ABSTRACT: In ligand-based drug design, quantitative structure–activity relationship (QSAR) models play an important role in activity prediction. One of the major end points of QSAR models is half-maximal inhibitory concentration (IC$_{50}$). Experimental IC$_{50}$ data from various research groups have been accumulated in publicly accessible databases, providing an opportunity for us to use such data in predictive QSAR models. In this study, we focused on using a ranking-oriented QSAR model as a predictive model because relative potency strength within the same assay is solid information that is not based on any mechanical assumptions. We conducted rigorous validation using the ChEMBL database and previously reported data sets. Ranking support vector machine (ranking-SVM) models trained on compounds from similar assays were as good as support vector regression (SVR) with the Tanimoto kernel trained on compounds from all the assays. As effective ways of data integration, for ranking-SVM, integrated compounds should be selected from only similar assays in terms of compounds. For SVR with the Tanimoto kernel, entire compounds from different assays can be incorporated.

INTRODUCTION

In ligand-based drug design, quantitative structure–activity relationship (QSAR) models play an important role in predicting compounds' biological activity and/or toxicity.\textsuperscript{1,2} Thanks to accumulated experimental data registered in publicly available databases and integration of these databases,\textsuperscript{3–7} collecting heterogeneous data for building QSAR models has become easy, and large-scale analyses have become popular. These large-scale analyses are indispensable for new development of QSAR modeling strategies.\textsuperscript{8–10} As objective end points of QSAR models, the half-maximal inhibitory concentration (IC$_{50}$), the inhibition constant (K$_i$), the half-maximum effective concentration (EC$_{50}$), and the dissociation constant (K$_d$) are commonly employed.\textsuperscript{11–13} However, IC$_{50}$ and EC$_{50}$ are theoretically assay-specific measurements, and so in contrast to equilibrium constants, these values are comparable only when the data are from compounds in a single assay. Nevertheless, several studies have used mixed IC$_{50}$ data for QSAR model construction because of the limited numbers of data points from single assays, and such QSAR models have sometimes shown satisfactory predictive ability.\textsuperscript{6,14,15} These results are in agreement with the finding by Kalliokoski et al., mixing IC$_{50}$ from different assays add a moderate amount of noise to the entire data set.\textsuperscript{16} Noises by data integration partly arise from experimental uncertainty. Using heterogeneous compounds from the ChEMBL database, the uncertainty in K$_i$ was estimated as a mean unsigned error of between 0.44 and 0.48 pK$_i$ units.\textsuperscript{17} Whether mixed IC$_{50}$ data contribute to the performance improvement of QSAR models or not was previously investigated using a case study for HIV-1 reverse transcriptase inhibitors by Tarasova and co-workers.\textsuperscript{18} They used the Integrity\textsuperscript{19} database with an assay description for building global and assay-type-based QSAR models. They observed that the predictive ability of QSAR models using compounds from a specific type of biological assay was better than using mixed IC$_{50}$ data, although the trend was not clear when the ChEMBL database was employed.

The method of using end point data from multiple assays without any mathematical transformation can be regarded as a data integration. Ideally, all IC$_{50}$ values from different assays should be standardized to a comparable form (e.g., K$_i$). However, this requires precise assay information and assumptions about the inhibition mechanism,\textsuperscript{20} which are impossible to access or difficult to obtain for many data sets.

Finding a proper data integration strategy for QSAR modeling is important because it could compensate for the lack of experimental data from single assays. Furthermore, as long as a biological target–inhibitor pair annotated with an
IC50 value is found in publicly available databases, data integration strategies could also provide a way of taking advantage of these accessible databases.

A method for integrating compounds with IC50 values from different assays to predict IC50 values for a specific target assay was previously proposed. In this method, IC50 values from multiple assays are calibrated to the targeted one based on a shared active compound, for which IC50 values were measured in both assays to be integrated (see the scaling method in the Materials and Methods section). Although this method is reasonable, it has the disadvantage of requiring a shared (control) active compound among all assays.

Predicting exact end point values is sometimes too excessive, particularly in virtual screening. Rather, prioritizing compounds to be tested or synthesized next is sometimes sufficient. The concept of learning to rank (LTR) fits this purpose. LTR is a class of algorithms for building ranking models in which test compounds are ranked on the basis of ranking information about training compounds (i.e., the relative strength of their inhibition ability). LTR-based models have been reported to outperform the traditional regression-based models in both assays to be integrated (see the scaling method in the Materials and Methods section). This is partly because model training based on ranking information can eliminate the experimental uncertainty of end point values and compensate for small sample size by transforming activity values into pairwise orderings. Furthermore, LTR provides a natural method to integrate ranking information from different assays.

Herein, we propose prediction strategies to create ranking-oriented QSAR models by integrating data from different assays to improve the models’ predictive ability. Ranking-SVM has been used for this purpose, and various criteria for data integration have been tested to derive the best strategy. The logarithm of inverse IC50 (pIC50) was the tested end point in this study. Our validation scheme using the ChEMBL databases and bootstrap test repetition revealed that ranking-support vector machine (ranking-SVM) with the integration of compounds across similar assays and support vector regression (SVR) with the Tanimoto kernel with the incorporation of compounds across all the assays gave more stable and better performance than other methods.

### MATERIALS AND METHODS

**Assay-Wise Activity Data Sets.** Assay-based (AB) active compounds for various biological targets were extracted from the ChEMBL database (version 24). As an activity measurement, IC50 was measured in nanomolar standard units (nM).

- **MATERIALS AND METHODS**
remaining assays with the corresponding biological target names and the number of compounds are reported in Table 1.

Cyclooxygenase 1 Data Sets. For direct comparison of the ranking-oriented QSAR models for data integration with previously proposed data integration methods published using the same data, human cyclooxygenase 1 (COX 1) data sets were prepared by downloading assay-wise compounds from the ChEMBL database (version 25) based on the compounds’ CHEMBL Assay IDs. Consequently, 12 assays were collected, and the average number of compounds per assay was 29. Only 9 assays are eligible for applying the scaling method. The assays for these COX 1 data sets are also reported in Table 1.

Validation Scheme. A validation scheme was established to evaluate whether data integration of different assays improves the QSAR methods’ performance (Figure 1). For each assay of a biological target, the compounds with IC50 values were equally split into training and test data sets (Assay α in Figure 1). Other assays can also be incorporated into Assay α depending on the integration strategies (see the Data Integration Strategies section) used to form an integrated training set for ranking-oriented QSAR models. Ranking-oriented QSAR models predict the relative rankings of test compounds. Although individual compounds’ IC50 values vary based on assay conditions, the relative potencies of a set of compounds are preserved throughout each assay. Therefore, compound rankings can serve as trustworthy information over different assays. Compounds are ranked in the ascending order of IC50 values (compounds with low IC50 values have high rankings). These ranking-oriented QSAR models predict compounds’ rankings on the basis of their IC50 values. This procedure was repeated 10 times by using a random number generator to change the random seed for splitting of the training/test data sets, unless the number of training and integrated compounds was <5. The training and integrated compounds were excluded from the test set, preventing duplication between the training and test data sets.

Model performance was evaluated using test data sets in terms of the Kendall rank correlation coefficient between the compounds’ rankings based on the measured IC50 values and their predicted rankings:

\[
\eta_k = \frac{G - H}{\frac{n(n - 1)}{2}}
\]

where \(n\) is the number of test compounds and \(G\) and \(H\) are the total number of concordant and discordant pairs, respectively.

Molecular Representations. As a molecular representation, the extended connectivity fingerprints with a bond diameter of 4 (ECFP4)26 were folded into a 2048-bit vector by a modulo operation. The ECFP4 fingerprints were generated using in-house Python scripts with the aid of the OEChem toolkit.

Modeling Methods Combined with Data Integration. There are several LTR model construction methods and approaches. In this study, we focused on ranking-SVM with the linear kernel function as an LTR method. As a regression method, partial least-squares regression (PLS-R) and SVR with the Tanimoto kernel function were employed to create the IC50 prediction models. The target end point was pIC50. These models were trained on (non)integrated data sets based on data integration strategies. Furthermore, for PLS-R and SVR models, the previously proposed “scaling” and “nonscaling” (NSC) potency integration methods were applied. Afterward, the predicted pIC50 values from the regression models were converted to corresponding rankings.

Representations of Assay Information for Regression Modeling. Two types of representations were introduced to the PLS-R and SVR models to describe the assays: one was one-hot encoding and the other was the statistical values of IC50 from the assays. In the latter case, the statistics of within-assay IC50 values (i.e., the minimum, first quartile, median, third quartile, and maximum) were used as descriptors. These descriptors and ECFP4 fingerprints were concatenated to form a set of descriptors including assay information. The statistical values of IC50 from assays were used only for PLS-R models to avoid including additional parameters of kernel functions in SVR.

Ranking-SVM. Ranking-SVM, which is a variant of SVM, is a commonly used LTR modeling method. Ranking-SVM uses a pairwise ranking algorithm. The idea of the pairwise algorithm is to transform the ranking problem into a pairwise classification problem, as follows:

Figure 1. Validation protocols for assay integration in QSAR modeling.
minimize $\frac{1}{2}w^Tw + C\sum \xi_{i,j,k}$

s.t. $\forall i, j, k, \forall \xi_{i,j,k} \geq 0$

$\forall (c_i, c_j) \in R_k$

$w^T(\Phi(a_k, c_i) - \Phi(a_k, c_j)) \geq 1 - \xi_{i,j,k}$

Suppose that $c_i$ and $c_j$ are two compounds. If the rank of $c_j$ is higher than that of $c_i$ in the assay $k$ using a certain ranking metric $r$, we can denote $(c_i, c_j) \in R_k$, $w$ is the weight vector to be optimized. $\Phi(a, c)$ is a mapping onto features that describe a feature, taking into account both the chemical structure $c$ and assay $k$. $C$ is a hyperparameter that determines the tradeoff between the margin size and training error. $\xi$ is the SVM slack variable, which allows misclassification during SVM model training. Within assay $k$, when compound $i$ had a smaller IC50 value than compound $j$, the ranking-SVM model was trained so that $w^T\Phi(a_k, c_i)$ became greater than $w^T\Phi(a_k, c_j)$. In the objective function for the training error, $\xi_{i,j,k}$ is only applied for compounds within the same assay. In the ranking-SVM method, designing the mapping function of $\Phi(a, c)$ is important yet difficult. In this study, only a compound-based (CB) representation (i.e., ECFP4) was used as the mapping function. Adding a constant value to each assay values (e.g., one-hot representations) does not change the output of ranking-SVM models that employ the linear kernel function because training errors solely depend on within-assay data, and such constant values would be cancelled out in the constraint function above. SVM-rank by Cornell University was used to create the ranking-SVM models.

**Scaling and NSC Methods.** In the NSC method, a QSAR model is built using compounds from different assays without any corrections to their IC50 values. This corresponds to creating a QSAR model of pIC50 for a biological target using active compounds extracted from different assay sources, such as the ChEMBL database. In contrast, the scaling method merges compound data sets from different assays by calibrating their IC50 values using a reference compound tested in the multiple assays. In a previous study, celecoxib was employed as the reference compound for COX 1 data sets, and IC50 values from nine training assays were calibrated so that the IC50 values of celecoxib became identical across the assays. In this study, the scaling method was also used combined with PLS-R and SVR, followed by ranking based on the predicted pIC50 values. To apply the scaling method, compounds with the same identity are necessary to calibrate the IC50 values across different assays.

**Data Integration Strategies.** Before constructing a QSAR model of IC50 (or rankings) for a specific target assay, IC50 data from different assays can be merged into the training data set to enhance the model predictive ability. When merging data from different assays, similar compounds play an important role according to the concept of the applicability domain. The concept of the applicability domain relates to the similarity principle of QSAR, where similar compounds exhibit similar activity. Limiting the compounds or assays to be integrated is preferable to using the entire library of compounds in all assays. Here, we propose two types of data integration strategies: AB and CB.

**AB Integration.** In AB integration, for predicting IC50 values for compounds in a specific assay, whole compounds from similar assays were merged into the training data. The concept of AB integration is illustrated in Figure 2A. Whether or not compounds are integrated is determined on the basis of assays. In Figure 2A, compounds from different assays are plotted in different colors. Compounds in the assays are integrated when they are within a certain distance threshold from the test assay. The distance between two assays is determined by the Euclidean distance between the means of the assay compounds using ECFP4 fingerprints (Figure 2B). Euclidean distances of 3.0, 5.0, and 7.0 were established as threshold values. For example, when compounds in assay D were tested using a threshold of 3.0, the compounds in assay E were integrated into the construction of the model (Figure 2).

**CB Integration.** In contrast to merging whole compounds in assays, CB integration selects compounds from multiple assays on the basis of individual compound distances to test compounds (Figure 3). As a result, only similar compounds from multiple assays are merged into the training data set. This similarity premise is not guaranteed when using AB integration. In this study, the Tanimoto similarity using ECFP4 was employed to define the similarity between compounds. The Tanimoto similarity values of 0.15 and 0.3 were tested as thresholds for data integration. The integration

![Figure 2](https://doi.org/10.1021/acsomega.1c00463)

**Figure 2.** Depiction of assay distance for data integration. As an exemplary case, the chemical space of ErbB1 inhibitors from several assays is visualized (A) using t-distributed stochastic neighbor embedding (t-SNE). Extended connectivity fingerprints with diameter of 4 (ECFP4) were used as a molecular representation. The corresponding distance matrix is shown in (B).

![Figure 3](https://doi.org/10.1021/acsomega.1c00463)

**Figure 3.** CB integration for QSAR model training. For test compounds in the red-shaded circle, integrated compounds from various assays are specified on the t-SNE surface. The yellow-shaded circles represent distance thresholds within which compounds are integrated into training of QSAR models of potency rankings.
criteria were also applied to the training compounds in the test assay, resulting in some of the training compounds unused.

**Model Selection Scheme.** In practical application, the best regression method with a data integration strategy should be determined for a given data set. For example, for some data sets, a simple PLS without data integration might work better than a rather complicated ranking-SVM method with the AB data integration strategy. To identify a proper model and a data integration strategy based solely on a training data set consisting of compounds with IC_{50} values, leave-one-out cross validation (LOOCV) was conducted. The Kendall rank correlation coefficients were calculated between the LOOCV-based rankings and actual rankings of the training compounds. When the ranking of a single compound was predicted using ranking-SVM, the single compound was ranked along with the rest of the holdout training compounds.

## RESULTS AND DISCUSSION

### Overall Performance Comparison

For each assay of a biological target, 20 strategies for rank prediction were tested. These strategies comprised ranking-SVM (nonintegration), ranking-SVM using CB-integrated data with three different similarity thresholds, ranking-SVM using AB-integrated data with three different distance thresholds, PLS-R and SVR using training data (nonintegrated), PLS-R and SVR using CB-integrated data with three different thresholds, PLS-R and SVR using the scaling method, PLS-R and SVR using one-hot vectors for the representation of assays, and PLS-R using AB descriptors. SVR using AB descriptors was not adopted because the Tanimoto kernel could not be applied. When CB-integrated data were used with PLS-R and SVR, the NSC method (i.e., simply using the nominal IC_{50} values from different assays) was employed. For each test data set, 20 strategies were ranked based on the Kendall rank correlation coefficient, and these rankings were collected over all 23 targets (excluding COX 1 from Table 1) to observe the distribution of method rankings. The results are presented in Figure 4 as violin plots and in Table 2 as representative statistics (mean and standard deviation). In Table 2, the statistics based on the median ranking for each assay (across 10 repetitions) are also reported to avoid including outliers. Furthermore, for comparing the Kendall rank correlation coefficients between two methods, paired t-tests were model performance using each of the 125 assays over the 23 biological targets were conducted. Among the nonintegrated strategies, ranking-SVM showed superior performance to PLS-R (p-value: 0.035 from two-tailed t-test). Because both modeling methods were linear, model construction with the LTR method was better than linear regression followed by ranking transformation based on predicted pIC_{50} values. Furthermore, comparing SVR with the Tanimoto kernel to ranking-SVM without data integration, SVR–Tanimoto was also less accurate than ranking-SVM, and the margin was larger than PLS (p-value: 0.008).

Focusing on different data integration strategies using PLS-R, the scaling method exhibited almost the same performance (mean of median rankings: 12.15) as the NSC method (PLS/NSC/CB/thres0.0: 12.24) and other traditional strategies (PLS/categorized by one-hot: 11.58, PLS/potency information: 12.35). Thus, the scaling method was not observed to have any advantages for ranking compounds.

However, using the SVR–Tanimoto kernel showed merits by data integration. When including the rest of assays in training data sets, the method placed a higher position (8.49) than simple SVR (10.93) supported by p-value: 0.000 from the two-tailed t-test. For this method, introducing similar compounds to training data sets did not increase the predictive ability of the models (SVR/NSC/CB/thres0.3: 8.65).

Ranking-SVM without data integration showed superior performance (mean of median rankings: 8.50) to PLS models, and integrating all the compounds from all assays for the analysis of a target drastically deteriorated the model’s performance (12.54) (p-value: 0.000). When CB integration
was applied to this method, performance improved significantly (SVM-rank/CB/thres0.3: 9.66) (p-value: 0.000).

In contrast, when AB methods were used, the smaller the neighborhood similarity threshold used, the better the ranking-SVM model performed. This result implies that integrating all compounds in a similar assay, a choice that is also defined by the compound distance, assured the selection of very similar compounds and also a sufficient number of compounds into the training data set. The best performance of ranking-SVM was observed when using ranking-SVM with AB integrated with a threshold distance of 3.0 (8.46). However, the

Figure 5. Performance comparison with/without data integration. Kendall rank correlation coefficients for estrogen receptor alpha as a target by (a) SVM-rank-based methods, (b) PLS-based (numerical prediction) methods including the scaling method, (c) SVR-based (numerical prediction) including the scaling method.
difference in performance from nonintegrated ranking-SVM was marginal (8.50) and not significant (p-value: 0.824).

We observed that whether or not data integration improved the models’ predictive ability depended on the target. Entire target-wise methodological comparison results are presented in Figures S1–S23 and Tables S2–S4 in the Supporting Information. Figure 5 shows the Kendall rank correlation coefficients of each integration strategy for assays on estrogen receptor alpha. As shown in Figure 5a, for assay 306245, ranking-SVM using all compounds in all assays (CB/thres0.0) performed better than the nonintegrated strategy. In contrast, for assay 371967, this integration strategy performed worse than the nonintegrated strategy. Overall, the CB-integrated strategy with a threshold of 0.3 worked well for this target. The results also support our recommendation of integrating a similar but sufficient number of compounds into the models to generate stable predictive performance. Among the PLS-R methods, the best-performing methods were not consistent among the assays for this target (Figure 5b). The threshold value of 0.3 showed better results than other thresholds, but this was not true for some targets (e.g., assay 306245). For this target, the scaling method did not show much advantage over the other integrated strategies but showed better performance than the methods without data integration. When SVR with the Tanimoto kernel was used, integrating all the compounds from the assays performed better than nonintegrated models (Figure 5c). No sudden performance deterioration was observed for assay 371967 by data integration in contrast to ranking-SVM. Comparing with ranking-SVM models, SVR models gave a stabler performance based on the variance of the Kendall rank correlation coefficient values. The means of the standard deviations across the assays by ranking-SVM and SVR using all the compounds from all the assays for this specific target were 0.06 and 0.03, respectively.

**Chemical Space of Two Exemplary Targets.** Our next question was in which cases AB-integrated training data could improve the model performance in ranking-SVM. To answer this question, two extreme cases (targets) were extracted and their chemical spaces were visualized using t-SNE. One target was hERG, for which the AB integration method effectively worked. The other was the epidermal growth factor receptor erbB1, for which AB integration impaired model performance. For hERG, the integrated model improved significantly without similarity restriction (SVM-rank/CB/thres0.0) (Figure 6A), as did AB integration strategies with distance thresholds of 7.0 and 5.0. In contrast, for ErbB1, AB and CB data integration impaired ranking-SVM model performance (Figure 6B). Interestingly, SVR with the Tanimoto kernel showed a stable performance by simply merging all the assays into training data to construct models. For hERG, the scaling method with SVR worked in contrast to PLS. For ErbB1, for which data integration did not contribute to performance improvement when ranking-SVM was used, data integration gave predictive performance at the same level as without data integration. This implies that traditional SVR with the Tanimoto kernel with CB integration (or one-hot representation) could be a first choice for predicting rankings of compounds for a specific target assay.

Two t-SNE maps using ECFP4 fingerprints for inhibitors of hERG (Figure 7a) and ErbB1 (Figure 7b) showed discrepancies of their compound distributions in the chemical space. hERG inhibitors from multiple assays occupied a wide range of chemical space. In contrast, ErbB1 inhibitors formed clusters based on their source assays, meaning that the compounds tested by the assays were different chemotypes. This also implies that it is difficult to find shared features that can increase model performance when the compounds in the assays do not share the same chemical space. Our current AB integration criterion (i.e., distances between means of compounds across assays) was not sufficient in this way. Further improvement of the integration criterion should be tested to consider information from compound distributions.

**Selection of the Best Predictive Model.** Our analysis revealed that the model performance was highly dependent on the data set and assays employed. Therefore, selecting the optimal method based on a certain validation scheme was
important. For this purpose, model selection using LOOCV was tested to identify the optimal model from the 14 modeling strategies employed without “SVR with Tanimoto kernels” for saving the calculation cost. However, models selected using LOOCV did not perform well on the test compounds at all. Only 13 of 125 assays outperformed the best-performing model (i.e., ranking-SVM using AB-integrated data with a threshold of 3.0). An exemplary target (hERG) is presented in Figure 8, which shows that the model selected by LOOCV was inferior to nonintegrated or AB-integrated data for most assays.

Performance Comparison Using COX 1 Data Sets.

Overall, ranking-SVM using AB-integrated training data with a distance threshold of 5.0 and SVR with the scaling method showed the best performance of all methods (mean of rankings: 6.77 for ranking-SVM/AB/thres5.0, 7.10 for SVR/scaling) (Figure 9). PLS-R with the scaling method was slightly superior to the NSC method by 0.57 in terms of the mean of rankings. In PLS-R, performing data integration using one-hot vectors or potency information gave higher predictive ability than using the scaling method. The means of performance rankings produced using one-hot vectors and potency information were 10.97 and 9.97, respectively. The average number of assays integrated when using AB integration with thresholds of 5.0 and 3.0 were 1.78 and 0.13, respectively. Collecting assay data including similar compounds improved

the predictive ability of the ranking-SVM models. In SVR with the Tanimoto kernel, the scaling method gave the highest performance (7.10), and the difference between nonintegrated and scaling-SVR was slightly significant (p-value: 0.104). This strategy showed almost identical performance with SVR with the one-hot vector (8.88) and NSC (8.70), and p-values were 0.346 and 0.362, respectively.

■ CONCLUSIONS

Integrating heterogeneous data from various assays to predict IC₅₀ values for a target assay is important because it could compensate for a lack of experimental data from single assays and could make use of accumulated data from different sources. In this study focusing on ranking-oriented QSAR models, our proposed ranking-SVM method using AB-integrated data with a limited distance threshold outperformed
other nonintegrated strategies. It performed as good as SVR with the Tanimoto kernel using all the compounds from different assays. The performance validation of 20 different prediction strategies using ChEMBL data sets across 23 biological targets revealed that many compounds from similar assays, as defined by distances between compounds, can provide important benefits in terms of data integration when combined with ranking-SVM. Although the overall performance of ranking-SVM using only compounds from a target assay was comparable, the AB integration method was statistically better than the nonintegrated method for some assays.

Ranking-SVM might work better when compound distributions of different assays share regions of chemical space, judging by our visual inspection of t-SNE-based chemical space. Nevertheless, when employing SVR with the Tanimoto kernel as a predictive model, simple data integration worked as good as ranking-SVM-based strategies. In this case, the scaling method and using one-hot vector encoding of assay information were also preferable.

When using the COX 1 data sets, ranking-SVM with the integration of AB data with a distance threshold of 5.0 outperformed PLS-based methods, but not SVR with the Tanimoto kernel-based methods. This result also supports that using compounds from similar assays is important when employing ranking-SVM models. One possible improvement could be the refinement of the definition of similarity between assays. The mean of compound descriptor values (i.e., the assay center) does not provide information about the distribution of compounds across the assay. This is our next challenge to improve the practicality of ranking-oriented QSAR models.

**ASSOCIATED CONTENT**

1. Supporting Information

   The Supporting Information is available free of charge [here](https://pubs.acs.org/doi/10.1021/acsomega.1c00463).

   Biological target information extracted from the ChEMBL database (Table S1), modeling results including predictive rankings of 20 QSAR modeling methods (Figures S1–S23), and statistics of modeling methods (Tables S2–S4) [PDF](#)

**AUTHOR INFORMATION**

**Corresponding Author**

Kimito Funatsu — Data Science Center, Nara Institute of Science and Technology, Nara 630-0192, Japan; Department of Chemical System Engineering, School of Engineering, The University of Tokyo, Tokyo 113-8656, Japan; orcid.org/0000-0002-9368-0302; Phone: +81-3-5841-7751; Email: funatsu@dsc.naist.jp; Fax: +81-3-5841-7771

**Authors**

Katsuhisa Matsumoto — Graduate School of Science and Technology, Nara Institute of Science and Technology, Nara 630-0192, Japan

Tomoyuki Miyao — Graduate School of Science and Technology, Nara Institute of Science and Technology, Nara 630-0192, Japan; Data Science Center, Nara Institute of Science and Technology, Nara 630-0192, Japan; orcid.org/0000-0002-8769-2702

Complete contact information is available at: [here](https://pubs.acs.org/doi/10.1021/acsomega.1c00463)

**Notes**

The authors declare no competing financial interest.

**ACKNOWLEDGMENTS**

We thank the OpenEye Scientific Software, Inc. for providing a free academic license of the OpenEye chemistry toolskits. This work was supported by the JSPS KAKENHI Grant number JP20K19922.

**REFERENCES**

(1) Alberto Castillo-Garit, J.; Abad, C.; Enrique Rodriguez-Borges, J.; Marrero-Ponce, Y.; Torrens, F. A Review of QSAR Studies to Discover New Drug-like Compounds Actives Against Leishmaniasis and Trypanosomiasis. *Curr. Top. Med. Chem.* **2012**, *12*, 852–865.

(2) Muratov, E. N.; Bajorath, J.; Sheridan, R. P.; Tetko, I. V.; Filimonov, D.; Poroikov, V.; Oprea, T. I.; Baskin, I. I.; Varnek, A.; Roßberg, A.; et al. QSAR without Borders. *Chem. Soc. Rev.* **2020**, *49*, 3525–3565.

(3) Mendes, D.; Gaulton, A.; Bento, A. P.; Chambers, J.; De Veij, M.; Felix, E.; Magariños, M. P.; Mosquera, J. F.; Mutovo, F.; Nowotka, M.; et al. ChEMBL: Towards Direct Deposition of Bioassay Data. *Nucleic Acids Res.* **2019**, *47*, D930–D940.

(4) Kim, S.; Chen, J.; Cheng, T.; Gindulyte, A.; He, J.; He, S.; Li, Q.; Shoemaker, B. A.; Thiessen, P. A.; Yu, B.; et al. PubMed 2019 Update: Improved Access to Chemical Data. *Nucleic Acids Res.* **2019**, *47*, D1102–D1109.

(5) Gilson, M. K.; Liu, T.; Battalk, M.; Nicola, G.; Hwang, L.; Chong, J. BindingDB in 2015: A Public Database for Medicinal Chemistry, Computational Chemistry and Systems Pharmacology. *Nucleic Acids Res.* **2016**, *44*, D1045–D1053.

(6) Law, V.; Knox, C.; Djoumbou, Y.; Jewison, T.; Guo, A. C.; Liu, Y.; Maclejewski, A.; Arndt, D.; Wilson, M.; Neveu, V.; et al. DrugBank 4.0: Shedding New Light on Drug Metabolism. *Nucleic Acids Res.* **2014**, *42*, D1091–D1097.

(7) Muresan, S.; Petrov, P.; Southan, C.; Kjellberg, M. J.; Kogej, T.; Tyrrchan, C.; Varkonyi, P.; Xie, P. H. Making Every SAR Point Count: The Development of Chemistry Connect for the Large-Scale Integration of Structure and Bioactivity Data. *Drug Discov. Today* **2011**, *16*, 1019–1030.

(8) Cortés-Ciriano, I.; Škuta, C.; Bender, A.; Svozil, D. QSAR-Derived Affinity Fingerprints (Part 2): Modeling Performance for Potency Prediction. *J. Cheminf.* **2020**, *12*, 41.

(9) Alberga, D.; Trisciu zig, D.; Montaruli, M.; Leonetti, F.; Mangiatordi, G. F.; Nicolotti, O. A New Approach for Drug Target and Bioactivity Prediction: The Multifingerprint Similarity Search Algorithm (MuSSeL). *J. Chem. Inf. Model.* **2019**, *59*, 586–596.

(10) Svensson, F.; Aniceto, N.; Norinder, U.; Cortes-Ciriano, I.; Spith, O.; Carlsson, L.; Bender, A. Conformal Regression for Quantitative Structure-Activity Relationship Modeling - Quantifying Prediction Uncertainty. *J. Chem. Inf. Model.* **2018**, *58*, 1132–1140.

(11) Duchowicz, P. Linear Regression QSAR Models for Polo-Like Kinase-1 Inhibitors. *Cells* **2018**, *7*, 13.

(12) Das, N. R.; Mishra, S. P.; Achary, P. G. R. Evaluation of Molecular Structure Based Descriptors for the Prediction of PEC50(M) for the Selective Adenosine A2A Receptor. *J. Mol. Struct.* **2021**, *1232*, No. 130080.

(13) López, A. F. F.; Martínez, O. M. M.; Hernández, H. F. C. Evaluation of Amaryllidaceae Alkaloids as Inhibitors of Human Acetylcholinesterase by QSAR Analysis and Molecular Docking. *J. Mol. Struct.* **2021**, *1225*, No. 129142.

(14) Ziaja, A.; Žuk, J.; Bartuzi, D.; Matosicki, D.; Poso, A.; Kaczor, A. A. The Universal 3D QSAR Model for Dopamine D2 Receptor Antagonists. *Int. J. Mol. Sci.* **2019**, *20*, 4555.

(15) Kovalishyn, V.; Tanchuk, V.; Charochkina, L.; Semenuta, I.; Prokopenko, V. Predictive QSAR Modeling of Phosphodiesterase 4 Inhibitors. *J. Mol. Graph. Model.* **2012**, *32*, 32–38.
(16) Kalliokoski, T.; Kramer, C.; Vulpetti, A.; Gedeck, P. Comparability of Mixed IC50 Data—A Statistical Analysis. PLoS One 2013, 8, e61007.

(17) Kramer, C.; Kalliokoski, T.; Gedeck, P.; Vulpetti, A. The Experimental Uncertainty of Heterogeneous Public K i Data. J. Med. Chem. 2012, 55, 5165−5173.

(18) Tarasova, O. A.; Urusova, A. F.; Filimonov, D. A.; Nicklaus, M. C.; Zakharov, A. V.; Poroikov, V. V. QSAR Modeling Using Large-Scale Databases: Case Study for HIV-1 Reverse Transcriptase Inhibitors. J. Chem. Inf. Model. 2015, 55, 1388−1399.

(19) Prous Science launches Integrity drug discovery portal. Reuters Events, Pharma.

(20) Yung-Chi, C.; Prusoff, W. H. Relationship between the Inhibition Constant (KI) and the Concentration of Inhibitor Which Causes 50 per Cent Inhibition (I50) of an Enzymatic Reaction. Biochem. Pharmacol. 1973, 22, 3099−3108.

(21) Lagunin, A. A.; Geronikaki, A.; Eleftheriou, P.; Pogodin, P. V.; Zakharov, A. V. Rational Use of Heterogeneous Data in Quantitative Structure-Activity Relationship (QSAR) Modeling of Cyclooxygenase/Lipoxygenase Inhibitors. J. Chem. Inf. Model. 2019, 59, 713−730.

(22) Liu, T. Y. Learning to Rank for Information Retrieval. Found. Trends Inf. Retr. 2009, 3, 225−331.

(23) Zhang, W.; Ji, L.; Chen, Y.; Tang, K.; Wang, H.; Zhu, R.; Jia, W.; Cao, Z.; Liu, Q. When Drug Discovery Meets Web Search: Learning to Rank for Ligand-Based Virtual Screening. J. Comput. Chem. 2015, 36, 86−108.

(24) Suzuki, S. D.; Ohue, M.; Akiyama, Y. PKRank: A Novel Learning-to-Rank Method for Ligand-Based Virtual Screening Using Pairwise Kernel and RankSVM. Artif. Life Robot 2018, 23, 205−212.

(25) Joachims, T. Optimizing Search Engines Using Clickthrough Data. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2002. DOI: DOI: 10.1145/775047.775067

(26) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. J. Chem. Inf. Model. 2010, 50, 742−754.

(27) OpenEye. OEChem, version 1.9.1; OpenEye Scientific Software Inc.: SantaFe, NM.

(28) Agarwal, S.; Dugar, D.; Sengupta, S. Ranking Chemical Structures for Drug Discovery: A New Machine Learning Approach. J. Chem. Inf. Model. 2010, 50, 716−731.

(29) Rathke, F.; Hansen, K.; Brefeld, U.; Müller, K. R. Structrank: A New Approach for Ligand-Based Virtual Screening. J. Chem. Inf. Model. 2011, 51, 83−92.

(30) Wold, S.; Sjöström, M.; Eriksson, L. PLS-Regression: A Basic Tool of Chemometrics. Chemom. Intell. Lab. Syst. 2001, 58, 109−130.

(31) Drucker, H.; Surges, C. J. C.; Kaufman, L.; Smola, A.; Vapnik, V. Support Vector Regression Machines. Advances in Neural Information Processing Systems, 1997.

(32) Vapnik, V. Pattern Recognition Using Generalized Portrait Method. Autom. Remote Control 1963, 24, 774−780.

(33) Joachims, T.; Training Linear SVMs in Linear Time. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2006. DOI: DOI: 10.1145/1150402.1150429

(34) Tetko, I. V.; Sushko, I.; Pandey, A. K.; Zhu, H.; Tropsha, A.; Papa, E.; Oberg, T.; Todeschini, R.; Fourches, D.; Varnek, A. Critical Assessment of QSAR Models of Environmental Toxicity against Tetrahymena pyriformis: Focusing on Applicability Domain and Overfitting by Variable Selection. J. Chem. Inf. Model. 2008, 48, 1733−1746.

(35) Dragos, H.; Gilles, M.; Alexandre, V. Predicting the Predictability: A Unified Approach to the Applicability Domain Problem of Qsar Models. J. Chem. Inf. Model. 2009, 49, 1762−1776.

(36) A.H.-L.; Y. G. S.. Concepts and Applications of Molecular Similarity. J. Mol. Struct. 1992, 269, 376−377.

(37) Rogers, D. J.; Tanimoto, T. T. A Computer Program for Classifying Plants. Science 1960, 132, 1115−1118.

(38) Student. The Probable Error of a Mean. Biometrika 1908, 6, 1.