0.90, mean sensitivities of 0.84−0.91, and mean specificities of 0.81−0.90, while mean values of the statistics of the latter classifiers are 0.90−0.96 for AUCs, 0.84−0.91 for accuracies, 0.84−0.92 for sensitivities, and 0.83−0.90 for specificities. When the mean values are compared between the IC_{50}- or K_i-based classifiers and the BEI-based classifiers, the differences of the means are statistically significant with a P value < 0.01 or < 0.05 for most comparisons, although the differences are very small. The BEI-based classifiers have larger values of the statistics than the IC_{50}- or K_i-based classifiers (Figure 1 and Supporting Information Table S16). In particular, in Figures 1A and C, none of the IC_{50GPCR}- and IC_{50PK}-based classifiers show higher means of the statistics than the BEI-based classifiers. Also, in Figure 1B, there are only five comparisons where the K_{iGPCR}-based classifier outperforms the BEI_{GPCR,K_i}-based classifier. These results clearly show that, in the cross-validation tests, the BEI-based classifiers can better discriminative the actives and

dx.doi.org/10.1021/ci400240u | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

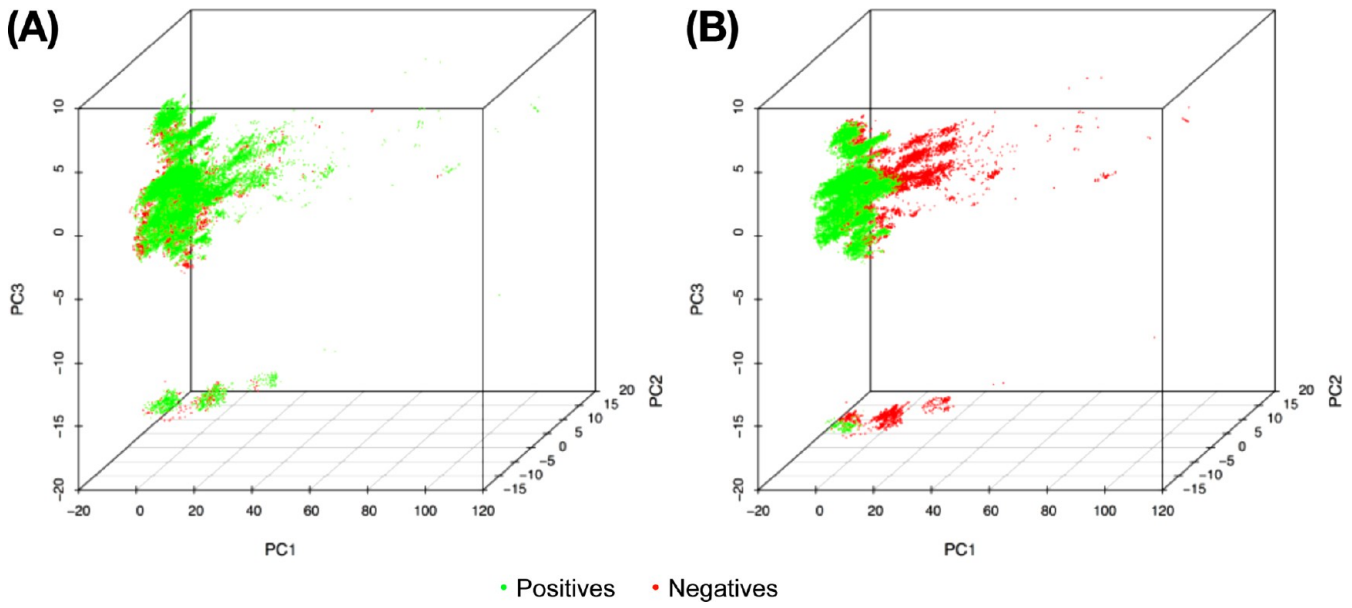


Figure 5. Two examples of the distributions of the positives and negatives in the training data using the three principal components. (A) IC_{50GPCR} training data and MOE2D+GPS66 descriptor. (B) $BEI_{GPCR,IC_{50}}$ training data and MOE2D+GPS66 descriptor. The *x*-, *y*-, and *z*-axes indicate the first principal component (PC1), the second principal component (PC2), and the third principal component (PC3), respectively. The positives and negatives are shown by green or red, respectively. For the distributions of other training data, see Supporting Information Figure S2. The principal component analyses were conducted using the software package R.[36]

inactives in the training data than the IC_{50}- or K_i-based classifiers.

To investigate which descriptor excels, we compared the mean values of the statistics between two SVM classifiers in which only the descriptors of interest differed. For example, the means of the statistics of the classifier, IC_{50GPCR} using MACCS +GPS66, were compared with those of the statistics of the classifier, IC_{50GPCR} using MACCSF+GPS66, to test which compound descriptor, MACCS or MACCSF, offers better discrimination. Figure 2 shows that the performance of the descriptors used in this study is not homogeneous; some descriptors are more discriminative than others in the cross validation tests. In Figure 2A (comparisons of the mean AUCs), 12 comparisons (2 kinds of descriptor concatenation × 6 kinds of data type) are practicable for each pair of the compound descriptors. If the classifiers using MACCS and those using MACCSF are compared with respect to the mean AUCs, 5 out of 12 comparisons show that the classifiers using MACCS have large AUCs with statistical significance. On the other hand, 4 out of 12 comparisons indicate that the mean AUCs of the classifiers using MACCSF are higher than those using MACCS. With respect to all statistics (AUC, accuracy, sensitivity, and specificity), one-half or more of the 12 comparisons show that the classifiers using MACCS and MACCSF outperform those using TGT and MOE2D (see cells colored pink) (Figure 2 and Supporting Information Table S17). Among the latter two, the classifiers using MOE2D overall show better discriminative power than those using TGT. It should be noted, however, that which descriptor is superior to others may depend on the training data used for creating the SVM classifiers. For example, many classifier comparisons between TGT and MOE2D indicate that the classifiers using TGT have higher means of the statistics when the IC_{50}- or K_i-based training data are used, but the classifiers using MOE2D have higher means when the BEI-based training data are used (Supporting Information Table S17). For protein descriptors, 24 comparisons (4 kinds of descriptor concatenation × 6 kinds of data type) are practicable. With respect to all statistics, the classifiers using GPS66 or ABS36 descriptor based on the tertiary structures of GPCRs and PKs perform better in the cross-validation tests than those using diAA descriptor based on the amino acid sequences of the proteins (Figure 2 and Supporting Information Table S17). The results of the descriptor comparisons imply that the compound descriptors, MACCS and MACCSF, and the protein descriptors, GPS66 and ABS36, may exhibit more discriminative performance than the others in the cross validation tests (see Discussion).

Validation of the SVM Classifiers. By applying the SVM classifiers created from the training data based on the previous database versions to the new data in the latest versions, we validated the predictive performance of the classifiers. AUC, accuracy, sensitivity, and specificity were calculated for each training set in the 1000 random data sets and, then, averaged to the mean values over the 1,000 data sets.

Figure 3 shows that, in the prediction of the validation data, the AUCs and specificities of the BEI-based SVM classifiers are considerably improved in comparison with those of the IC_{50}- or K_i-based classifiers. In particular, the specificities dramatically increase in the BEI-based classifiers, resulting in balanced sensitivities and specificities. The specificities of the BEI-based classifiers range from 0.64 ($BEI_{GPCR,IC_{50}}$ using MACCS+diAA) to 0.86 ($BEI_{GPCR,IC_{50}}$ using MOE2D+GPS66). In contrast, most of the specificities of the IC_{50}- or K_i-based classifiers are below 0.6, with a maximum value of 0.66 (K_{iGPCR} using MOE2D +GPS66). The differences in the mean AUCs and specificities between the BEI-based and the IC_{50}- or K_i-based classifiers are statistically significant with a P value < 0.01 for all classifier comparisons (Figure 3 and Supporting Information Table S18). The BEI-based and the IC_{50}- or K_i-based classifiers have comparable accuracy and sensitivity. These results clearly

dx.doi.org/10.1021/ci400240u | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

demonstrate that the BEI-based classifiers can substantially outperform the IC_{50}- or K_i-based classifiers, especially in predicting negative (inactive) bioactivity data in the validation data.

Statistical tests of the classifier comparisons with respect to the compound descriptors indicate that which descriptor has more predictive power than the others is dependent on the statistics used for the comparisons (Figure 4 and Supporting Information Table S19). When the AUCs are compared, MACCSF, TGT, and MOE2D have better performance than MACCS, and MACCSF and MOE2D are better than TGT (Figure 4A). On the other hand, MACCS, MACCSF, and TGT perform much better than MOE2D if the classifiers are compared with respect to accuracy and sensitivity (Figures 4B and C). The classifiers using MOE2D show higher specificities than those using other descriptors, and thus they seem to be superior to the others in predicting the negatives in the validation data (Figure 4D) but to be inferior to the others in predicting the positives (Figure 4C). For the protein descriptors, GPS66 and ABS36 are somewhat better than diAA in comparisons of AUCs, accuracies, and specificities, which is in almost agreement with the results in the cross validation tests.

Explanation for the Superior Predictive Performance of the BEI-Based SVM Classifiers. Why is the predictive power of the SVM classifiers improved, and are the values of sensitivity and specificity poised, when the BEI-based training data are used for creating the SVM classifiers? Close examination of the nature of the training data created from the BEI data type and the IC_{50} or K_i data type provides us with two plausible explanations for the improvement.

One explanation is the distinct distribution of the positives and negatives in a feature space of SVMs between the IC_{50}- or K_i-based training data and the BEI-based data. Figure 5 shows two examples of the distributions of the positives and negatives in the training data, one from the IC_{50GPCR} training data (Figure 5A) and another from the $BEI_{GPCR,IC_{50}}$ training data (Figure 5B), obtained from the principal component analyses of the data. The distributions of the positives and negatives in the IC_{50GPCR} training data largely overlap each other. In contrast, the positives and negatives in the $BEI_{GPCR,IC_{50}}$ training data are more separately distributed. Indeed, the Euclidean distance (1.541) between the centroid of the positives and of the negatives in the distribution of the $BEI_{GPCR,IC_{50}}$ training data is over twice as large as the distance (0.742) in the IC_{50GPCR} training data. To investigate whether this holds for other pairs (BEI-based versus IC_{50}- or K_i-based) of the training data, we calculated the distance between the centroid of the positives and of the negatives in each training set. Table 2 clearly shows that, with respect to all pairs of the training data, the positives and negatives in the BEI-based data are more distantly separated in a feature space than those in the IC_{50}- or K_i-based training data. More separated distribution of the positives and negatives in the training data can lead to improvement of the statistics in the cross validation tests. This explains why the statistics of the BEI-based classifiers are higher than those of the IC_{50}- or K_i-based classifiers in Figure 1. Figure 5 also shows that the diversity of the positives in the IC_{50GPCR} training data is higher than that of the positives in the $BEI_{GPCR,IC_{50}}$ training data and that the $BEI_{GPCR,IC_{50}}$ training data has higher diversity in the distribution of the negatives than the IC_{50GPCR} training data. To

Table 2. Distance between the Centroids of Positives and Negatives in the Training Data

	distance between the centroids[a]	
descriptor concatenation	IC_{50GPCR}	$BEI_{GPCR,IC_{50}}$
MACCS+GPS66	0.819 ± 0.027	1.869 ± 0.031
MACCSF+GPS66	0.621 ± 0.022	1.028 ± 0.019
TGT+GPS66	1.659 ± 0.086	7.544 ± 0.140
MOE2D+GPS66	0.742 ± 0.024	1.541 ± 0.028
MACCS+diAA	0.763 ± 0.024	1.812 ± 0.031
MACCSF+diAA	0.545 ± 0.020	0.913 ± 0.020
TGT+diAA	1.623 ± 0.096	7.534 ± 0.137
MOE2D+diAA	0.671 ± 0.024	1.474 ± 0.032
	K_{iGPCR}	BEI_{GPCR,K_i}
MACCS+GPS66	0.713 ± 0.028	1.636 ± 0.026
MACCSF+GPS66	0.436 ± 0.021	1.026 ± 0.024
TGT+GPS66	1.492 ± 0.091	5.024 ± 0.098
MOE2D+GPS66	0.480 ± 0.021	1.394 ± 0.026
MACCS+diAA	0.658 ± 0.031	1.475 ± 0.028
MACCSF+diAA	0.344 ± 0.025	0.738 ± 0.029
TGT+diAA	1.453 ± 0.086	4.973 ± 0.094
MOE2D+diAA	0.385 ± 0.023	1.196 ± 0.030
	IC_{50PK}	$BEI_{PK,IC_{50}}$
MACCS+ABS36	0.774 ± 0.033	1.303 ± 0.039
MACCSF+ABS36	0.400 ± 0.018	0.515 ± 0.028
TGT+ABS36	1.897 ± 0.078	3.219 ± 0.083
MOE2D+ABS36	0.403 ± 0.022	0.942 ± 0.026
MACCS+diAA	0.786 ± 0.030	1.306 ± 0.038
MACCSF+diAA	0.401 ± 0.027	0.520 ± 0.030
TGT+diAA	1.891 ± 0.084	3.210 ± 0.087
MOE2D+diAA	0.406 ± 0.029	0.943 ± 0.036

[a]Values of mean ± SD (over the 1000 random training data sets) are shown.

evaluate the diversity of the positives and the negatives in the training data, we calculated the mean and SD of the distances between each instance and the centroid for every training set in the 1000 random data set. Then, means of the means and SDs over the 1000 random data sets were calculated. Larger values of the means and SDs signify that the instances in the positives (or negatives) are more dispersed in a feature space. Figure 6A summarizes the comparisons of the means and SDs between the IC_{50}- or K_i-based training data and the BEI-based data. When one focuses on the "Mean" (Figure 6B), in the "Positives" column in Figure 6A, 22 of the IC_{50}- or K_i-based data show a larger value than the BEI-based data with statistical significance (P value < 0.01), while no BEI-training data have a larger value. In the "Negatives" column, larger means are obtained in the 16 BEI-based training data, and the 6 IC_{50}- or K_i-based data have larger means. Fisher's exact test for this contingency table indicates that null hypothesis (the number of both types of the training data is equally distributed in the Positives and Negatives columns in the contingency table) is rejected with statistical significance (P value = 3.6×10^{-7}). The biased distribution of the number of the training data is observed also in the contingency table about the SDs of the distances (Figure 6C), although this observation is not statistically significant (P value = 0.069). These results demonstrate that the positives of the IC_{50}- or K_i-based training data tend to be more enriched in their diversity than those of the BEI-based data, while the diversity of the negatives in the BEI-based data has a tendency to be higher than that of the

dx.doi.org/10.1021/ci400240u | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

(A)

Data type 1	Data type 2	Descriptor concatenation	Training data showing larger mean or SD			
			Positives		Negatives	
			Mean	SD	Mean	SD
$IC50_{GPCR}$	$BEI_{GPCR,IC50}$	MACCS+GPS66	$IC50_{GPCR}$	$BEI_{GPCR,IC50}$	$IC50_{GPCR}$	$BEI_{GPCR,IC50}$
$IC50_{GPCR}$	$BEI_{GPCR,IC50}$	MACCSF+GPS66	$IC50_{GPCR}$	$IC50_{GPCR}$	$IC50_{GPCR}$	$BEI_{GPCR,IC50}$
$IC50_{GPCR}$	$BEI_{GPCR,IC50}$	TGT+GPS66	$IC50_{GPCR}$	$IC50_{GPCR}$	$BEI_{GPCR,IC50}$	$IC50_{GPCR}$
$IC50_{GPCR}$	$BEI_{GPCR,IC50}$	MOE2D+GPS66	$IC50_{GPCR}$	$IC50_{GPCR}$	$BEI_{GPCR,IC50}$	$BEI_{GPCR,IC50}$
$IC50_{GPCR}$	$BEI_{GPCR,IC50}$	MACCS+diAA	$IC50_{GPCR}$	$BEI_{GPCR,IC50}$	$IC50_{GPCR}$	$BEI_{GPCR,IC50}$
$IC50_{GPCR}$	$BEI_{GPCR,IC50}$	MACCSF+diAA	$IC50_{GPCR}$	$BEI_{GPCR,IC50}$	$IC50_{GPCR}$	$IC50_{GPCR}$
$IC50_{GPCR}$	$BEI_{GPCR,IC50}$	TGT+diAA	$IC50_{GPCR}$	$IC50_{GPCR}$	$BEI_{GPCR,IC50}$	$IC50_{GPCR}$
$IC50_{GPCR}$	$BEI_{GPCR,IC50}$	MOE2D+diAA	$IC50_{GPCR}$	$BEI_{GPCR,IC50}$	$IC50_{GPCR}$	$IC50_{GPCR}$
Ki_{GPCR}	$BEI_{GPCR,Ki}$	MACCS+GPS66	Ki_{GPCR}		$BEI_{GPCR,Ki}$	$BEI_{GPCR,Ki}$
Ki_{GPCR}	$BEI_{GPCR,Ki}$	MACCSF+GPS66	Ki_{GPCR}	Ki_{GPCR}	$BEI_{GPCR,Ki}$	Ki_{GPCR}
Ki_{GPCR}	$BEI_{GPCR,Ki}$	TGT+GPS66	Ki_{GPCR}	Ki_{GPCR}	$BEI_{GPCR,Ki}$	$BEI_{GPCR,Ki}$
Ki_{GPCR}	$BEI_{GPCR,Ki}$	MOE2D+GPS66	Ki_{GPCR}	Ki_{GPCR}	$BEI_{GPCR,Ki}$	Ki_{GPCR}
Ki_{GPCR}	$BEI_{GPCR,Ki}$	MACCS+diAA	Ki_{GPCR}	Ki_{GPCR}	Ki_{GPCR}	$BEI_{GPCR,Ki}$
Ki_{GPCR}	$BEI_{GPCR,Ki}$	MACCSF+diAA		Ki_{GPCR}		$BEI_{GPCR,Ki}$
Ki_{GPCR}	$BEI_{GPCR,Ki}$	TGT+diAA	Ki_{GPCR}	Ki_{GPCR}	$BEI_{GPCR,Ki}$	$BEI_{GPCR,Ki}$
Ki_{GPCR}	$BEI_{GPCR,Ki}$	MOE2D+diAA		Ki_{GPCR}		$BEI_{GPCR,Ki}$
$IC50_{PK}$	$BEI_{PK,IC50}$	MACCS+ABS36	$IC50_{PK}$	$BEI_{PK,IC50}$	$BEI_{PK,IC50}$	$BEI_{PK,IC50}$
$IC50_{PK}$	$BEI_{PK,IC50}$	MACCSF+ABS36	$IC50_{PK}$	$IC50_{PK}$	$BEI_{PK,IC50}$	$IC50_{PK}$
$IC50_{PK}$	$BEI_{PK,IC50}$	TGT+ABS36	$IC50_{PK}$	$IC50_{PK}$	$BEI_{PK,IC50}$	$BEI_{PK,IC50}$
$IC50_{PK}$	$BEI_{PK,IC50}$	MOE2D+ABS36	$IC50_{PK}$	$IC50_{PK}$	$BEI_{PK,IC50}$	$IC50_{PK}$
$IC50_{PK}$	$BEI_{PK,IC50}$	MACCS+diAA	$IC50_{PK}$	$BEI_{PK,IC50}$	$BEI_{PK,IC50}$	
$IC50_{PK}$	$BEI_{PK,IC50}$	MACCSF+diAA	$IC50_{PK}$	$BEI_{PK,IC50}$	$BEI_{PK,IC50}$	
$IC50_{PK}$	$BEI_{PK,IC50}$	TGT+diAA	$IC50_{PK}$	$IC50_{PK}$	$BEI_{PK,IC50}$	$BEI_{PK,IC50}$
$IC50_{PK}$	$BEI_{PK,IC50}$	MOE2D+diAA	$IC50_{PK}$		$BEI_{PK,IC50}$	

(B)

	Positives	Negatives
Number of the IC50- or Ki-based data	22	6
Number of the BEI-based data	0	16

Mean

SD

(C)

	Positives	Negatives
Number of the IC50- or Ki-based data	15	8
Number of the BEI-based data	7	13

Figure 6. Comparisons of the diversity of the positives and negatives in the training data. (A) Means of the means and SDs of the distances between each instance and the centroid were calculated for the positives and negatives in the training data. The means were compared between the classifier from "Data type 1" and from "Data type 2". If a classifier from Data type 1 has a large value with statistical significance (P value < 0.01), "IC_{50GPCR}", "K_{iGPCR}", or "IC_{50PK}" is shown. If a classifier from "Data type 2" has a large value, "$BEI_{GPCR,IC_{50}}$", "BEI_{GPCR,K_i}", or "$BEI_{PK,IC_{50}}$" is shown. When no statistically significant difference between the means is observed, the cell is void. For more details regarding the statistical tests, see Supporting Information Table S20. (B) Contingency table of the number of the training data in the "Positives" and "Negatives" columns in part A with respect to the "Mean" of the distances. (C) Contingency table of the number of the training data in the Positives and Negatives columns in part A with respect to the "SD" of the distances.

negatives in the IC_{50}- or K_i-based data. The higher diversity of the negatives could culminate in the improvement of the specificities of the BEI-based SVM classifiers in the prediction of the validation data. We postpone an explanation for the results showing that the lower diversity in the positives in the BEI-based data does not cause the decrease in the sensitivity of the data, because we currently have no method for further investigating the results with respect to the distribution patterns.

Another explanation is a bias in the number of the positives and negatives in the training data. As shown in Table 1, the training data sets differ in the number of the positives and negatives, and the ratios of the positives to negatives are not homogeneous. In general, a higher diversity of instances and better predictive performance can be expected if there are many instances in a training data. Therefore, the number of positives and negatives could influence the predictive power of SVM classifiers. Figure 7 shows a plot of the logarithm of the ratio of the sensitivity (Sens) to specificity (Spec) of the SVM classifier in the prediction of the validation data to the logarithm of the ratio of the number of the positives (no. positives) to that of the negatives (no. negatives) in the training data. This figure demonstrates that there is a strong positive correlation (r = 0.708, P value = 1.8×10^{-8}) between log(Sens/Spec) and log(no. positives/no. negatives). This indicates that a SVM classifier created from a training data set with a highly biased no. positives/no. negatives ratio shows a highly biased Sens/Spec ratio in the prediction of the validation data. Intriguingly,

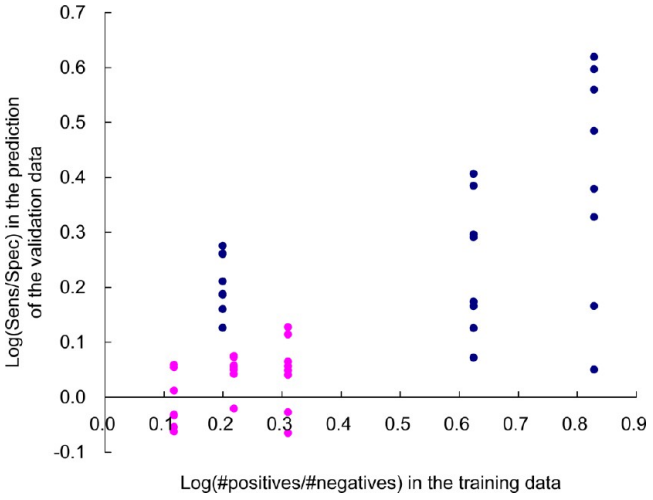


Figure 7. Plot of the 48 training data types. The vertical axis is the logarithm of the ratio of the sensitivity (Sens) to specificity (Spec) in the prediction of the validation data. The horizontal axis is the logarithm of the ratio of the number of positives to that of negatives in the training data. The IC_{50}- and K_i-based training data are colored blue, and the BEI-based data are in magenta.

while the training data based on IC_{50} or K_i (shown by blue) have a large bias in the no. positives/no. negatives ratio, the ratio is relatively small in the training data based on BEI (shown by magenta). These low biases of the no. positives/no.

dx.doi.org/10.1021/ci400240u | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

negatives ratio in the BEI-based training data can result in the more poised sensitivity and specificity in the prediction of the validation data.

DISCUSSION

The SVM methods, compound descriptors, and protein descriptors used here, and their relatives, have been frequently adopted in previous studies. What is unique about the current study is the use of ligand efficiency when creating the training data. Using the BEI-based and IC_{50}- or K_i-based training data derived from GPCRSARfari and KinaseSARfari, we showed that the SVM classifiers created from the BEI-based training data have more discriminative power than the classifiers from the IC_{50}- or K_i-based data in the cross validation tests. The values of AUC, accuracy, sensitivity, and specificity of 0.84−0.96 are comparable to the results provided in previous studies.[1−10] Application of the SVM classifiers to the validation data created from the latest versions of the databases demonstrated that the BEI-based classifiers offered better prediction than the IC_{50}- or K_i-based classifiers, especially in the prediction of the negative data as reflected by the increase in specificities. Furthermore, the sensitivities and specificities are more poised in the prediction using the BEI-based classifiers. The results indicating that the BEI-based classifiers are superior to the IC_{50}- or K_i-based classifiers are independent of the compound and protein descriptors used, the measurement type (IC_{50} and K_i), and the target protein (GPCRs and PKs) (Figures 1 and 3).

The results showing the superiority of the BEI-based classifiers are independent also of the setting of threshold for defining positives and negatives used here. As shown in the Methods section, we defined compound−protein pairs with IC_{50} (or K_i) ≤ 1 μM as positives and pairs with IC_{50} (or K_i) $\geq$ 10 μM as negatives. To check whether a change of the threshold influences the results here, we created training data based on two other types of thresholds and conducted cross validation tests and validation of the constructed SVM classifiers. One training data defines compound−protein pairs with IC_{50} (or K_i) $\leq$ 100 nM as positives and pairs with IC_{50} (or K_i) $\geq$ 10 μM as negatives. Another training data defines compound−protein pairs with IC_{50} (or K_i) $\leq$ 500 nM as positives and pairs with IC_{50} (or K_i) ≥ 5 μM as negatives. For the BEI-based training data, BEIs corresponding to the IC_{50} (or K_i) values above were transferred from IC_{50}s or K_is based on the BEI−pIC_{50} (or pK_i) plots in Supporting Information Figure S1. Cross-validation tests indicate that, in both types of the training data, the BEI-based classifiers outperform the IC_{50}- or K_i-based classifiers with statistical significance (Supporting Information Tables S21 and S23). Also in validation of the SVM classifiers, the BEI-based classifiers show better performance than the IC_{50}- or K_i-based classifiers (Supporting Information Tables S22 and S24). Furthermore, like in Figure 3, AUCs and specificities are considerably improved and sensitivities and specificities are more balanced (Supporting Information Tables S22 and S24). Although only three settings of threshold are tested in this study, the superiority of the BEI-based training data could be robust to the change of the setting of threshold for defining positives and negatives. A possible drawback of our method for creating training data is the discard of intermediate bioactivity data. To comprehensively treat all range of bioactivity data without discarding intermediates, machine learning methods using a regression model such as support vector regression may be more suitable.

Among the compound descriptors used in this study, the classifiers using MACCS and MACCSF perform better than those using TGT and MOE2D in the cross-validation tests (Figure 2 and Supporting Information Table S17). The good discriminating power of MACCS and MACCSF in the cross-validation tests may be partially attributed to the simplicity of these descriptors. MACCS and MACCSF fingerprints are composed of only 166 elements, and they represent a compound as a vector in terms of the presence/absence (in MACCSF) or the observed number (in MACCSF) of 166 simple 2D structures of the compounds.[26] Thus, it is highly probable that two compounds that appear to be very different to a chemist display high similarity. This can lead to the two instances reciprocally more similar to each other in a training data set using MACCS or MACCSF than in a data set using other compound descriptors and, thus, to higher discriminative performance of the classifiers created from the training data using MACCS or MACCSF. On the other hand, the prediction of the validation data does not show that classifiers using MACCS and MACCSF are superior to those using TGT and MOE2D. Which descriptor has more predictive power is dependent on the statistics used for comparing the SVM classifiers (Figure 4 and Supporting Information Table S19). For example, MOE2D is better at predicting the negatives than the other descriptors (high specificity) but weak in the prediction of the positives (low sensitivity). The disagreement between the results in the cross-validation tests and those in the prediction of the validation data implies that which compound descriptor is better cannot be determined only by cross-validation tests. Any compound descriptor should be applied to the prediction target data, even though a result of a cross-validation test using a descriptor is not better than other descriptors. In contrast to the compound descriptors, the classifiers using the protein descriptor based on ligand binding sites in the tertiary structures of GPCRs and PKs (GPS66 and ABS36) show better performance both in the cross-validation tests and in the prediction of the validation data compared with the classifiers using diAA based on amino acid sequences. This result clearly demonstrates that the ligand binding site-based descriptors are more informative than the sequence-based descriptor in machine learning methods using SVMs. Many crystal structures of drug target proteins, especially of GPCRs, have been rapidly solved in recent years. As a lot of tertiary structures of drug target proteins will be stored in public databases in future, the importance of tertiary structure-based descriptors will further increase.

Although we used BEI as a measure of ligand efficiency, it is highly probable that SVM classifiers reflecting ligand efficiency offer better prediction than classifiers reflecting IC_{50} or K_i, even though other measures such as the original LE,[21] NBEI,[23] nBEI,[23] and mBEI[23] are used to create training data instead of BEI. This is because BEI and these measures are similar to each other in that they combine ligand potency (or binding energy) with the total size of the ligand represented by molecular weight and the number of non-hydrogen atoms. Molecular weight and the number of non-hydrogen atoms in compounds highly and positively correlate (data not shown), thus leading to highly positive correlations between BEI and other BEI-similar measures. As a consequence, when positive and negative instances are defined by a threshold, most instances in two training data sets (one based on BEI and another based on a BEI-similar measure) can be identical. Further studies, however, will be needed to investigate whether using a training data set

dx.doi.org/10.1021/ci400240u | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

based on another BEI-dissimilar measure of ligand efficiency can influence the results obtained in this study. For example, in ChEMBL, SEI (defined as pIC_{50} (or pK_i) divided by the polar surface area ($Å^2$) of the compound) is available as another measure of ligand efficiency.[20] In this study, we did not utilize SEI because very low positive correlation was observed between SEI and pIC_{50} (or pK_i) (data not shown) and thus SEIs corresponding to IC_{50}s (or K_is) of 1 and 10 μM were not determined as the thresholds for positives and negatives. To create a training data set for machine learning methods using SEI thresholds, other criteria for setting the thresholds will have to be developed. Using the SEI-based training data, researchers can validate whether the results presented here hold when other measures of ligand efficiency are used.

SVM methods have been most frequently applied to predicting interactions between compounds and drug target proteins based on bioactivity databases in previous studies,[1−13] and we also used the methods in this study to compare the values of AUCs, accuracies, sensitivities, and specificities with those in previous studies. However, there are many other machine learning methods such as boosting, naïve Bayes, neural network, and random forest. Further studies will be needed whether the superiority of ligand efficiency observed here holds when other machine learning methods are used.

CONCLUSION

In this study, we demonstrated that when SVM classifiers were created from training data based on BEI, they displayed higher discriminative or predictive power than classifiers created from training data based on IC_{50} or K_i. There are two plausible explanations for this observation. One is the more separated distribution of the positives and negatives in the BEI-based training data than in the IC_{50}- or K_i-based data in a feature space of SVMs, and the higher diversity of the negatives in the BEI-based training data. Another is the more balanced number of the positives and negatives in the BEI-based training data than in the IC_{50}- or K_i-based data. Although the results presented here are independent of the descriptors used, the measurement type (IC_{50} or K_i), the target protein (GPCRs or PKs), and the setting of threshold for defining positives and negatives, there is nevertheless a possibility that the results may depend on the nature of the bioactivity data used in this study, that is, GPCRSARfari and KinaseSARfari. For example, the plots of pIC_{50} (or pKi) versus BEI used to define the thresholds of BEI for the positives and negatives in Figure S1 are dependent on the current versions of GPCRSARfari and KinaseSARfari. The addition of a huge amount of new data in future or the use of a different bioactivity database may influence the plots and thus the results obtained here. It is essential to investigate whether the result that classifiers incorporating ligand efficiency are more predictive than classifiers incorporating IC_{50} or K_i is robust to the addition of new data in future or to a variety of bioactivity databases utilized for machine learning. Although several matters require further study, our findings strongly suggest that a training data set based on ligand efficiency as well as classical IC_{50}, EC_{50}, K_d, K_i values is important when creating a classifier using a machine learning approach based on bioactivity data.

ASSOCIATED CONTENT

Supporting Information

Figure S1 shows the plots of the bioactivity data used in this study. Figure S2 is the distributions of the positives and negatives in the training data using the three principal components. Tables S1 and S3 provide the multiple alignments of the 66 or 36 amino acid residues of the GPCRs or PKs used in this study. Table S2 shows the physicochemical properties of the amino acids used as the protein descriptors. Tables S4−S15 provide the training and validation data used in this study. Tables S16−S24 show the results of the statistical tests of the difference of the mean values of the statistics. This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

*E-mail: sugaya@pharmadesign.co.jp.

Notes

The authors declare no competing financial interest.

REFERENCES

(1) Bock, J. R.; Gough, D. A. Virtual screen for ligands of orphan G protein-coupled receptors. *J. Chem. Inf. Model.* **2005**, *45*, 1402−1414.
(2) Jacob, L.; Vert, J. P. Protein-ligand interaction prediction: an improved chemogenomics approach. *Bioinformatics* **2008**, *24*, 2149−2156.
(3) Jacob, L.; Hoffmann, B.; Stoven, V.; Vert, J. P. Virtual screening of GPCRs: an *in silico* chemogenomics approach. *BMC Bioinform.* **2008**, *9*, 363.
(4) Weill, N.; Rognan, D. Development and validation of a novel protein-ligand fingerprint to mine chemogenomic space: application to G protein-coupled receptors and their ligands. *J. Chem. Inf. Model.* **2009**, *49*, 1049−1062.
(5) Wang, F.; Liu, D.; Wang, H.; Luo, C.; Zheng, M.; Liu, H.; Zhu, W.; Luo, X.; Zhang, J.; Jiang, H. Computational screening for active compounds targeting protein sequences: methodology and experimental validation. *J. Chem. Inf. Model.* **2011**, *51*, 2821−2828.
(6) Wang, Y.-C.; Zhang, C.-H.; Deng, N.-Y.; Wang, Y. Kernel-based data fusion improves the drug-protein interaction prediction. *Comput. Biol. Chem.* **2011**, *35*, 353−362.
(7) Yabuuchi, H.; Niijima, S.; Takematsu, H.; Ida, T.; Hirokawa, T.; Hara, T.; Ogawa, T.; Minowa, Y.; Tsujimoto, G.; Okuno, Y. Analysis of multiple compound-protein interactions reveals novel bioactive molecules. *Mol. Syst. Biol.* **2011**, *7*, 472.
(8) Yu, H.; Chen, J.; Xu, X.; Li, Y.; Zhao, H.; Fang, Y.; Li, X.; Zhou, W.; Wang, W.; Wang, Y. A systematic prediction of multiple drug-target interactions from chemical, genomic, and pharmacological data. *PLoS One* **2012**, *7*, e37608.
(9) Cao, D. S.; Liu, S.; Xu, Q. S.; Lu, H. M.; Huang, J. H.; Hu, Q. N.; Liang, Y. Z. Large-scale prediction of drug-target interactions using protein sequences and drug topological structures. *Anal. Chim. Acta* **2012**, *752*, 1−10.
(10) Buchwald, F.; Richter, L.; Kramer, S. Predicting a small molecule-kinase interaction map: A machine learning approach. *J. Cheminform.* **2011**, *3*, 22.
(11) Geppert, H.; Humrich, J.; Stumpfe, D.; Gärtner, T.; Bajorath, J. Ligand prediction from protein sequence and small molecule information using support vector machines and fingerprint descriptors. *J. Chem. Inf. Model.* **2009**, *49*, 767−779.
(12) Wassermann, A. M.; Geppert, H.; Bajorath, J. Ligand prediction for orphan targets using support vector machines and various target-ligand kernels is dominated by nearest neighbor effects. *J. Chem. Inf. Model.* **2009**, *49*, 2155−2167.
(13) Faulon, J.-L.; Misra, M.; Martin, S.; Sale, K.; Sapra, R. Genome scale enzyme-metabolite and drug-target interaction predictions using the signature molecular descriptor. *Bioinformatics* **2008**, *24*, 225−233.
(14) *PDSP*. http://pdsp.med.unc.edu/indexR.html (accessed March 14, 2013).
(15) Kanehisa, M.; Goto, S.; Sato, Y.; Furumichi, M.; Tanabe, M. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res.* **2012**, *40*, D109−D114.

(16) Okuno, Y.; Tamon, A.; Yabuuchi, H.; Niijima, S.; Minowa, Y.; Tonomura, K.; Kunimoto, R.; Feng, C. GLIDA: GPCR–ligand database for chemical genomics drug discovery–database and tools update. *Nucleic Acids Res.* **2008**, *36*, D907−D912.
(17) Liu, T.; Lin, Y.; Wen, X.; Jorissen, R. N.; Gilson, M. K. BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities. *Nucleic Acids Res.* **2007**, *35*, D198−D201.
(18) *GVK BIO Kinase Inhibitor Databases Reviews, Pricing, Demos*. http://www.jazdlifesciences.com/pharmatech/company/GVK-BIO/Kinase-Inhibitor-Databases.htm?supplierId=30003173&productId=695879 (accessed March 14, 2013).
(19) Fabian, M. A.; Biggs, W. H., 3rd; Treiber, D. K.; Atteridge, C. E.; Azimioara, M. D.; Benedetti, M. G.; Carter, T. A.; Ciceri, P.; Edeen, P. T.; Floyd, M.; Ford, J. M.; Galvin, M.; Gerlach, J. L.; Grotzfeld, R. M.; Herrgard, S.; Insko, D. E.; Insko, M. A.; Lai, A. G.; Lélias, J. M.; Mehta, S. A.; Milanov, Z. V.; Velasco, A. M.; Wodicka, L. M.; Patel, H. K.; Zarrinkar, P. P.; Lockhart, D. J. A small molecule-kinase interaction map for clinical kinase inhibitors. *Nat. Biotechnol.* **2005**, *23*, 329−336.
(20) Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* **2012**, *40*, D1100−D1107.
(21) Hopkins, A. L.; Groom, C. R.; Alex, A. Ligand efficiency: a useful metric for lead selection. *Drug Discov. Today* **2004**, *9*, 430−431.
(22) Abad-Zapatero, C.; Metz, J. T. Ligand efficiency indices as guideposts for drug discovery. *Drug Doscov. Today* **2005**, *10*, 464−469.
(23) Abad-Zapatero, C.; Perišić, O.; Wass, J.; Bento, A. P.; Overington, J.; Al-Lazikani, B.; Johnson, M. E. Ligand efficiency indices for an effective mapping of chemico-biological space: the concept of an atlas-like representation. *Drug Discov. Today* **2010**, *15*, 804−811.
(24) *GPCRSARfari ftp site*. ftp://ftp.ebi.ac.uk/pub/databases/chembl/GPCRSARfari/releases/ (accessed November 1, 2012).
(25) *KinaseSARfari ftp site*. ftp://ftp.ebi.ac.uk/pub/databases/chembl/KinaseSARfari/releases/ (accessed November 1, 2012).
(26) *Molecular Operating Environment*. http://www.chemcomp.com/software.htm (accessed November 1, 2012).
(27) Rose, P. W.; Bi, C.; Bluhm, W. F.; Christie, C. H.; Dimitropoulos, D.; Dutta, S.; Green, R. K.; Goodsell, D. S.; Prlic, A.; Quesada, M.; Quinn, G. B.; Ramos, A. G.; Westbrook, J. D.; Young, J.; Zardecki, C.; Berman, H. M.; Bourne, P. E. The RCSB Protein Data Bank: new resources for research and education. *Nucleic Acids Res.* **2013**, *41*, D475−D482.
(28) Punta, M.; Coggill, P. C.; Eberhardt, R. Y.; Mistry, J.; Tate, J.; Boursnell, C.; Pang, N.; Forslund, K.; Ceric, G.; Clements, J.; Heger, A.; Holm, L.; Sonnhammer, E. L.; Eddy, S. R.; Bateman, A.; Finn, R. D. The Pfam protein families database. *Nucleic Acids Res.* **2012**, *40*, D290−D301.
(29) Eddy, S. R. Accelerated Profile HMM Searches. *PLoS Comput. Biol.* **2011**, *7*, e1002195.
(30) Wallace, A. C.; Laskowski, R. A.; Thornton, J. M. LIGPLOT: a program to generate schematic diagrams of protein-ligand interactions. *Protein Eng.* **1995**, *8*, 127−134.
(31) Vroling, B.; Sanders, M.; Baakman, C.; Borrmann, A.; Verhoeven, S.; Klomp, J.; Oliveira, L.; de Vlieg, J.; Vriend, G. GPCRDB: information system for G protein-coupled receptors. *Nucleic Acids Res.* **2011**, *39*, D309−D319.
(32) Huang, D.; Zhou, T.; Lafleur, K.; Nevado, C.; Caflisch, A. Kinase selectivity potential for inhibitors targeting the ATP binding site: a network analysis. *Bioinformatics* **2010**, *26*, 198−204.
(33) *LIBSVM—A Library for Support Vector Machines*. http://www.csie.ntu.edu.tw/~cjlin/libsvm/ (accessed November 1, 2012).
(34) *GPU-accelerated LIBSVM*. http://mklab.iti.gr/project/GPU-LIBSVM (accessed November 7, 2012).
(35) *LIBSVM tools*. http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/#roc_curve_for_binary_svm (accessed November 7, 2012).
(36) *The Comprehensive R Archive Network*. http://cran.r-project.org/ (accessed March 13, 2013).