JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

pubs.acs.org/jcim Article

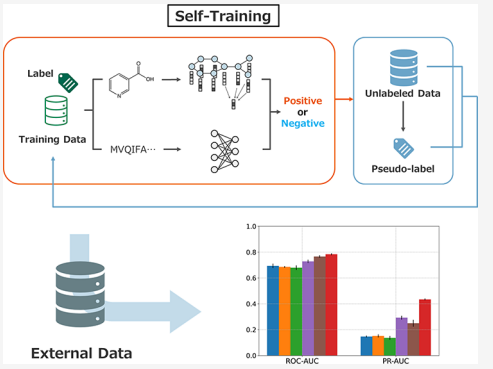
Improving Compound–Protein Interaction Prediction by Self-Training with Augmenting Negative Samples

Takuto Koyama, Shigeyuki Matsumoto,* Hiroaki Iwata, Ryosuke Kojima, and Yasushi Okuno*

Cite This: J. Chem. Inf. Model. 2023, 63, 4552–4559 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Identifying compound-protein interactions (CPIs) is crucial for drug discovery. Since experimentally validating CPIs is often time-consuming and costly, computational approaches are expected to facilitate the process. Rapid growths of available CPI databases have accelerated the development of many machine-learning methods for CPI predictions. However, their performance, particularly their generalizability against external data, often suffers from a data imbalance attributed to the lack of experimentally validated inactive (negative) samples. In this study, we developed a self-training method for augmenting both credible and informative negative samples to improve the performance of models impaired by data imbalances. The constructed model demonstrated higher performance than those constructed with other conventional methods for solving data imbalances, and the improvement was prominent for external datasets. Moreover, examination of the prediction score thresholds for pseudo-labeling during self-training revealed that augmenting the samples with ambiguous prediction scores is beneficial for constructing a model with high generalizability. The present study provides guidelines for improving CPI predictions on real-world data, thus facilitating drug discovery.



INTRODUCTION

Identifying compound-protein interactions (CPIs) is of great importance for drug discovery.[1] Especially in the early drug development stage, many CPIs are experimentally evaluated to find hit compounds and avoid undesired off-target effects using various biological assay techniques. These experimental approaches are often time-consuming and expensive.[2] Therefore, computational approaches are expected to accelerate the identification of CPIs in drug discovery.

Computational approaches are roughly categorized into two groups: structure-based and structure-free (chemical genomics-based) methods. Structure-based methods evaluate CPIs by calculating physical chemistry-based scores from three-dimensional (3D) complex models generated by molecular docking simulations.[3−5] When reliable 3D structures of the target proteins are available, docking methods can be applied without prior information on known interactions. Their performance heavily depends on the scoring function's accuracy and predicted docking pose. Structure-free methods are primarily chemoinformatic approaches that utilize prior information on known CPIs to predict unknown interactions. They can accurately predict CPIs with relatively low computational costs when the prior information covers sufficient pharmacological space.

In recent years, publicly available CPI databases, such as ChEMBL, BindingDB, PubChem, DrugBank, and PDBbind,[6−10] have been rapidly growing, thereby accelerating the development of various structure-free methods using machine learning (ML) algorithms.[11,12] These ML methods consider compound information, protein information, and their interactions in a unified framework.[13−16] More recently, the CPI prediction models using deep learning (DL) techniques, such as convolutional neural networks (CNNs), graph convolutional networks (GCNs),[17] and transformer algorithms,[18] have shown substantially improved predictive performance and interpretability.[19−24] These models can extract the feature representations of compounds and proteins during end-to-end learning of their interactions.

The performance of structure-free methods using ML techniques is often hampered by the quality of the training data derived from known interactions available in CPI databases. In many cases, experimentally validated inactive (negative) samples are lacking in public databases, which leads to class imbalances in the available CPI data. This insufficiency results in poor performance of ML models on out-of-domain samples and overestimation caused by the majority class in CPI predictions. To compensate for the insufficiency of negative samples, the random pairing method (referred to as

Received: February 21, 2023
Published: July 17, 2023

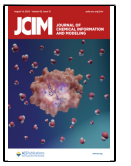


Table 1. Summary of the ChEMBL, BioPrint, Davis, and BindingDB Datasets

	ChEMBL		BioPrint	Davis	BindingDB	
	GPCR	kinase	GPCR	kinase	GPCR	kinase
compound	70,545	66,652	2621	68	43,293	119,518
protein	106	113	89	442	134	307
interaction	111,064	99,398	232,334	30,056	54,619	171,721
positive	107,908	94,911	15,630	9125	47,774	152,010
negative	3156	4487	216,704	20,931	6845	19,711

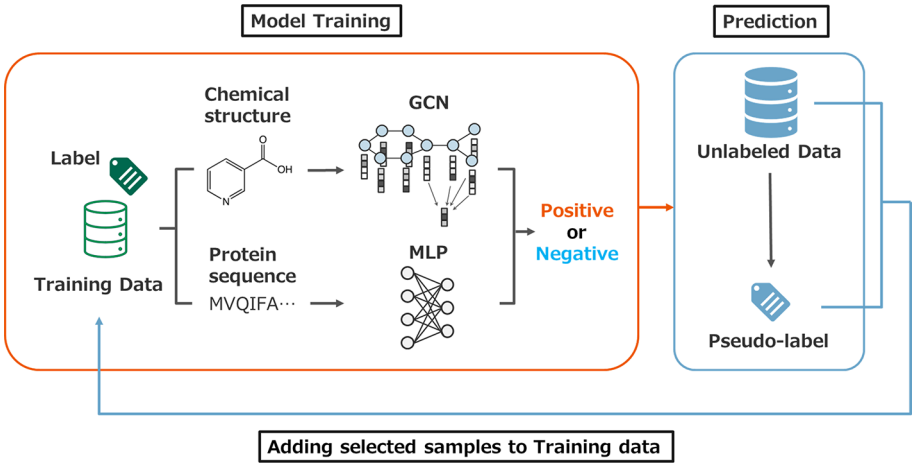


Figure 1. The overall workflow of our proposed method using self-training.

Random Negative henceforth)[21,23,25] has been employed in previous studies. This method randomly generates negative samples from compound−protein pairs that have not been experimentally confirmed to interact. However, the extracted samples could contain potentially active (positive) samples, resulting in low credibility. To address this problem, Liu et al. proposed a method for generating highly credible negative samples using a systematic screening framework to select samples distant from positive CPIs.[26] Nevertheless, a recent study demonstrated that similarity-controlled negative samples that are overly distant from positives are "easy" to learn for the model training, and the model's generalization performance on external datasets is significantly lower than that of the *Random Negative* model.[27]

To overcome these limitations, we propose a new self-training[28] method for effectively augmenting negative samples. This method was applied to graph-based CPI prediction models and successfully improved the model performance, including generalizability. Concretely, the CPI prediction models trained with our approach showed higher robustness on the external dataset than other methods addressing the class imbalance. Furthermore, examination of the parameters in our approach revealed that the pseudo-labeled samples predicted with "near-boundary" scores were more beneficial for model generalizability than the samples readily classified as negative.

MATERIALS AND METHODS

Dataset. The dataset for constructing the binary classification model and evaluating its performance was derived from ChEMBL.[10] The external datasets for verifying generalizability were constructed from BioPrint (Eurofins, Luxemburg City, Luxemburg), Davis,[29] and BindingDB.[9] The threshold for annotating positive and negative labels was set to 10 μM. At the threshold, the ChEMBL and BindingDB datasets primarily consisted of positive interactions, while the BioPrint and Davis datasets contained a large volume of experimentally validated negative samples. CPI data of G protein-coupled receptor (GPCR) and kinase families were used in the present study. The details of these four datasets are described below, and their statistics are summarized in Table 1.

ChEMBL. Activity data targeting human proteins were collected from ChEMBL version 31.[10] We extracted CPI data with "IC50," "Ki," "EC50," or "Kd" as the standard type, "single protein" as the protein type, and "B" as the assay type, and then we selected the target proteins with over 200 interacted compounds. The GPCR and kinase datasets were extracted from the preprocessed ChEMBL dataset based on the family names obtained from Swiss-Prot.[30] The data points with pChEMBL values $\geq$ 5 (activity $\leq$ 10 μM) and <5 (activity > 10 μM) were labeled as positive and negative, respectively. pChEMBL is defined as −log (molar IC50, XC50, EC50, AC50, Ki, Kd, or Potency).

BioPrint. Real-world screening data targeting GPCR families were obtained from BioPrint, which contains experimental data on the inhibitory activity at a dose of 10 μM. CPIs were assigned as positive if the inhibition rate at 10 μM was 50% or more.

Davis. Real-world screening data targeting kinase families were obtained from the Davis[29] dataset, composed of binding affinity information with dissociation constant (Kd) values for 68 drugs and 442 kinase proteins. This version was the same as found in the Github repository of DeepDTA[31] (https://github.com/hkmztrk/DeepDTA/). Data points were annotated as positive if Kd < 10 μM and negative if Kd = 10 μM (indicating weak or inactive).

BindingDB. The external dataset in which positive interactions outnumber negatives was derived from the BindingDB.[9] The activity data with IC50 values targeting

https://doi.org/10.1021/acs.jcim.3c00269
J. Chem. Inf. Model. 2023, 63, 4552−4559

human proteins were extracted and labeled into two classes in the same way as in the ChEMBL.

Workflow of Our Proposed Method Using Self-Training. Self-training begins with training a teacher model with labeled data. The teacher model is used to generate pseudo-labels by predicting unlabeled data. After adding the selected pseudo-labeled data to the training data, a student model is trained with the updated data. This process is iterated using the student model as the next teacher model. Figure 1 illustrates the workflow of our proposed method using self-training, and the following steps were implemented.

Step 1: Build a Teacher Model. A teacher model $f_T(x)$ is trained by minimizing the binary cross-entropy loss on the labeled data $\{(x_1,y_1), (x_2,y_2), ...(x_n,y_n)\}$

$$\frac{1}{n}\sum_{i}^{n} \mathrm{BCE}(y_i, f_T(x_i)) \tag{1}$$

where the binary cross entropy is defined as $\mathrm{BCE}(y_i, \hat{y}_i) = y_i \cdot \log(\hat{y}_i) + (1 - y_i)\cdot\log(1 - \hat{y}_i)$, with y_i and $\hat{y}_i$ representing a label and prediction score, respectively.

Step 2: Predict Unlabeled Data. The teacher model $f_T(x)$ is used to generate pseudo-negative labels for the unlabeled data $\{\tilde{x}_1, \tilde{x}_2, ... \tilde{x}_n\}$. If $\tilde{x}_i$ satisfies $\phi \le f_T(\tilde{x}_i) \le 0.5$, $\tilde{x}_i$ is regarded as a pseudo-negative sample, where $\phi \in [0,0.5)$ is a threshold parameter and $i \in \{1,2,...,n\}$ is an index of the unlabeled data.

Step 3: Add Selected Samples to the Training Data. Pseudo-labeled negative samples are added to the labeled data.

Step 4: Build a Student Model. A student model $f_S(x)$ is trained by minimizing the binary cross-entropy loss on both labeled and pseudo-labeled data

$$\frac{1}{n}\sum_{i}^{n} \mathrm{BCE}(y_i, f_S(x_i)) + \frac{1}{m}\sum_{i}^{m} \mathrm{BCE}(\tilde{y}_i, f_S(\tilde{x}_i)) \tag{2}$$

where $\tilde{y}_i = 0$ represents the pseudo-negative label.

Step 5. Substitute the student model for the teacher and return to Step 2.

$$f_S \rightarrow f_T \tag{3}$$

In CPI predictions, $x = (c,p)$ indicates compound-protein pairs, where c is a compound and p is a protein. Unlabeled data $\{\tilde{x}_1, \tilde{x}_2, ... \tilde{x}_n\}$ were generated by randomly pairing compounds and target proteins. These compounds were derived from the preprocessed ChEMBL dataset (317,244 compounds). A subset of the data to be pseudo-labeled was randomly sampled from all the unlabeled data at each iteration. We set the sample size of a subset to 750 and 500 K for the GPCR and kinase datasets, respectively. For the parameter ϕ in Step 2, we set $\phi = 0.20$ for both GPCR and kinase datasets. The iterations were terminated if the negative sample size of every target protein reached the same number of positive interactions or if the number of iterations reached the predetermined maximum. We set the maximum number of iterations to nine.

CPI Prediction Model. We used a binary classification model for the CPI prediction to determine whether a given compound and protein interact. A model with a multimodal neural network was constructed using kMoL (https://github.com/elix-tech/kmol), which is an open-source chemoinformatics library based on kGCN[32] and can combine various architectures using several input features, such as molecular graphs and extended connectivity fingerprints (ECFP)[33] for compounds and Bag-of-Words and tokens for protein sequences. This study mainly employed molecular graphs for compounds and Bag-of-Words for protein sequences as input. The output of kMoL was activated by a sigmoid function, resulting in a score ranging from 0 to 1. Additional detailed information on the model architecture and model training is provided in the Supporting Information Text S1, Tables S1−S3.

Performance Evaluation. Our proposed method was internally evaluated by 5-fold cross-validation using the ChEMBL dataset. The ChEMBL dataset was randomly divided into five folds so that the training and test splits were stratified according to the sample size of each target. Twenty percent of the randomly selected training data was used as validation data. We then evaluated the model performances on the external datasets Davis (kinase), BioPrint (GPCR), and BindingDB (GPCR and kinase). The area under the receiver operating characteristic curve (ROC-AUC) and the area under the precision−recall curve (PR-AUC) were used for performance evaluation. The ROC curve plots the true positive rate against the false positive rate, and the PR curve plots the precision against the recall.

The effectiveness of our proposed method was compared with that of four other methods addressing class imbalance along with the baseline model, which trains the original imbalanced data.

Weighted Loss. This method generally applies to a DL framework trained with a class-imbalanced dataset.[34] We adopted the simplest weighted loss approach, which weights the binary cross-entropy loss according to the sample size of the training dataset

$$l(x, y) = \frac{1}{n}\sum_{i}^{n} \left\{ \frac{n_0}{n_1} y_i \cdot \log(\hat{y}_i) + (1 - y_i)\cdot\log(1 - \hat{y}_i) \right\} \tag{4}$$

where n_0 and n_1 represent the number of negative and positive samples ($n_0 + n_1 = n$).

Random Undersampling. The majority-class sample size was randomly reduced to be the same as the minority-class samples to balance the two classes.[35]

Random Negative. This method generates negative samples by randomly pairing a compound and protein to eliminate the data imbalance.[21,23,25] Negative samples were generated until they compensated for the shortage in the original training dataset.

Similarity Controlled. Negative samples are selected from unlabeled pairs based on the dissimilarity from positive samples. We adopted the same algorithm as employed in the previous study by Yaseen et al.[27] and set the inter-class similarity α as 0.10.

RESULTS AND DISCUSSION

Performance on the Internal Dataset: ChEMBL. To validate the effectiveness of self-training in CPI prediction, we conducted 5-fold cross-validations on the ChEMBL dataset. As shown in Table 2, the average ROC-AUC scores of the baseline models on the internal dataset for the GPCR and kinase families were 0.9139 and 0.9175, respectively. By applying our self-training method to the imbalanced data, the sizes of the positive and negative samples were progressively approximated, and the ratio of positive samples in each protein

https://doi.org/10.1021/acs.jcim.3c00269
J. Chem. Inf. Model. 2023, 63, 4552−4559

Table 2. Performance Evaluation on the ChEMBL Dataset

	GPCR		Kinase	
	ROC-AUC (Std[a])	PR-AUC (Std)	ROC-AUC (Std)	PR-AUC (Std)
baseline	0.9139 (0.0083)	0.9962 (0.0005)	0.9175 (0.0035)	0.9946 (0.0004)
weighted loss	0.9182 (0.0079)	0.9965 (0.0005)	0.9149 (0.0089)	0.9945 (0.0006)
random undersampling	0.9026 (0.0046)	0.9958 (0.0002)	0.8982 (0.0066)	0.9933 (0.0006)
random negative	0.9034 (0.0063)	0.9961 (0.0004)	0.9053 (0.0012)	0.9940 (0.0002)
similarity controlled	0.9234 (0.0059)	0.9969 (0.0002)	0.9166 (0.0060)	0.9947 (0.0005)
self-training (ours)	**0.9336 (0.0035)**	**0.9974 (0.0003)**	**0.9336 (0.0021)**	**0.9960 (0.0002)**

[a]Standard deviation (Std) was calculated from the results of 5-fold cross-validation.

converged asymptotically to 0.5 toward the final iteration (Figure S1). The model trained with the updated data showed a slightly better performance based on the ROC-AUC scores of GPCR (0.9336) and kinase (0.9336). This improvement can be attributed to the expanded data distribution and resolved data imbalance. The performance of our method was better than that of the *Random Negative* models (GPCR: 0.9034 and kinase: 0.9053). This improvement was expected because the generated samples based on the predictive scores would be more credible than the randomly paired negative samples. Our models outperformed the *Random Undersampling*, *Weighted Loss*, and *Similarity-Controlled* methods (Table 2). We additionally implemented three types of 5-fold cross-validation recently adopted in CPI prediction and other fields:[36,37] Compound cross-validation (CV), protein CV, and compound−protein CV (see Supporting Information Text S2). As shown in Figure S2, our approach performed best in all the CVs, validating our self-training method's effectiveness.

Performance on the External Datasets: BioPrint, Davis, and BindingDB. To evaluate the generalizability of the currently constructed models, we compared the model performances on the external datasets, BioPrint for the GPCR families and Davis for the kinase families. Here, we employed the PR-AUC score as the primary metric, which is more appropriate for evaluating model performance on an imbalanced dataset dominated by negative samples.[38] Our model showed improved performances on the BioPrint and Davis datasets, with PR-AUC scores of 0.4344 and 0.5792, respectively, which were 28.7 and 17.5% better than the baseline model (Figure 2). Our method also performed better than other methods for addressing data imbalance. The PR-AUC scores of the second-best *Random Negative* model were 0.2927 and 0.5491. These results demonstrated that our self-training method could significantly improve model robustness on external datasets. Our method also outperformed other models in F1 scores that is a harmonic mean of Precision and Recall since other models were highly biased toward one side, emphasizing the advantages of our approach (Tables S4 and S5). It should be noted that our method is not biased toward negative predictions because the prediction performance on the BindingDB datasets, which are dominated by positive interactions, was not compromised (ROC-AUC, PR-AUC, and Precision: 0.7778, 0.9517, and 0.9585 for GPCR, 0.8009, 0.9606, and 0.9604 for kinase) (Figure S3, Tables S6 and S7).

We analyzed the model's generalizability using the Davis dataset to evaluate its predictive performance at the protein level. As shown in Figure 3, the distribution of PR-AUC scores for our model (median 0.6253) was significantly higher than that of the baseline (median: 0.4888; Mann−Whitney *U* test, $p = 1.40 \times 10^{-5}$) and *Random Negative* (median: 0.5426; Mann−Whitney *U* test, $p = 1.30 \times 10^{-23}$). We separately evaluated the prediction performance for both seen and unseen proteins, which were stratified according to their presence in the training dataset (ChEMBL). The average PR-AUC scores of the baseline model, *Random Negative*, and our model for unseen proteins were 0.4504, 0.4877, and 0.5573, respectively, while those of the seen proteins were 0.4975, 0.6658, and 0.7144, respectively. Although a performance gap between the seen and unseen proteins was observed, our model outperformed the other models in both cases. This observation further supports the effectiveness of our method in improving its generalizability. For the BioPrint dataset, our model (median: 0.1695) outperformed the baseline model (median: 0.0479; Mann−Whitney *U* test: $p = 6.33 \times 10^{-4}$), while statistical significance was not found from *Random Negative* (median: 0.0928; Mann−Whitney *U* test: $p = 0.157$) at the protein level

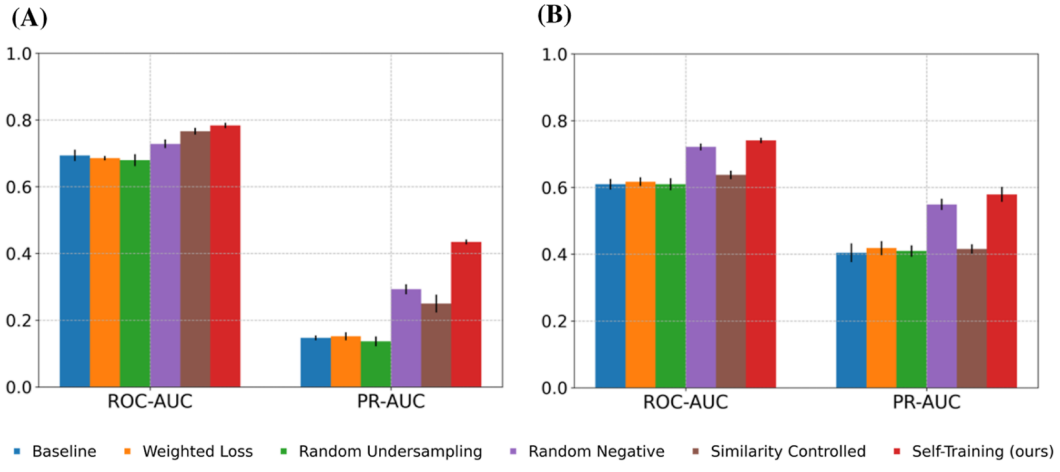


Figure 2. Performance evaluation on the external datasets. ROC-AUC and PR-AUC scores on the BioPrint dataset (A) and ROC-AUC and PR-AUC scores on the Davis dataset (B) for the evaluated six models. The error bars represent the standard deviations of the scores predicted by the six models in the cross-validation.

https://doi.org/10.1021/acs.jcim.3c00269
J. Chem. Inf. Model. 2023, 63, 4552−4559

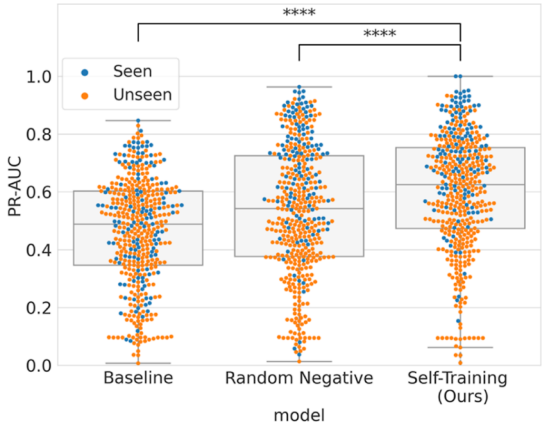


Figure 3. Performance evaluation at the protein level using the Davis dataset. The box plot shows the distribution of PR-AUC scores for individual proteins. Each dot in the swarm plot represents the PR-AUC score of each protein color-coded in blue and orange for seen and unseen proteins, respectively. The Mann−Whitney *U* test determined the statistical significance of the performance of our proposed method. ****$p < 0.0001$.

(Figure S4). The difficulty in observing the significant difference from *Random Negative* could be due to the small number of available proteins.

Applicability to Other CPI Prediction Models. To further demonstrate the usefulness of the self-training approach, we applied our method to a classical multi-layer perceptron (MLP) classifier using ECFP4 and Bag of Words as inputs and existing graph-based CPI prediction model, GraphDTA,[20] which uses graph isomorphism network[39] (GIN) and convolutional neural network (CNN). As shown in Tables 3 and 4, the PR-AUC scores for the BioPrint and Davis dataset were significantly improved by our method, and the improvements were superior to *Random Negative*. These results demonstrated that our self-training method widely applies to other CPI prediction models.

Table 3. PR-AUC Scores for the BioPrint Dataset Using Other Models

	kMoL (Std[a])	MLP (ECFP4 + BoW[b])	GraphDTA[20]
baseline	0.1471 (0.0072)	0.1319	0.1160
random negative	0.2927 (0.0149)	0.2328	0.2364
self-training (ours)	**0.4344 (0.0075)**	**0.4362**	**0.3279**

[a]Std: Standard deviation. [b]BoW: Bag of Words.

Table 4. PR-AUC Scores for the Davis Dataset Using Other Models

	kMoL (Std[a])	MLP (ECFP + BoW[b])	GraphDTA[20]
baseline	0.4042 (0.0279)	0.3629	0.4202
random negative	0.5491 (0.0171)	0.4909	0.4566
self-training (ours)	**0.5792 (0.0225)**	**0.5409**	**0.5136**

[a]Std: Standard deviation. [b]BoW: Bag of Words.

UMAP Visualization of Training Data. The improvement in the performance of our self-training approach can be attributed to clarifying the decision boundary by generating negative samples. To visualize the distribution of the training dataset and decision boundary, we mapped the 64-dimensional features extracted from the last hidden layer of the interaction module in kMoL (Supporting Information Text S1) onto a 2-dimensional space using UMAP.[40] The distribution of the originally labeled data in the baseline model showed that their decision boundary was ambiguous because most negative samples were embedded in the same space as the positive samples (Figure 4A). In contrast, the decision boundary in our self-training model was clarified by augmenting the negative samples covering known negatives (Figure 4B,C). The separation of the two classes into different spaces would contribute to improving the predictive performance. Similar results were obtained using the GPCR dataset (Figure S5).

Effect of the Score Threshold on Model Generalizability. The threshold of the score $f_T(\tilde{x}_i)$ for pseudo-labeling is an important parameter for model performance in a self-training algorithm. Conventional self-training approaches adopt unlabeled data with extremely confident scores in the pseudo-labeling step as the augmented samples.[28] Nevertheless, we selected unlabeled data with more ambiguous scores ($0.2 \leq f_T(\tilde{x}_i) \leq 0.5$ in the case of kinase families) as the added negative samples, which are hereafter referred to as near-boundary samples. To validate the effectiveness of the adopted score threshold, we compared the learning process on the ChEMBL dataset and the predictive performance on the Davis dataset with two other score thresholds: $f_T(\tilde{x}_i) \leq 0.2$ and $f_T(\tilde{x}_i) \leq 0.5$. The learning process of the model at the last iteration demonstrated that the validation ROC-AUC using confident negative samples ($f_T(\tilde{x}_i) \leq 0.2$) reached a plateau earlier than that using near-boundary samples (Figure 5A).

Meanwhile, the PR-AUC score of the model using near-boundary samples was better than that using confident samples (Figure 5B). The model self-trained with a threshold of $f_T(\tilde{x}_i) \leq 0.5$ showed intermediate behaviors (Figure 5). These results indicated that the near-boundary samples are more informative than the confident samples for model training; therefore, selecting near-boundary samples is crucial for achieving high generalizability. Confident samples are overly distant from positive interactions, thus constituting an easy-to-learn dataset, which is supported by the rapid convergence of the validation ROC-AUC (Figure 5A). This easy-to-learn situation would result in poor generalizability (Figure 5B), which is compatible with the situation discussed by Yaseen et al. when using similarity-controlled negative samples.[27] Similar behavior was also observed in the GPCR models (Figure S6). We further evaluate the effect of the hyperparameter in the pseudo-labeling process, ϕ, on the prediction performance, showing that higher ϕ leads to better prediction performance in the ChEMBL, Davis, and BioPrint datasets (Figure S7). This result also supports that near-boundary samples are more beneficial for improving the model performance.

CONCLUSIONS

In this study, we constructed a self-training method to improve model performance and generalizability suffering from a data imbalance in CPI prediction. Model evaluations demonstrated that our method outperformed the other methods for addressing class imbalance on both internal and external datasets. The analysis of the score thresholds for determining pseudo-labeled samples showed that improved generalizability

https://doi.org/10.1021/acs.jcim.3c00269
J. Chem. Inf. Model. 2023, 63, 4552−4559

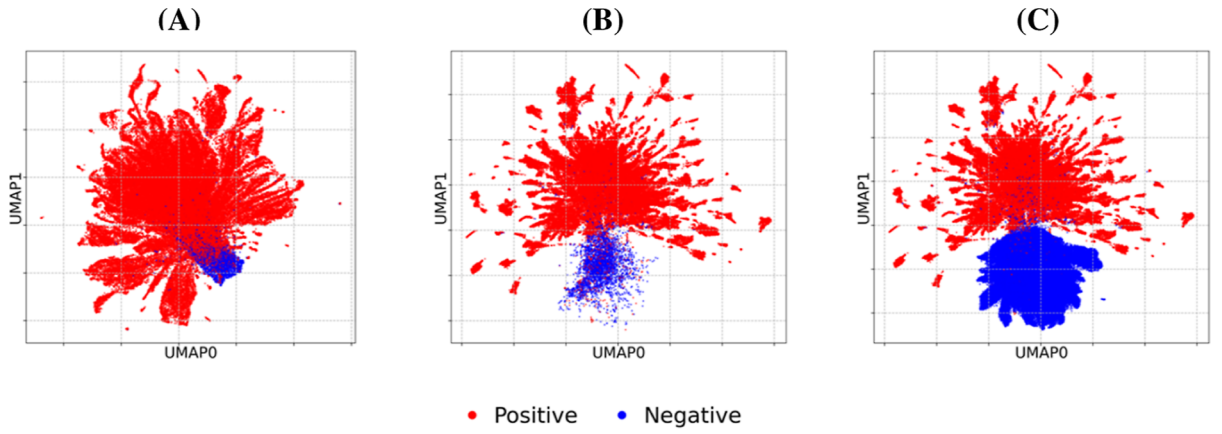


Figure 4. UMAP visualization of the kinase training dataset. Distribution of the originally labeled data extracted from the baseline model (A), labeled data extracted from our model (B), and labeled and pseudo-labeled data extracted from our model (C). Positive- and negative-labeled samples are represented by red and blue dots, respectively.

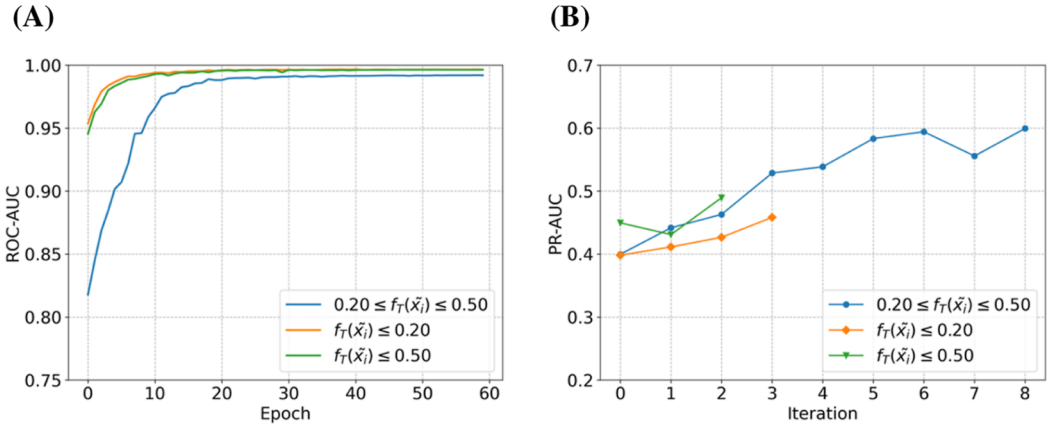


Figure 5. Model evaluation depending on the score thresholds. (A) Learning curves on validation data of kinase families at the last iteration. (B) PR-AUC scores on the Davis dataset during the self-training iterations.

could be achieved by using credible samples verified from predictive values and prioritizing the addition of informative samples near the boundary. In addition, we verified the effectiveness of self-training in other CPI prediction models, emphasizing that our method and the obtained insights could be broadly helpful in resolving data imbalance problems in other structure-free methods. The present study will provide guidelines for improving CPI prediction and facilitate the identification of novel interactions in real-world drug discovery.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jcim.3c00269.

Construction of compound protein interaction prediction model; performance evaluation by compound cross-validation (CV), protein CV, and compound−protein CV; node features of a graph representation; edge features of a graph representation; molecular properties calculated with RDKit; performance evaluation on the BioPrint dataset; performance evaluation on the Davis dataset; performance evaluation on the BindingDB GPCR dataset; performance evaluation on the BindingDB Kinase dataset; transition in the ratios of positive samples during self-training; performance evaluation for the ChEMBL datasets by three types of 5-fold cross-validation; performance evaluation on the BindingDB dataset; performance evaluation at a protein level on the BioPrint dataset; UMAP visualization of the GPCR training dataset; model evaluation depending on the score thresholds; and the effect of different parameters ϕ on the predictive performance (PDF)

■ AUTHOR INFORMATION

Corresponding Authors

Shigeyuki Matsumoto − *Graduate School of Medicine, Kyoto University, 606-8507 Kyoto, Japan;* orcid.org/0000-0001-9329-6362; Email: matsumoto.shigeyuki.4z@kyoto-u.ac.jp

Yasushi Okuno − *Graduate School of Medicine, Kyoto University, 606-8507 Kyoto, Japan; HPC- and AI-driven Drug Development Platform Division, RIKEN Center for Computational Science, Kobe 650-0047 Hyogo, Japan;* Email: okuno.yasushi.4c@kyoto-u.ac.jp

Authors

Takuto Koyama − *Graduate School of Medicine, Kyoto University, 606-8507 Kyoto, Japan;* orcid.org/0000-0002-9569-8370

https://doi.org/10.1021/acs.jcim.3c00269
J. Chem. Inf. Model. 2023, 63, 4552−4559

Hiroaki Iwata − *Graduate School of Medicine, Kyoto University, 606-8507 Kyoto, Japan;* orcid.org/0000-0001-9791-0008

Ryosuke Kojima − *Graduate School of Medicine, Kyoto University, 606-8507 Kyoto, Japan*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jcim.3c00269

Author Contributions

T.K., S.M., R.K., and Y.O. conceived and designed the study. T.K. performed the calculations and analyzed the data. T.K., H.I., S.M., and Y.O. contributed to interpreting the results. T.K. drafted the original manuscript, and the other authors revised the drafts. All the authors approved the final version of the manuscript.

Notes

The authors declare no competing financial interest.

Data and code are provided at the online public link https://github.com/clinfo/kMoL-ST.

■ ACKNOWLEDGMENTS

This research was supported by the Japan Agency for Medical Research and Development (AMED) under grant number JP22nk0101111 and JSPS KAKENHI grant number JP20K12063.

■ ABBREVIATIONS

CPI, compound−protein interaction; DL, deep learning; CNN, convolutional neural network; GCN, graph convolutional networks; ML, machine learning; GPCR, G protein-coupled receptors; ROC-AUC, area under the receiver operating characteristic curve; PR-AUC, area under the precision−recall curve

■ REFERENCES

(1) Keiser, M. J.; Setola, V.; Irwin, J. J.; Laggner, C.; Abbas, A. I.; Hufeisen, S. J.; Jensen, N. H.; Kuijer, M. B.; Matos, R. C.; Tran, T. B.; et al. Predicting new molecular targets for known drugs. *Nature* **2009**, *462*, 175−181.

(2) Macarron, R.; Banks, M. N.; Bojanic, D.; Burns, D. J.; Cirovic, D. A.; Garyantes, T.; Green, D. V.; Hertzberg, R. P.; Janzen, W. P.; Paslay, J. W.; et al. Impact of high-throughput screening in biomedical research. *Nat. Rev. Drug Discovery* **2011**, *10*, 188−195.

(3) Trott, O.; Olson, A. J. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J. Comput. Chem.* **2009**, *31*, 455−461.

(4) Meng, X. Y.; Zhang, H. X.; Mezei, M.; Cui, M. Molecular docking: a powerful approach for structure-based drug discovery. *Curr. Comput.-Aided Drug Des.* **2011**, *7*, 146−157.

(5) Meiler, J.; Baker, D. ROSETTALIGAND: Protein−small molecule docking with full side-chain flexibility. *Proteins: Struct., Funct., Bioinf.* **2006**, *65*, 538−548.

(6) Wishart, D. S.; Feunang, Y. D.; Guo, A. C.; Lo, E. J.; Marcu, A.; Grant, J. R.; Sajed, T.; Johnson, D.; Li, C.; Sayeeda, Z.; et al. DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res.* **2018**, *46*, D1074−D1082.

(7) Wang, R.; Fang, X.; Lu, Y.; Wang, S. The PDBbind database: Collection of binding affinities for protein− ligand complexes with known three-dimensional structures. *J. Med. Chem.* **2004**, *47*, 2977−2980.

(8) Kim, S.; Chen, J.; Cheng, T.; Gindulyte, A.; He, J.; He, S.; Li, Q.; Shoemaker, B. A.; Thiessen, P. A.; Yu, B.; et al. PubChem 2019 update: improved access to chemical data. *Nucleic Acids Res.* **2019**, *47*, D1102−D1109.

(9) Gilson, M. K.; Liu, T.; Baitaluk, 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.

(10) Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; et al. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* **2012**, *40*, D1100−D1107.

(11) Chen, X.; Yan, C. C.; Zhang, X.; Zhang, X.; Dai, F.; Yin, J.; Zhang, Y. Drug−target interaction prediction: databases, web servers and computational models. *Briefings Bioinf.* **2016**, *17*, 696−712.

(12) Zhou, Y.; Zhang, Y.; Lian, X.; Li, F.; Wang, C.; Zhu, F.; Qiu, Y.; Chen, Y. Therapeutic target database update 2022: facilitating drug discovery with enriched comparative data of targeted agents. *Nucleic Acids Res.* **2022**, *50*, D1398−D1407.

(13) Yamanishi, Y.; Araki, M.; Gutteridge, A.; Honda, W.; Kanehisa, M. Prediction of drug−target interaction networks from the integration of chemical and genomic spaces. *Bioinformatics* **2008**, *24*, i232−i240.

(14) 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.

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

(16) Bleakley, K.; Yamanishi, Y. Supervised prediction of drug−target interactions using bipartite local models. *Bioinformatics* **2009**, *25*, 2397−2403.

(17) Kipf, T. N.; Welling, M. Semi-supervised classification with graph convolutional networks. **2016**, arXiv:1609.02907. arXiv preprint.

(18) Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Advances in Neural Information Processing Systems*; Curran Associates, 2017; Vol. *30*.

(19) Tsubaki, M.; Tomii, K.; Sese, J. Compound−protein interaction prediction with end-to-end learning of neural networks for graphs and sequences. *Bioinformatics* **2019**, *35*, 309−318.

(20) Nguyen, T.; Le, H.; Quinn, T. P.; Nguyen, T.; Le, T. D.; Venkatesh, S. GraphDTA: predicting drug−target binding affinity with graph neural networks. *Bioinformatics* **2021**, *37*, 1140−1147.

(21) Lee, I.; Keum, J.; Nam, H. DeepConv-DTI: Prediction of drug-target interactions via deep learning with convolution on protein sequences. *PLoS Comput. Biol.* **2019**, *15*, No. e1007129.

(22) Huang, K.; Xiao, C.; Glass, L. M.; Sun, J. MolTrans: molecular interaction transformer for drug−target interaction prediction. *Bioinformatics* **2021**, *37*, 830−836.

(23) Hamanaka, M.; Taneishi, K.; Iwata, H.; Ye, J.; Pei, J.; Hou, J.; Okuno, Y. CGBVS-DNN: Prediction of Compound-protein Interactions Based on Deep Learning. *Mol. Inf.* **2017**, *36*, 1600045.

(24) Chen, L.; Tan, X.; Wang, D.; Zhong, F.; Liu, X.; Yang, T.; Luo, X.; Chen, K.; Jiang, H.; Zheng, M. TransformerCPI: improving compound−protein interaction prediction by sequence-based deep learning with self-attention mechanism and label reversal experiments. *Bioinformatics* **2020**, *36*, 4406−4414.

(25) Thafar, M. A.; Olayan, R. S.; Ashoor, H.; Albaradei, S.; Bajic, V. B.; Gao, X.; Gojobori, T.; Essack, M. DTiGEMS+: drug−target interaction prediction using graph embedding, graph mining, and similarity-based techniques. *J. Cheminf.* **2020**, *12*, 44.

(26) Liu, H.; Sun, J.; Guan, J.; Zheng, J.; Zhou, S. Improving compound−protein interaction prediction by building up highly credible negative samples. *Bioinformatics* **2015**, *31*, i221−i229.

(27) Yaseen, A.; Amin, I.; Akhter, N.; Ben-Hur, A.; Minhas, F. Insights into performance evaluation of compound−protein interaction prediction methods. *Bioinformatics* **2022**, *38*, ii75−ii81.

(28) Amini, M.-R.; Feofanov, V.; Pauletto, L.; Devijver, E.; Maximov, Y. Self-Training: A Survey. arXiv preprint arXiv:2202.12040 **2022**.

(29) Davis, M. I.; Hunt, J. P.; Herrgard, S.; Ciceri, P.; Wodicka, L. M.; Pallares, G.; Hocker, M.; Treiber, D. K.; Zarrinkar, P. P. Comprehensive analysis of kinase inhibitor selectivity. *Nat. Biotechnol.* **2011**, *29*, 1046−1051.
(30) Boeckmann, B.; Bairoch, A.; Apweiler, R.; Blatter, M.-C.; Estreicher, A.; Gasteiger, E.; Martin, M. J.; Michoud, K.; O'Donovan, C.; Phan, I. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.* **2003**, *31*, 365−370.
(31) Öztürk, H.; Özgür, A.; Ozkirimli, E. DeepDTA: deep drug−target binding affinity prediction. *Bioinformatics* **2018**, *34*, i821−i829.
(32) Kojima, R.; Ishida, S.; Ohta, M.; Iwata, H.; Honma, T.; Okuno, Y. kGCN: a graph-based deep learning framework for chemical structures. *J. Cheminf.* **2020**, *12*, 32.
(33) Rogers, D.; Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742−754.
(34) Özdemir, Ö.; Sönmez, E. B. Weighted cross-entropy for unbalanced data with application on covid x-ray images. *2020 Innovations in Intelligent Systems and Applications Conference (ASYU)*; IEEE, 2020; pp 1−6.
(35) Drummond, C.; Holte, R. C. C4. 5, class imbalance, and cost sensitivity: why under-sampling beats over-sampling. *Workshop on Learning from Imbalanced Datasets II*; ICML, 2003; Vol. *11*, pp 1−8.
(36) Bai, X.; Yin, Y. Exploration and augmentation of pharmacological space via adversarial auto-encoder model for facilitating kinase-centric drug development. *J. Cheminf.* **2021**, *13*, 95.
(37) Lihong, P.; Wang, C.; Tian, X.; Zhou, L.; Li, K. Finding lncrna-protein interactions based on deep learning with dual-net neural architecture. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*; IEEE, 2022; Vol. *19*, pp 3456−3468.
(38) Saito, T.; Rehmsmeier, M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS One* **2015**, *10*, No. e0118432.
(39) Xu, K.; Hu, W.; Leskovec, J.; Jegelka, S. How powerful are graph neural networks? **2018**, arXiv:1810.00826. arXiv preprint.
(40) McInnes, L.; Healy, J.; Melville, J. Umap: Uniform manifold approximation and projection for dimension reduction. **2018**, arXiv:1802.03426. arXiv preprint.

https://doi.org/10.1021/acs.jcim.3c00269
J. Chem. Inf. Model. 2023, 63, 4552−4559