JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

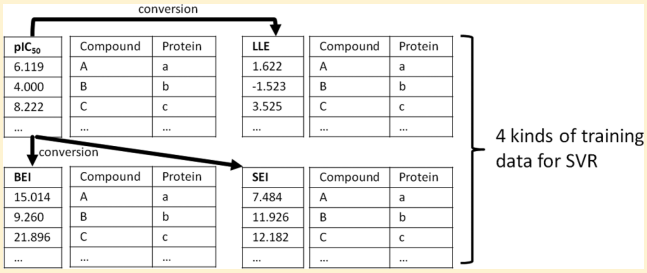
Ligand Efficiency-Based Support Vector Regression Models for Predicting Bioactivities of Ligands to Drug Target Proteins

Nobuyoshi Sugaya*

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

Supporting Information

ABSTRACT: The concept of ligand efficiency (LE) indices is widely accepted throughout the drug design community and is frequently used in a retrospective manner in the process of drug development. For example, LE indices are used to investigate LE optimization processes of already-approved drugs and to re-evaluate hit compounds obtained from structure-based virtual screening methods and/or high-throughput experimental assays. However, LE indices could also be applied in a prospective manner to explore drug candidates. Here, we describe the construction of machine learning-based regression models in which LE indices are adopted as an end point and show that LE-based regression models can outperform regression models based on pIC_{50} values. In addition to pIC_{50} values traditionally used in machine learning studies based on chemogenomics data, three representative LE indices (ligand lipophilicity efficiency (LLE), binding efficiency index (BEI), and surface efficiency index (SEI)) were adopted, then used to create four types of training data. We constructed regression models by applying a support vector regression (SVR) method to the training data. In cross-validation tests of the SVR models, the LE-based SVR models showed higher correlations between the observed and predicted values than the pIC_{50}-based models. Application tests to new data displayed that, generally, the predictive performance of SVR models follows the order SEI > BEI > LLE > pIC_{50}. Close examination of the distributions of the activity values (pIC_{50}, LLE, BEI, and SEI) in the training and validation data implied that the performance order of the SVR models may be ascribed to the much higher diversity of the LE-based training and validation data. In the application tests, the LE-based SVR models can offer better predictive performance of compound−protein pairs with a wider range of ligand potencies than the pIC_{50}-based models. This finding strongly suggests that LE-based SVR models are better than pIC_{50}-based models at predicting bioactivities of compounds that could exhibit a much higher (or lower) potency.



INTRODUCTION

Ligand efficiency (LE) indices are widely accepted in drug design community as a metric for prioritizing drug candidate compounds by associating their potency with physicochemical properties such as molecular weight (MW), lipophilicity (e.g., log *P*), and total polar surface area (TPSA). LE indices have been frequently used to investigate the LEs of already-approved drugs and the LE optimization steps in the drug development process.[1,2] For example, a recent study by Hopkins and colleagues[1] has shown that recently marketed oral drugs were highly optimized, with respect to their LEs for their targets. LE indices have also been used to re-evaluate the potential of hit compounds as drug candidates obtained from structure-based virtual screening methods and/or high-throughput experimental assays.[3−6] For example, Tanaka and colleagues[4] reported the effectiveness of LE-guided selection of leadlike compounds from virtual screening hits. In those studies, LE indices provided essential information for facilitating drug development processes. However, although many previous studies have utilized LE indices in a retrospective manner, only a few studies have used LE indices in a prospective manner or considered LE indices as an objective function for seed/lead optimization. By adopting LE indices as an end point, one can develop novel methods for drug candidate prediction. Rather than simply utilizing LEs as one of helpful information to prioritize compounds afterward, one can instead directly incorporate LEs for evaluating drug candidates beforehand.

Machine learning approaches are one of the appropriate approaches to develop a novel method that utilizes LEs as an objective function for seed/lead optimization. In particular, supervised machine learning approaches have been frequently adopted for constructing classification models for active and inactive compound−protein pairs or for creating regression models of quantitative structure−activity relationships of ligands based on chemogenomics data. These classification and/or regression models have been applied to prospectively and systematically predict bioactivities of novel compound−protein pairs.[7−19] In previous studies, various databases relevant to chemogenomics, such as BindingDB,[20] ChEMBL,[21]

Received: June 3, 2014
Published: September 13, 2014

GLIDA,[22] and KEGG,[23] have been utilized as data sources for constructing machine learning-based classification and regression models.[7−19] These databases register bioactivity data of compound−protein pairs as ligand potency, such as IC_{50}, EC_{50}, K_d, and K_i. As expected by this observation, most machine learning methods in previous studies have been potency-centric, i.e., simply serving IC_{50}, EC_{50}, K_d, and K_i values as end points in making training data for machine learning. Recently, however, we have reported that when training data was created using LE, rather than ligand potency, as the end point, the support vector machine (SVM) classification models yielded from the training data had better performance than the models from the training data using IC_{50} or K_i as the end point.[24] This observation strongly suggests that utilizing LE for machine learning can improve the performance of machine learning-based classification models.[24]

Although we have demonstrated that LE-based SVM classification models showed better predictive performance than IC_{50}- or K_i-based models in our previous study,[24] only BEI was used as a representative LE. Whether the superiority of LE-based classification models also holds when other LE indices, such as SEI and LLE, are used remains to be tested. Therefore, we created machine learning models to predict bioactivities of compound−protein pairs by using not only pIC_{50} and BEI, but also SEI and LLE as end points. The objective of this study is to reconfirm that machine learning models constructed using LE can outperform models based on pIC_{50}, and also to compare predictive performances among LE-based models. In our SVM-based classification method, the thresholds for defining positives and negatives (positives, IC_{50} (or K_i) < 1 μM; negatives, IC_{50} (or K_i) > 10 μM)[24] resulted in intermediate data between the thresholds being disregarded. To effectively utilize as much bioactivity data as possible, we adopted a regression method instead of a classification method in this study.

METHODS

Bioactivity Data. We retrieved bioactivity data from ChEMBL15,[8] GPCRSARfari ver. 2,[25] and KinaseSARfari ver. 4[26] databases. Bioactivity data associated with only ion channels were collected from ChEMBL. Then, this study focuses on three target protein families: G protein-coupled receptors (GPCRs), protein kinases (PKs), and ion channels (ICs). We only used bioactivity data recorded as IC_{50} values, as these were the most abundant in the databases. Furthermore, only "assay type B" (where "B" denotes "binding") bioactivity data were collected in order to exclude cell-based assay data that did not directly measure the "binding" between the compound and the protein.

Using these retrieved data, we created four types of training data by adopting each pIC_{50}, LLE, BEI, or SEI as an end point (Figure 1). The training data differ only with respect to their activity values. All IC_{50} values in the bioactivity data were converted to pIC_{50}s. As LE indices, LLE, BEI, and SEI are available in ChEMBL. These are defined as follows:[27−29]

$$\mathrm{LLE} = \mathrm{pIC}_{50} - A\log P$$

$$\mathrm{BEI} = \frac{\mathrm{pIC}_{50}}{\mathrm{MW}}$$

and

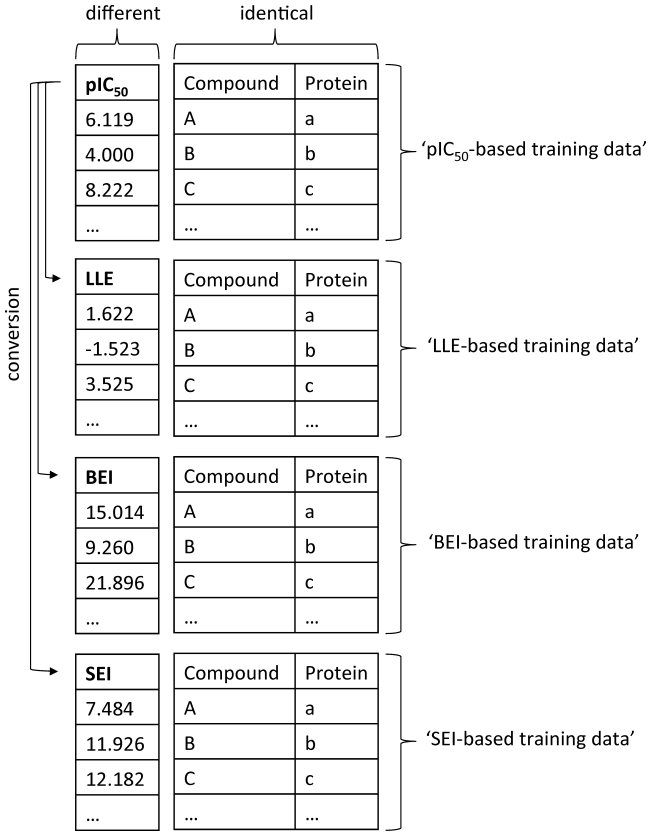


Figure 1. Procedure for creating four types of training data. pIC_{50}-, LLE-, BEI-, and SEI-based training data only differ in the activity values used for constructing the SVR models. The compound−protein pairs—and, therefore, feature vectors representing the pairs—are all identical among these training data. Validation data used for the application tests were created in the same way.

$$\mathrm{SEI} = \frac{\mathrm{pIC}_{50}}{\mathrm{TPSA}/100}$$

where A log P, MW, and TPSA are log P (calculated by the method of Ghose and Crippen[30]), molecular weight (provided in kDa), and total polar surface area (measured in Å^2) of compound, respectively. To convert pIC_{50} values to LE indices, MW and TPSA of compounds were calculated using the software Molecular Operating Environment (MOE)[31] (ver. 2011.10). A log P values of compounds were obtained from ChEMBL, GPCRSARfari, and KinaseSARfari. If the A log P value for a compound was absent from the databases, bioactivity data associated with that compound were discarded. If there were redundant instances (compound−protein pairs) in the data (i.e., the activity value, compound ID, and protein ID were all the same), all instances except one were removed. When a compound−protein pair had different activity values in two or more instances, all instances were retained, since we could not ascertain which activity value was correct.

New data in the later versions (ChEMBL16, GPCRSARfari ver. 3, and KinaseSARfari ver. 5.01) were used to validate our SVR models constructed from the previous versions. As with the training data, only bioactivity data recorded as IC_{50} and belonging to "assay type B" were collected. All redundant instances between the training data and validation data were removed from the validation data. Redundant instances within the validation data were also removed, except for one instance.

dx.doi.org/10.1021/ci5003262 | J. Chem. Inf. Model. 2014, 54, 2751−2763

If a compound−protein pair had different activity values in two or more instances between the training data and validation data, or within the validation data, all instances were retained.

The numbers of instances in the training and validation data are listed in Table 1. All training and validation data can be obtained from the Supporting Information.

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

target protein family	training data	validation data
GPCR	78 114	1553
PK	70 423	6266
IC	28 352	1999

Compound and Protein Descriptors. We adopted three types of compound descriptors: MACCS fingerprint,[32] physicochemical properties of two-dimensional (2D) structures of compounds calculated by MOE (called "MOE2D" hereafter), and compound fingerprint calculated by the software OpenBabel[33] using the option "−FP2" (called "OBFP2" hereafter). If a compound was registered as a salt in the original databases, the compound was desalted (an ion partner was eliminated using the "database wash" tool in MOE with default parameters) before calculating the descriptors. MACCS, MOE2D, and OBFP2 represent distinct profiles of compounds. MACCS converts a compound to a vector composed of the presence/absence of a series of substructures.[32] MOE2D is a set of physicochemical properties of compound.[31] OBFP2 is a path-based fingerprint that indexes compound fragments based on linear segments of up to 7 atoms.[34] MACCS, MOE2D, and OBFP2 descriptors are composed of 166, 186, and 1024 elements, respectively.

We represented drug target proteins in the bioactivity data as frequency of dimers of amino acids in protein amino acid sequence (called "diAA" (frequencies of dimers of amino acid)

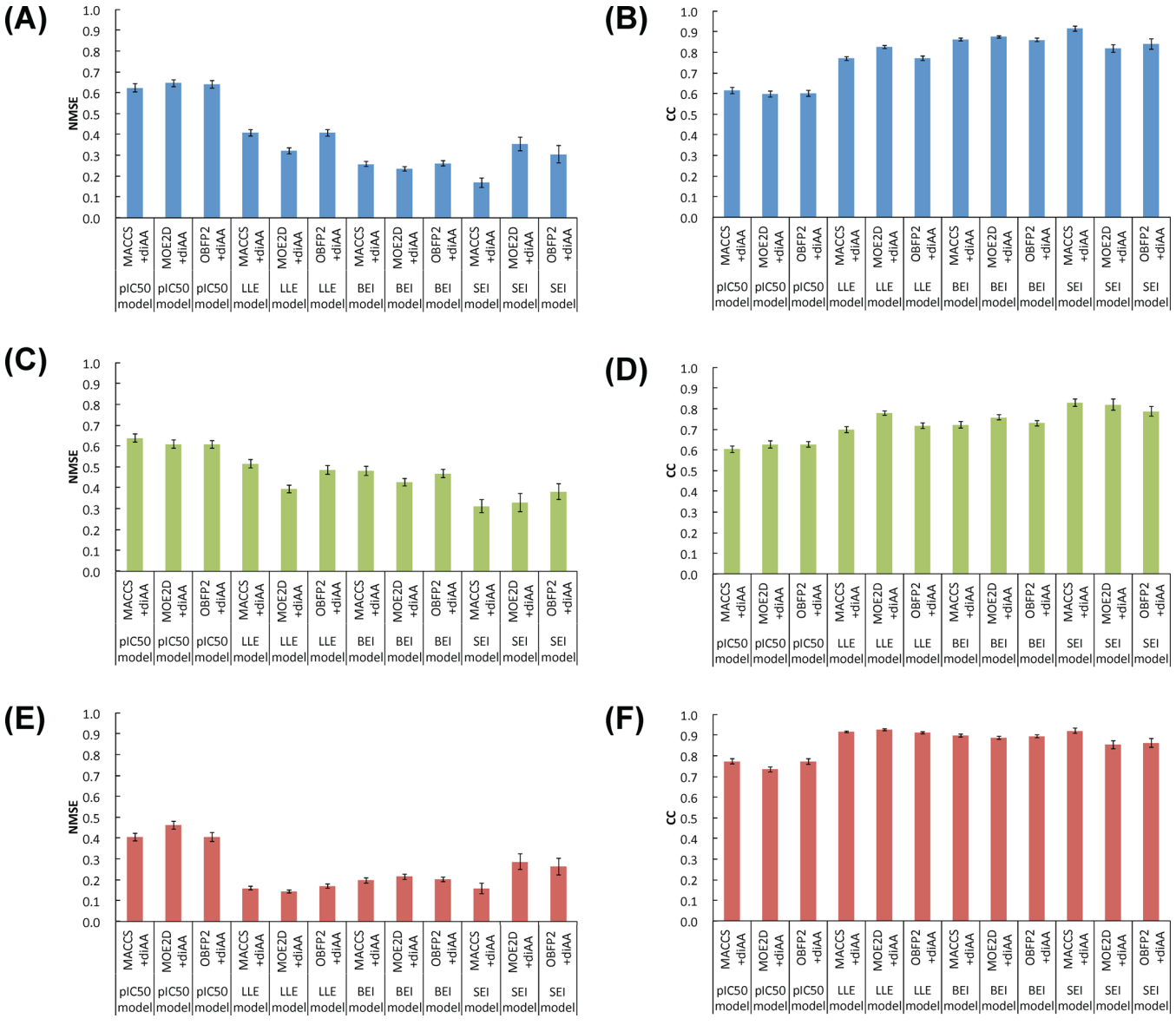


Figure 2. NMSEs and CCs in the cross-validation tests: (A) NMSEs for GPCR, (B) CCs for GPCR, (C) NMSEs for PK, (D) CCs for PK, (E) NMSEs for IC, and (F) CCs for IC. Standard deviations (SDs) are shown by error bars. See Table S1 in the Supporting Information for statistical tests on the differences in the mean values of NMSEs or CCs among the SVR models.

dx.doi.org/10.1021/ci5003262 | J. Chem. Inf. Model. 2014, 54, 2751−2763

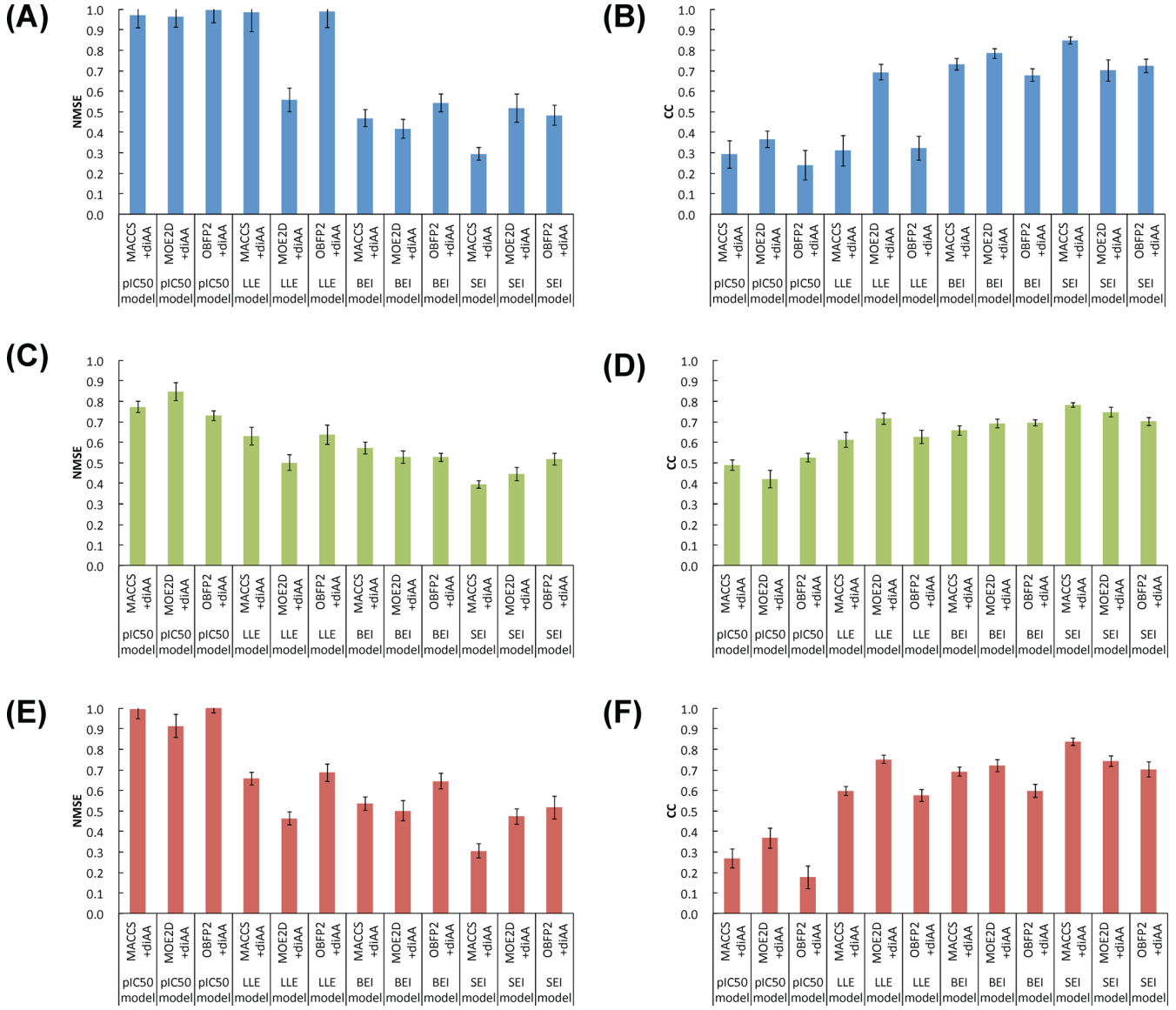


Figure 3. NMSEs and CCs in the application tests. (A) NMSEs for GPCR, (B) CCs for GPCR, (C) NMSEs for PK, (D) CCs for PK, (E) NMSEs for IC, and (F) CCs for IC. SDs are shown by error bars. See Table S2 in the Supporting Information for statistical tests on the differences in the mean values of NMSEs or CCs among the SVR models.

hereafter). The diAA descriptor has 400 elements. We downloaded amino acid sequences of GPCRs and PKs from GPCRSARfari and KinaseSARfari ftp sites.[35,36] These amino acid sequences focus on the transmembrane regions of GPCRs and the ATP-binding domains of PKs, respectively; other regions and domains were removed beforehand. Amino acid sequences of ICs were retrieved from ChEMBL. We calculated diAA descriptors based on these amino acid sequences. If a target IC protein in an instance was a protein complex or protein family, that instance was discarded, because we could not identify which protein in a complex or family was the true target of the compound. We represented compound−protein pairs in the bioactivity data by concatenating the compound and protein descriptors. Three types of descriptor concatenation (MACCS+diAA, MOE2D+diAA, and OBFP2+diAA) were generated, each having 566, 586, and 1424 elements, respectively.

SVR Implementation. We used the program package Libsvm[37] (ver. 3.1) and its graphics processing unit-accelerated package[38] (ver. 1.1). All elements of the descriptors in the training and validation data were scaled in the range 0 to 1 for every element. Radial basis function kernel was used. Parameters (C, γ, and p) for SVR were optimized by the python script "gridregression.py"[39] in the Libsvm package, then regression models were constructed with the optimized parameters.

To reduce computation time for constructing SVR models, we created 1000 random training datasets. Each dataset is composed of 3000 instances randomly chosen from the original training data. Using random training data sets instead of the original training data alone enabled us to not only reduce computation time, but also to conduct statistical tests of the constructed SVR models. The SVR models were evaluated using normalized mean square error (NMSE)[7] and Pearson's correlation coefficient (CC) calculated between the observed and predicted values. A lower NMSE and higher CC for a SVR model indicates that the SVR model has better predictive performance than other models. We calculated mean values and

dx.doi.org/10.1021/ci5003262 | J. Chem. Inf. Model. 2014, 54, 2751−2763

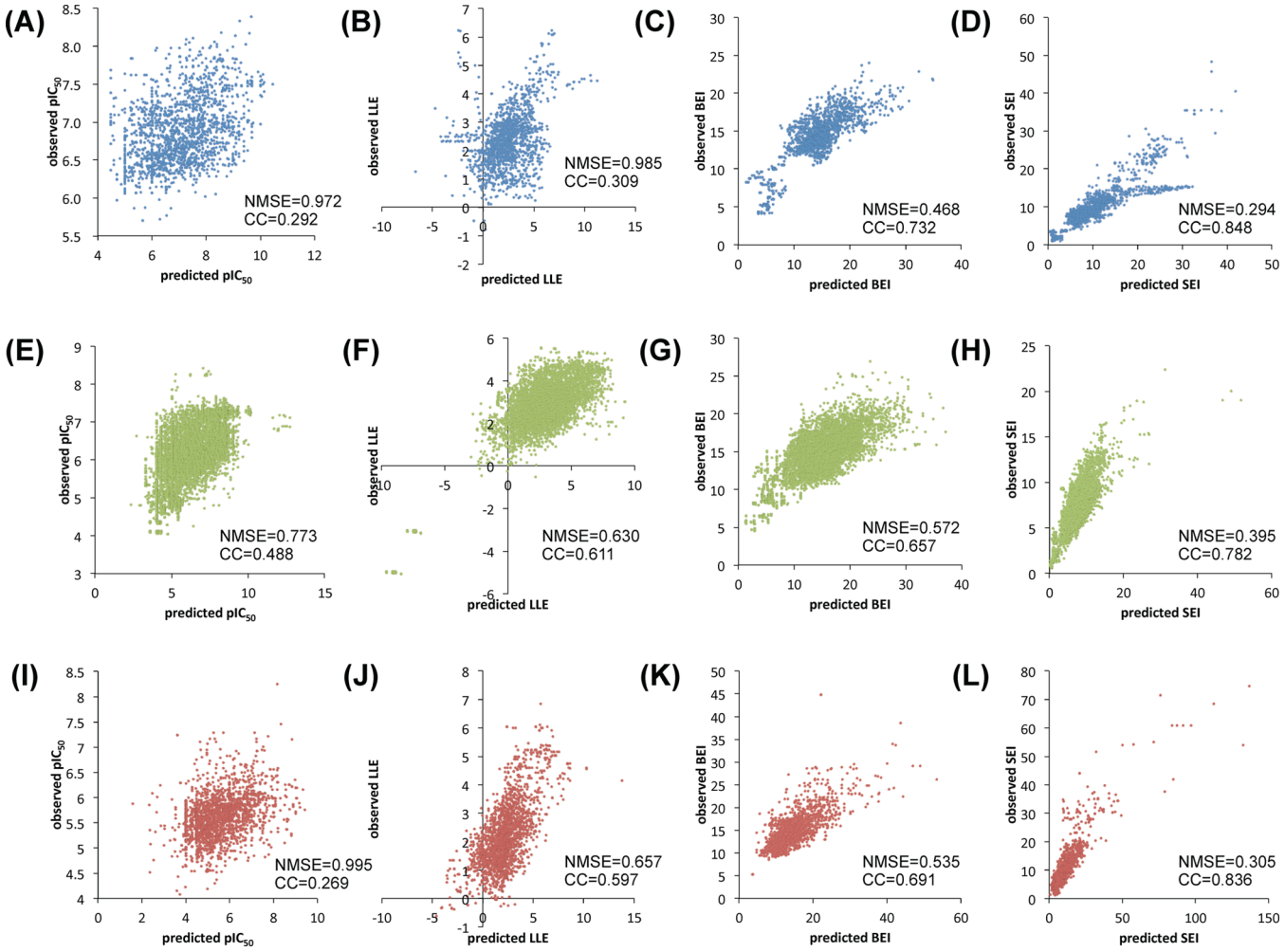


Figure 4. Plots of the observed values versus predicted values obtained from the application tests to the SVR models using the MACCS+diAA descriptor: (A) pIC_{50} model for GPCR, (B) LLE model for GPCR, (C) BEI model for GPCR, (D) SEI model for GPCR, (E) pIC_{50} model for PK, (F) LLE model for PK, (G) BEI model for PK, (H) SEI model for PK, (I) pIC_{50} model for IC, (J) LLE model for IC, (K) BEI model for IC, and (L) SEI model for IC. Horizontal axes indicate predicted values, and vertical axes indicate observed values. NMSEs and CCs are shown in the plots. See Figure S1 in the Supporting Information for the plots when other descriptors are used.

standard deviations (SDs) of NMSEs and CCs over 1000 random training datasets.

RESULTS

Cross-Validation Tests. Using the four types of training datasets based on pIC_{50}, LLE, BEI, or SEI (Figure 1), we conducted 10-fold cross-validation tests. Figure 2 shows that, for all target protein families and all descriptors, the LE-based SVR models (LLE, BEI, and SEI models) perform better than the pIC_{50}-based SVR models. The differences in the mean values of NMSEs (or CCs) between the pIC_{50} models and other LE-based models are statistically significant with P values of <0.01 for all target protein families and all descriptors (see Table S1 in the Supporting Information). Mean values of NMSEs (or CCs) from the LLE, BEI, and SEI models are considerably smaller (or larger) than those from the pIC_{50} models. For example, NMSEs and CCs of the pIC_{50} models for GPCR are both ~0.60 (Figures 2A and 2B). In contrast, all LLE, BEI, and SEI models for GPCR have NMSEs of <0.40 and CCs of >0.75. Of the LLE, BEI, and SEI models, the BEI and SEI models seem to be slightly better than the LLE models for GPCR and PK. On the other hand, all LE-based models have almost equal performance for IC.

Application of the SVR Models to New Data. To check the performance of the constructed SVR models when applied to new data (called "validation data" in this study), we conducted application tests of the SVR models. Figure 3 and Table S2 in the Supporting Information report that, as shown in the cross-validation tests, all LE-based SVR models outperform the pIC_{50} models for all target protein families and all descriptors. The differences in the mean values of NMSEs (or CCs) between the pIC_{50} models and the LE-based models (in particular, the BEI and SEI models) are more remarkable, when compared with the differences in the cross-validation tests. For example, the pIC_{50} models for GPCR show high NMSEs near 1.0 or very low CCs of <0.40 at best. In contrast, the BEI and SEI models have low NMSEs of <0.60 and high CCs of >0.65. Interestingly, among the three types of LE-based models, the BEI models are better than the LLE models and the SEI models are better than the BEI models (see Figure 3 and Table S2 in the Supporting Information) when SVR models are constructed using the MACCS+diAA and OBFP2+diAA descriptors. As a consequence, the predictive

dx.doi.org/10.1021/ci5003262 | J. Chem. Inf. Model. 2014, 54, 2751−2763

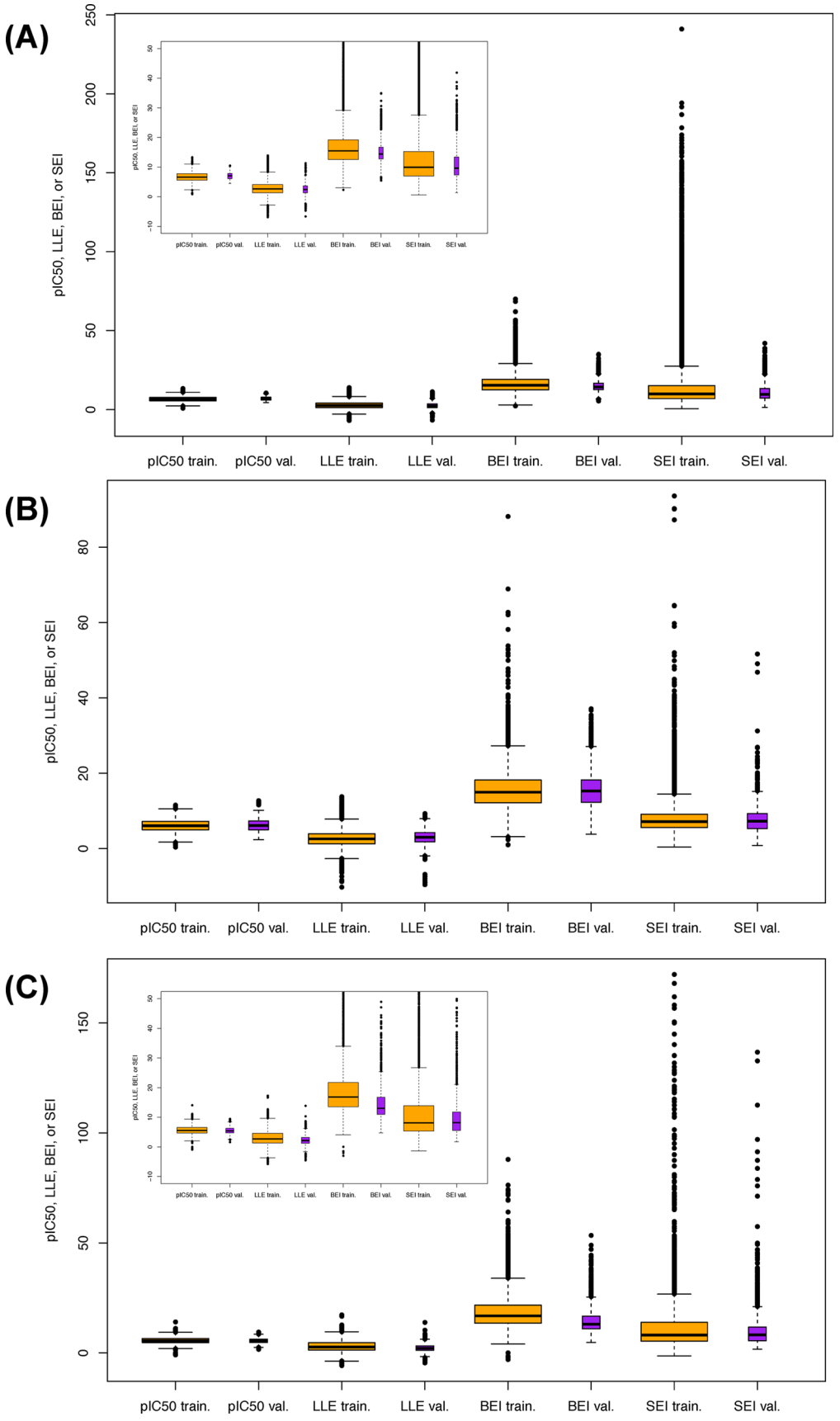


Figure 5. Box plots of the training and validation data for (A) GPCR, (B) PK, and (C) IC. In each plot, box widths are approximately proportional to the number of data points in each dataset. The term "train." approximates "training data" (training data are colored orange), whereas the term "val." approximates "validation data" (validation data are colored purple). To enhance the visual resolution, an expanded version of the plot is shown in the upper-left section of plots (A) and (C).

dx.doi.org/10.1021/ci5003262 | J. Chem. Inf. Model. 2014, 54, 2751−2763

performance of SVR models appears to follow the order SEI > BEI > LLE > pIC_{50}. This performance order also seems to hold, in part, in the SVR models using the MOE2D+diAA descriptor. This result can be visualized by plots of the observed values versus predicted values. Figure 4, as well as Figure S1 in the Supporting Information, clearly display that the observed and predicted values are most highly correlated in the SEI models, followed by the BEI, LLE, and pIC_{50} models, for all target protein families.

Distributions of the Activity Values in the Training and Validation Data. To further investigate why this performance order was observed among the SVR models, we examined the distributions of the activity values in the training and validation data. As shown in Figure 1, the training data (or validation data) differ from each other only in activity values. Thus, the distinct performance of SVR models constructed from the training data can likely be explained simply by the difference in the distributions of the activity values in the training and validation data.

Box plots in Figure 5 report that, for all target protein families, LEs (in particular, BEI and SEI) in the training data are more widely distributed than pIC_{50}s, as shown by wider interquartile regions (box height in the *y*-direction) and wider regions between whiskers. In addition, the distributions of the BEIs and SEIs are characterized by many outliers. The distributions of the pIC_{50}-based training data provide <1% outliers, while they are >4% in the distributions of the LE-based training data. These observations also hold for the distributions of the validation data as a whole. The distinct distribution patterns of the activity values in the training and validation data may, in part, explain the performance order of the SVR models (see Discussion).

Features of the Better Predicted Instances. The results of the application tests indicate that the LE models have better performance than the pIC_{50} models, and the BEI and SEI models are better than the LLE models of the three types of LE models. Therefore, one can ask several questions about the distinctive predictive performance of SVR models.

- Are instances predicted better by each SVR model similar or dissimilar?
- If dissimilar, what types of compound−protein pairs are each SVR model better at predicting the bioactivities of?

To further examine the distinct performance of each SVR model in the application tests, we investigated features of compound−protein pairs (in the validation data) that were better predicted by each SVR model. In this study, better predicted instances (BPIs) are defined as instances (compound−protein pairs) in which the observed activity value falls within a region of predicted value ±0.25SD; i.e., BPIs are data points located near the regression line in observed-versus-predicted plots such as Figure 4. All BPIs can be obtained from the Supporting Information.

Table 2 shows the numbers of BPIs from each SVR model. The percentages of BPIs are limited to 13% at most, and, as expected by the results from the application tests, the BEI and SEI models have more BPIs than the pIC_{50} and LLE models, because of the high correlation of the observed and predicted values. The number of BPIs overlapping between any two SVR models is small (see Figure S2 in the Supporting Information). For GPCR, the percentages of overlapping BPIs between any two models are in the range of 4.8%−25.2%. For PK and IC, the percentages are 18.8% or less and 16.4% or less, respectively. Either no or only one BPI is shared among all SVR models when the MACCS+diAA or OBFP2+diAA descriptor is used. Although the number of BPIs that are common to all SVR models increases when the MOE2D+diAA descriptor is used, it is still limited to 3 for GPCR, 18 for PK, and 6 for IC (see Figure S2 in the Supporting Information). These results imply that the SVR models constructed here could display better performances for compound−protein pairs that are partially distinct from each other. We then investigated compound−protein pairs in the BPIs by focusing on the observed ligand potency and physicochemical properties (*A* log *P*, MW, and TPSA closely associated with the SVR models in this study) of the compounds and the functional classification of the target proteins.

Table 2. Number of the BPIs from Each SVR Model

		Number of BPIs[a]		
target protein family	SVR model	MACCS +diAA	MOE2D +diAA	OBFP2+diAA
GPCR	pIC_{50} model	86 (5.5%)	90 (5.8%)	92 (5.9%)
GPCR	LLE model	131 (8.4%)	168 (10.8%)	92 (5.9%)
GPCR	BEI model	178 (11.5%)	206 (13.3%)	111 (7.1%)
GPCR	SEI model	189 (12.2%)	145 (9.3%)	138 (8.9%)
PK	pIC_{50} model	351 (5.6%)	386 (6.2%)	323 (5.2%)
PK	LLE model	441 (7.0%)	470 (7.5%)	433 (6.9%)
PK	BEI model	449 (7.2%)	498 (7.9%)	453 (7.2%)
PK	SEI model	591 (9.4%)	744 (11.9%)	665 (10.6%)
IC	pIC_{50} model	134 (6.7%)	149 (7.5%)	94 (4.7%)
IC	LLE model	157 (7.9%)	221 (11.1%)	167 (8.4%)
IC	BEI model	167 (8.4%)	215 (10.8%)	112 (5.6%)
IC	SEI model	218 (10.9%)	283 (14.2%)	157 (7.9%)

[a]Percentages of the number of BPIs, relative to that of all instances in the validation data, are shown in parentheses.

Observed Ligand Potency. For each BPI, the observed activity value in the validation data was retrieved, then the observed LLEs, BEIs, and SEIs were reconverted to pIC_{50}s. Figure 6A shows the distributions of the observed pIC_{50}s of the compounds in the BPIs from the pIC_{50}, LLE, BEI, and SEI SVR models using the MACCS+diAA descriptor for GPCR. The validation data are also plotted (see Figure S3 in the Supporting Information for other descriptors). This plot indicates that the observed pIC_{50}s of the compounds in the BPIs from the LE models are more widely distributed than those of the compounds in the BPIs from the pIC_{50} model. This is also observed in other target proteins (see Figures 7A and 8A) and the SVR models using other descriptors (see Figure S3 in the Supporting Information). This result implies that the LE models can offer better performance both for groups of compound−protein pairs showing higher or lower ligand potency, as well as for pairs showing midrange potencies that can be predicted by the pIC_{50} models. For example, in Figures 6A and 7A, all LE models can successfully predict the bioactivities of compound−protein pairs with an observed

dx.doi.org/10.1021/ci5003262 | J. Chem. Inf. Model. 2014, 54, 2751−2763

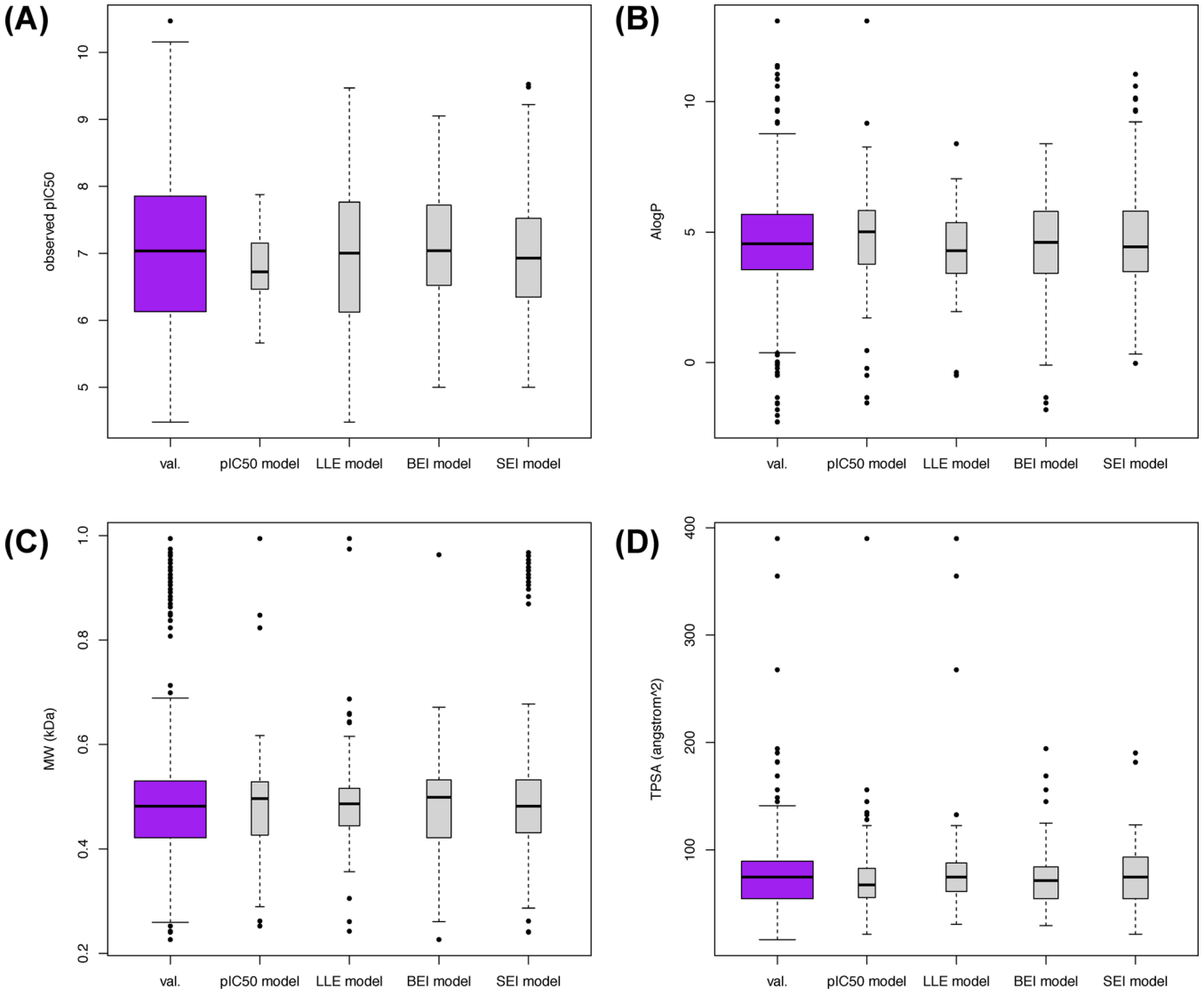


Figure 6. Box plots of the (A) observed pIC_{50}, (B) *A* log *P*, (C) MW, and (D) TPSA values of the compounds in the BPIs from the SVR models using the MACCS+diAA descriptor for GPCR. In each plot, box widths are approximately proportional to the number of data points in each dataset. The term "val." approximates "validation data" (validation data are colored purple). See Figure S3 in the Supporting Information for other descriptors.

pIC_{50} of >8, while the pIC_{50} models fail to predict the bioactivities of pairs with that potency. Furthermore, compound−protein pairs with an observed pIC_{50} of <4 can be successfully predicted by the LE models, as shown in Figures 7A and 8A, but the pIC_{50} models again fail to predict the bioactivities.

AlogP, MW, and TPSA. Unlike the observed ligand potency, the distributions of the *A* log *P*, MW, and TPSA values of the compounds in the BPIs from the LE-based SVR models showed no characteristic trend by which the LE-based models were clearly distinguished from the pIC_{50} models (recall Figures 6−8, and Figure S3 in the Supporting Information). Although some plots in Figures 6−8 show a distribution pattern from a SVR model that is distinct from other distributions, the pattern is not consistent over all target protein families. For example, the BEI model in Figure 6C can better predict the bioactivities of compound−protein pairs with a wider range of compound MWs than other SVR models, but this does not hold for other target proteins (Figure 7C and 8C). These results imply that the SVR models constructed in this study would be insensitive to the *A* log *P*, MW, and TPSA values of compounds in predicting the bioactivities of compound−protein pairs.

Functional Classification of the Target Proteins. We investigated the distributions of the target proteins in the BPIs, with respect to their functional classification. GPCRSARfari, KinaseSARfari, and ChEMBL have a hierarchical classification system of target proteins. We adopted hierarchical levels in the classification systems in which target proteins could be classified to an appropriate number of subfamilies (not too small or large) for the following statistical analyses. Then, we classified the target proteins in the BPIs and validation data according to the adopted hierarchical level of the classification systems. Hierarchical levels of 3, 2, and 6 were adopted for GPCR, PK, and IC, respectively.

Figure 9 shows that, both in the BPIs and validation data, there are considerable biases, with respect to the percentages of the target proteins that belong to each subfamily. The most abundant subfamily in GPCR is "Peptide|Short Peptide" followed by "SmallMol|Monoamine receptor". "TK" is the

dx.doi.org/10.1021/ci5003262 | J. Chem. Inf. Model. 2014, 54, 2751−2763

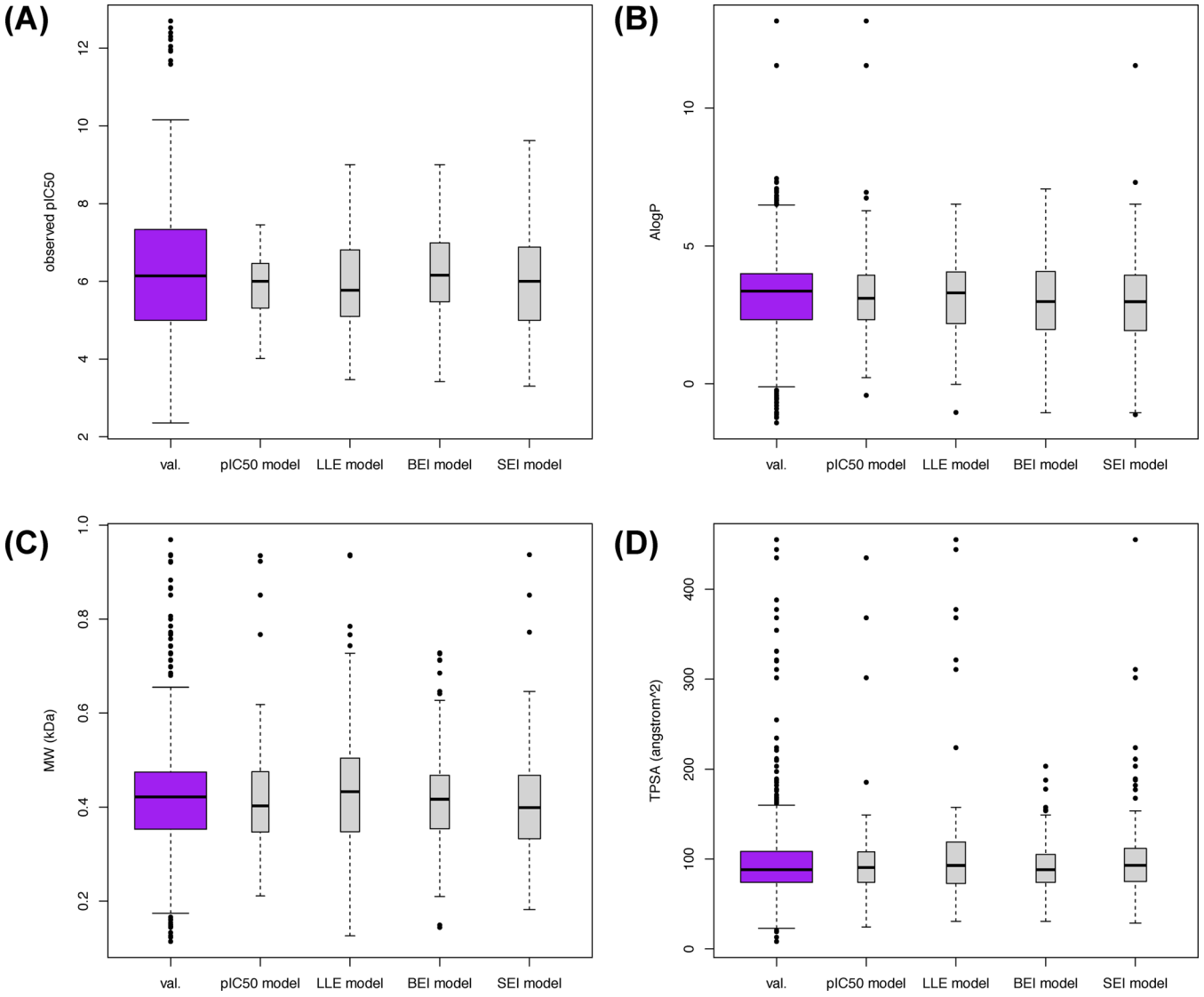


Figure 7. Box plots of the (A) observed pIC_{50}, (B) *A* log *P*, (C) MW, and (D) TPSA values of the compounds in the BPIs from the SVR models using the MACCS+diAA descriptor for PK. In each plot, box widths are approximately proportional to the number of data points in each dataset. The term "val." approximates "validation data" (validation data are colored purple). See Figure S3 in the Supporting Information for other descriptors.

most abundant subfamily in PK, followed by "CMGC" and "AGC". Most compound−protein pairs in the BPIs and validation data of IC are associated with human ether-à-go-go-related gene, or hERG, belonging to the "VGC|VGC|VOLT|CATIONIC|K" subfamily. Chi-square (χ^2) tests for independence between the validation data and BPIs showed that the BPIs from several SVR models have a distribution of the target proteins that differ from the validation data with a statistical significance of $P < 0.05$ (see Table S3 in the Supporting Information). To investigate which subfamily has the number of proteins that is distinct from an expected value, we calculated the adjusted residual[40] of each subfamily when a *P* value via the χ^2 test was <0.05. A large adjusted residual (>2 or <−2) of a category (subfamily) indicates lack of fit of the null hypothesis (BPIs and validation data are derived from the same distribution) in that category.[40] Table S4 in the Supporting Information reports that some subfamilies ("SmallMol|Nucleotide-like receptor" in GPCR, "AGC" in PK, and "LGIC|CYS_LOOP|GLY|ANIONIC|CL" in IC) show adjusted residuals of >2 or <−2 in several SVR models. On the other hand, several other subfamilies (for example, "Peptide|Chemokine receptor" in GPCR, "CK1" in PK, and "LGIC|CYS_LOOP|5HT3|CATIONIC|NS" in IC) have small adjusted residuals of <2 or >−2 in all SVR models. In all target protein families, however, it is not observed that any types of SVR model or any descriptors consistently show a large positive or negative adjusted residual for a specific subfamily. This implies that the SVR models constructed in this study have no better (or worse) predictive performance for a specific subfamily, although, in some cases, it is likely that each SVR model shows better (or worse) performance for a subfamily by chance.

DISCUSSION

In this study, we constructed SVR models based on LE indices (LLE, BEI, and SEI) as well as pIC_{50} values traditionally used in chemogenomics data-based machine learning studies. The LE-based SVR models can offer better performance than the pIC_{50}

dx.doi.org/10.1021/ci5003262 | J. Chem. Inf. Model. 2014, 54, 2751−2763

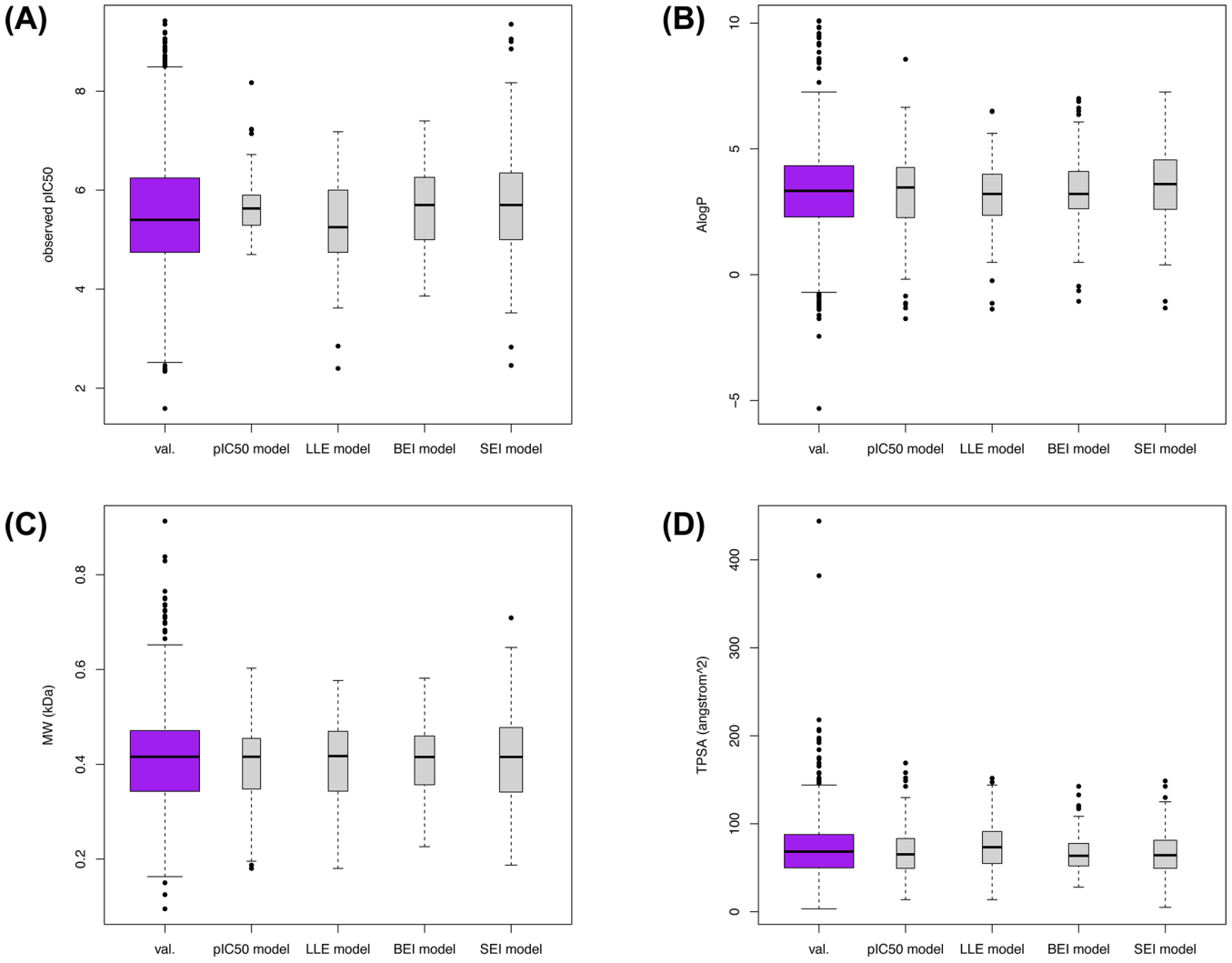


Figure 8. Box plots of the (A) observed pIC_{50}, (B) *A* log *P*, (C) MW, and (D) TPSA values of the compounds in the BPIs from the SVR models using the MACCS+diAA descriptor for IC. In each plot, box widths are approximately proportional to the number of data points in each dataset. The term "val." approximates "validation data" (validation data are colored purple). See Figure S3 in the Supporting Information for other descriptors.

models for all target protein families and all descriptors used. NMSEs of ∼0.6 and CCs of ∼0.6 yielded by the pIC_{50}-based SVR models for GPCR in the cross-validation tests (Figures 2A and 2B) are comparable with values obtained in a pioneering study by Bock and Gough.[7] In contrast, the LE-based SVR models can yield much lower NMSEs (<0.5) and higher CCs (>0.75) in the cross-validation tests. These results support our earlier findings that BEI-based SVM classification models have better performance than IC_{50} (or K_i)-based classification models.[24] In our previous study, we used BEI as a representative measurement of LE; others, such as LLE and SEI, had not previously been investigated in this context. Thus, it was not clear from our previous study whether SVM classification models (or SVR models) based on LLE and SEI can also outperform IC_{50} (or K_i)-based classification/regression models. The results of the cross-validation tests and application tests, shown in Figures 2 and 3, clearly demonstrate that the LLE- and SEI-based SVR models, as well as the BEI-based models, perform better than the pIC_{50} models. Therefore, the superiority of LE indices to ligand potencies (IC_{50} and K_i) when used in machine learning holds true for all LLE, BEI, and SEI.

The performance of SVR models follows the order SEI > BEI > LLE > pIC_{50} in the application tests. Our previous study attributed the better performance of BEI-based SVM classification models, compared to IC_{50} (or K_i)-based models to the observation that positives and negatives in the BEI-based training data are more separately distributed than those in the IC_{50} (or K_i-based training data in the SVM hyperspace.[24] In this study, because the feature vectors of the compound−protein pairs are all identical among the four types of training data and among validation data (Figure 1), the observed performance order can likely be explained based simply on the distributions of the activity values in the training and validation data (Figure 5). The distributions of the activity values showed that the LEs in the training and validation data are more widely distributed than the pIC_{50}s. This is particularly remarkable in the BEIs and SEIs. This result indicates that data points (activity values) in the LE-based training and validation data have a higher diversity than data points in the pIC_{50}-based data. Generally speaking, in order to create machine learning-based regression models that have as high a predictive performance as possible, training data used to create regression models would have to be composed of data points that are as diverse as possible. The higher diversity

dx.doi.org/10.1021/ci5003262 | J. Chem. Inf. Model. 2014, 54, 2751−2763

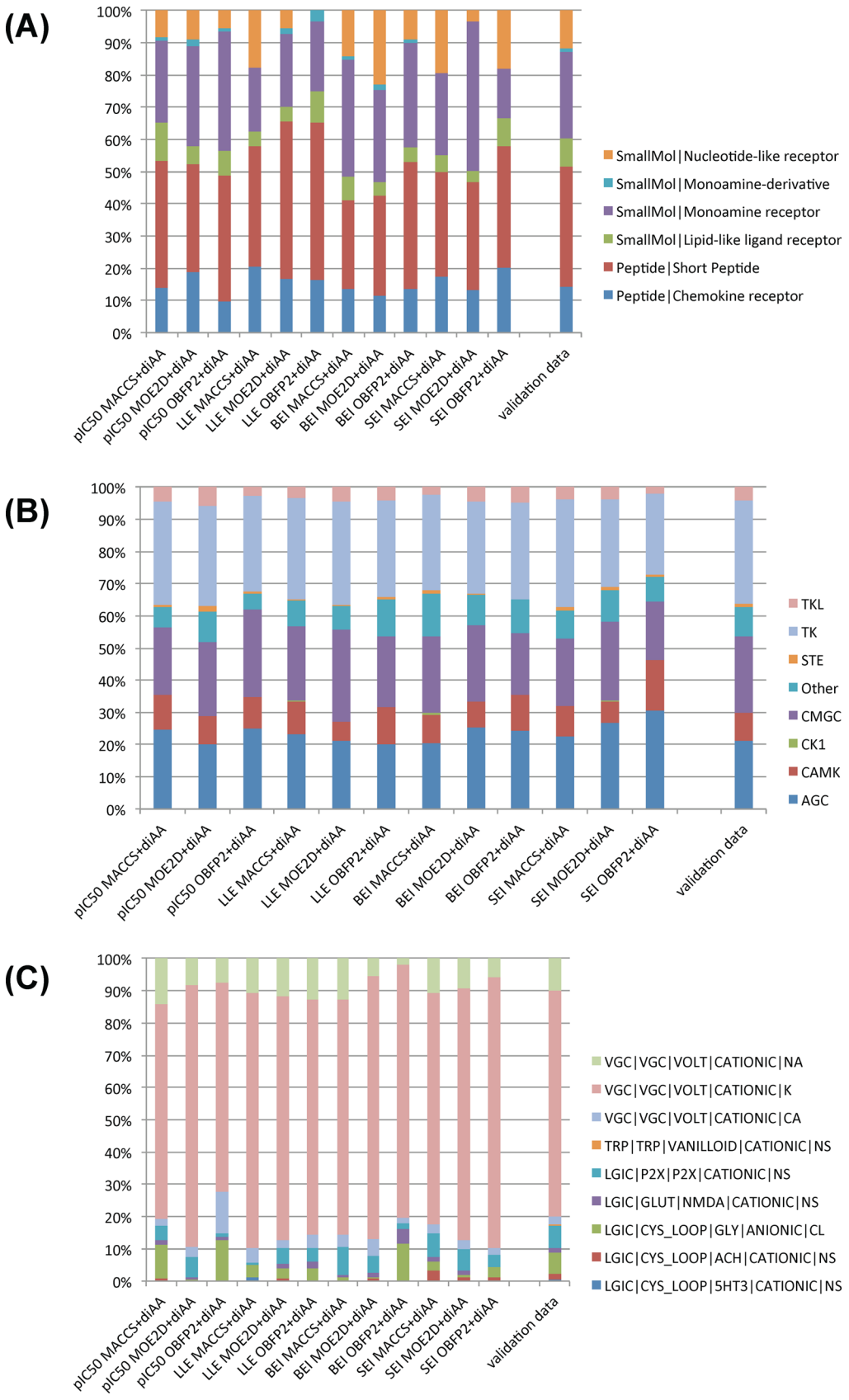


Figure 9. Functional classification of the target proteins in the BPIs and validation data for (A) GPCR, (B) PK, and (C) IC. In the notes, vertical bars (shown as "|") designate partitions of hierarchical levels in the classification systems of target proteins in GPCRSARfari, KinaseSARfari, or ChEMBL. For example, in "LGIC|CYS_LOOP|5HT3|CATIONIC|NS", LGIC is level 2, CYS_LOOP is level 3, 5HT3 is level 4, and so on.

dx.doi.org/10.1021/ci5003262 | J. Chem. Inf. Model. 2014, 54, 2751−2763

of the LEs, especially of the BEIs and SEIs, in the training data may explain the higher performance of the SVR models created from the LE-based training data.

Another explanation takes into consideration the features of activity values and descriptors and the relationships among them. While pIC_{50} evaluates the bioactivity of a ligand, with respect to the entire structure of the compound, BEI and SEI evaluate the bioactivity per unit size by definition and, thus, they necessarily focus on partial structures of the ligand. Two out of three compound descriptors used here, MACCS and OBFP2, are created based on partial structures of compound. Concentrating on partial structures of compound is the feature that is common to activity values of BEI and SEI and compound descriptors of MACCS and OBFP2. This may lead to good affinities between BEI/SEI and MACCS/OBFP2 and result in an improvement of the predictive performance of SVR models when activity values are converted from pIC_{50} to BEI and SEI. However, BEI and SEI models using the MOE2D descriptor also showed improvement over pIC_{50} models. MOE2D is calculated based on the entire 2D structure, rather than partial structures, of the compound.[31] This result indicates that the higher performance of BEI and SEI models is independent of the features of compound descriptors used in this study and implies that there may be no or very low affinities among BEI or SEI and MACCS or OBFP2. Furthermore, LLE-based SVR models, as well as BEI and SEI models, also showed slightly better performance than pIC_{50} models, irrespective of the compound descriptor used. By definition, LLE relates to the overall lipophilicity of a ligand and does not evaluate bioactivity of the ligand per unit size. Although the affinities between activity values and descriptors may be possibly associated with the superiority of LE-based SVR models, further studies will be needed to determine whether the features of descriptors are responsible for the superiority of LE-based SVR models.

The small number of overlapping BPIs among the SVR models suggests that each SVR model might display its predictive power for different groups of compound−protein pairs. For example, one might expect that BEI models are suitable for predicting bioactivities of compounds with a large (or small) MW. Close examination of the distributions of the observed ligand potencies and physicochemical properties (*A* log *P*, MW, and TPSA) of compounds in the BPIs from the SVR models demonstrates that the LE-based SVR models have better predictive power both for groups of compound−protein pairs with higher or lower ligand potencies, as well as for pairs with midrange potencies that can be predicted by the pIC_{50} models (Figures 6A, 7A, and 8A). However, all types of SVR models, are insensitive to the *A* log *P*, MW, and TPSA values of the compounds. As for the functional classification of the target proteins, the SVR models display no tendency toward a better performance for a specified target protein subfamily. Although we defined BPIs as instances in which an observed activity value fell within a region of predicted value of ±0.25SD, another definition can also be adopted. We defined BPIs based on other thresholds, such as "predicted value ±0.1SD" and "predicted value ±0.5SD", and reinvestigated the features of the BPIs. The number of BPIs decreased when the "predicted value ±0.1SD" threshold was used and increased when "predicted value ±0.5SD" was used, but the results obtained were not essentially different from the results for the BPIs based on the original threshold "predicted value ±0.25SD" (see Tables S5−S14 in the Supporting Information). Thus, the results here for the BPIs are robust to changes in definition. These observations suggest that compounds with much higher (or lower) potency can be more reliably predicted using LE models rather than pIC_{50} models.

Although the findings here hold true for all target protein families (GPCR, PK, and IC) and descriptors (MACCS+diAA, MOE2D+diAA, and OBFP2+diAA) tested, some issues remain to be addressed. For example, are the findings robust to bioactivity data from other drug target proteins such as proteases and nuclear receptors, and to other descriptors used as feature vector? Moreover, it should be noted that the results in this study may be dependent on the machine learning method SVR or bioactivity data in GPCRSARfari, KinaseSARfari, and ChEMBL. These issues will be addressed in future studies.

CONCLUSIONS

Using GPCRSARfari, KinaseSARfari, and ChEMBL, we created the pIC_{50}-, LLE-, BEI-, and SEI-based SVR models in order to predict ligand bioactivities to target proteins and compare the performance of the SVR models. The cross-validation tests reported that the predictive performance of the LLE-, BEI-, and SEI-based models were better than the pIC_{50}-based models. Moreover, the SEI models are better than the LLE and BEI models, and the BEI models are better than the LLE models, in the application tests. Overall, the performance order is SEI > BEI > LLE > pIC_{50}. This performance order can be partially explained by the greater diversity in activity values in the LE-based training and validation data. Close examination of the BPIs from the viewpoint of observed ligand potency and physicochemical properties of the compounds, and the functional classification of the target proteins, indicated that the LLE, BEI, and SEI models have better performance than the pIC_{50} models at predicting bioactivities of compounds with a wide range of potencies. However, all SVR models seem to be insensitive to the *A* log *P*, MW, and TPSA values of the compounds and to the functional classification of the target proteins. These findings strongly suggest that LE-based SVR models are better than pIC_{50} models at predicting bioactivities of compounds with much higher (or lower) potencies.

ASSOCIATED CONTENT

Supporting Information

Figure S1 shows the plots of the observed activity value versus predicted value from the application tests. Figure S2 shows the overlaps of BPIs among the four types of SVR models. Figure S3 is the box plots of the observed pIC_{50}, *A* log *P*, MW, and TPSA values of the compounds in the BPIs in the application tests. Table S1 shows the results of statistical tests on the difference in the mean values of NMSEs or CCs in the cross-validation tests. Table S2 shows the results of statistical tests on the difference in the mean values of NMSEs or CCs in the application tests. Table S3 provides the results of statistical tests on the distributions of the functional classification of the target proteins between the BPIs and validation data. Table S4 indicates the adjusted residuals of the target protein subfamilies in χ^2 tests between the BPIs and validation data. Tables S5−S14 report the results of investigations into the BPIs, based on other thresholds. Training data, validation data, and BPIs used in this study are provided in ZIP-format files. 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) Hopkins, A. L.; Keserü, G. M.; Leeson, P. D.; Rees, D. C.; Reynolds, C. H. The role of ligand efficiency metrics in drug discovery. *Nat. Rev. Drug Discovery* **2014**, *13*, 105−121.

(2) Perola, E. An analysis of the binding efficiencies of drugs and their leads in successful drug discovery programs. *J. Med. Chem.* **2010**, *53*, 2986−2997.

(3) Katritch, V.; Jaakola, V. P.; Lane, J. R.; Lin, J.; Ijzerman, A. P.; Yeager, M.; Kufareva, I.; Stevens, R. C.; Abagyan, R. Structure-based discovery of novel chemotypes for adenosine A(2A) receptor antagonists. *J. Med. Chem.* **2010**, *53*, 1799−1809.

(4) Tanaka, D.; Tsuda, Y.; Shiyama, T.; Nishimura, T.; Chiyo, N.; Tominaga, Y.; Sawada, N.; Mimoto, T.; Kusunose, N. A practical use of ligand efficiency indices out of the fragment-based approach: ligand efficiency-guided lead identification of soluble epoxide hydrolase inhibitors. *J. Med. Chem.* **2011**, *54*, 851−857.

(5) Dunkern, T.; Prabhu, A.; Kharkar, P. S.; Goebel, H.; Rolser, E.; Burckhard-Boer, W.; Arumugam, P.; Makhija, M. T. Virtual and experimental high-throughput screening (HTS) in search of novel inosine 5′-monophosphate dehydrogenase II (IMPDH II) inhibitors. *J. Comput. Aided Mol. Des.* **2012**, *26*, 1277−1292.

(6) Vass, M.; Schmidt, E.; Horti, F.; Keserû, G. M. Virtual fragment screening on GPCRs: a case study on dopamine D3 and histamine H4 receptors. *Eur. J. Med. Chem.* **2014**, *77*, 38−46.

(7) Bock, J. R.; Gough, D. A. Virtual screen for ligands of orphan G protein-coupled receptors. *J. Chem. Inf. Model.* **2005**, *45*, 1402−1414.

(8) 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.

(9) Jacob, L.; Hoffmann, B.; Stoven, V.; Vert, J. P. Virtual screening of GPCRs: An *in silico* chemogenomics approach. *BMC Bioinformatics* **2008**, *9*, 363.

(10) Jacob, L.; Vert, J. P. Protein-ligand interaction prediction: an improved chemogenomics approach. *Bioinformatics* **2008**, *24*, 2149−2156.

(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) 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.

(14) Buchwald, F.; Richter, L.; Kramer, S. Predicting a small molecule−kinase interaction map: A machine learning approach. *J. Cheminform.* **2011**, *3*, 22.

(15) 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.

(16) 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.

(17) 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.

(18) 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.

(19) 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.

(20) 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.

(21) 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.

(22) 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.

(23) 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.

(24) Sugaya, N. Training based on ligand efficiency improves prediction of bioactivities of ligands and drug target proteins in a machine learning approach. *J. Chem. Inf. Model.* **2013**, *53*, 2525−2537.

(25) *GPCR SARfari*; https://www.ebi.ac.uk/chembl/sarfari/gpcrsarfari (accessed May 7, 2014).

(26) *Kinase SARfari*; https://www.ebi.ac.uk/chembl/sarfari/kinasesarfari (accessed May 7, 2014).

(27) Leeson, P. D.; Springthorpe, B. The influence of drug-like concepts on decision-making in medicinal chemistry. *Nat. Rev. Drug Discovery* **2007**, *6*, 881−890.

(28) Abad-Zapatero, C.; Metz, J. T. Ligand efficiency indices as guideposts for drug discovery. *Drug Discovery Today* **2005**, *10*, 464−469.

(29) Abad-Zapatero, C.; Perišiæ, 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 Discovery Today* **2010**, *15*, 804−811.

(30) Ghose, A. K.; Crippen, G. M. Atomic physicochemical parameters for three-dimensional-structure-directed quantitative structure-activity relationships. 2. Modeling dispersive and hydrophobic interactions. *J. Chem. Inf. Comput. Sci.* **1987**, *27*, 21−35.

(31) *MOE: Molecular Operating Environment*; http://www.chemcomp.com/MOE-Molecular_Operating_Environment.htm (accessed May 13, 2014).

(32) *MACCS Structural Keys*; Accelrys: San Diego, CA.

(33) *Open Babel*; http://openbabel.org/wiki/Main_Page (accessed May 7, 2014).

(34) *FP2—Open Babel*; http://openbabel.org/wiki/FP2 (accessed April 8, 2014).

(35) *GPCRSARfari ftp site*; ftp://ftp.ebi.ac.uk/pub/databases/chembl/GPCRSARfari/releases/ (accessed May 13, 2014).

(36) *KinaseSARfari ftp site*; ftp://ftp.ebi.ac.uk/pub/databases/chembl/KinaseSARfari/releases/ (accessed May 13, 2014).

(37) *LIBSVM—A Library for Support Vector Machines*; http://www.csie.ntu.edu.tw/∼cjlin/libsvm/ (accessed May 13, 2014).

(38) *GPU-accelerated LIBSVM*; http://mklab.iti.gr/project/GPU-LIBSVM (accessed May 13, 2014).

(39) *LIBSVM tools*; http://www.csie.ntu.edu.tw/∼cjlin/libsvmtools/#grid_parameter_search_for_regression (accessed April 8, 2014).

(40) Agresti, A. In *An Introduction to Categorical Data Analysis*, Second Edition; John Wiley & Sons: Hoboken, NJ, 2007; Chapter 2, pp 38−39.