ACS Medicinal Chemistry Letters

pubs.acs.org/acsmedchemlett

Letter

# Integrating Incompatible Assay Data Sets with Deep Preference Learning

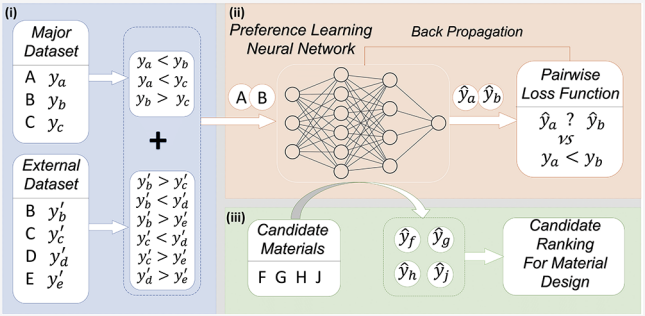
Xiaolin Sun, Ryo Tamura, Masato Sumita, Kenichi Mori, Kei Terayama, and Koji Tsuda*

Cite This: *ACS Med. Chem. Lett.* 2022, 13, 70−75

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

**ABSTRACT:** A large amount of bioactivity assay data is already accumulated in public databases, but the integration of these data sets for quantitative structure−activity relationship (QSAR) studies is not straightforward due to differences in experimental methods and settings. We present an efficient deep-learning-based approach called Deep Preference Data Integration (DPDI). For integrating outcome variables of different assay types, a surrogate variable is introduced, and a neural network is trained such that the total order induced by the surrogate variable is maximally consistent with given data sets. In a task of predicting efficacy of factor Xa inhibitors, DPDI successfully integrated 2959 molecules distributed in 129 assay data sets. In most of our experiments, data integration improved prediction accuracy strongly in interpolation and extrapolation tasks, indicating that DPDI is an effective tool for QSAR studies.




**KEYWORDS:** *Data integration, Preference learning, Deep learning, Bioactivity data*

A large number of bioassay data sets are accumulated in public databases, but their use is limited due to differences in experimental methods and settings. We propose a new deep learning model called Deep Preference Data Integration (DPDI) to enable the integration of incompatible data sets. Our method increases the value of public data sets by providing the means to reuse them.

a) Outcome values

| | A | B | C | D | E | F | G |
|---|---|---|---|---|---|---|---|
| Assay Type 1 | 0.1 | 1.3 | 0.9 | | | | |
| Assay Type 2 | | 0.05 | 0.01 | 0.08 | 0.001 | 0.02 | 0.11 |

b) Induced rankings

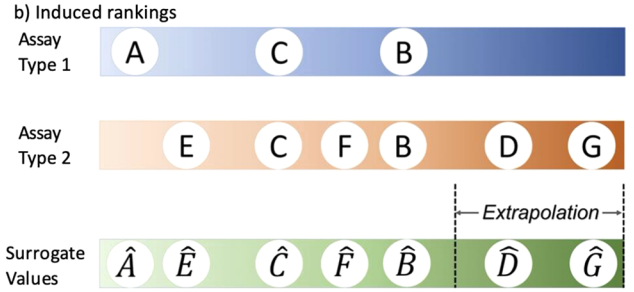
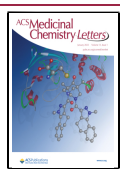



**Figure 1.** Data integration with a surrogate variable. (a) For ligands A−G, the outcome values for two different assay types are shown. (b) The first and second rows show the rankings according to the outcome values of corresponding assay types. The third row shows the ranking due to the surrogate values predicted with a neural network.

In quantitative structure−activity relationship (QSAR) studies, researchers are interested in investigating structural features of molecules that determine their bioactivities. Machine learning models are an essential part of QSAR studies, where bioactivities of a large number of molecules are induced from training examples. To maximize the size of a training set, one may consider combining multiple bioactivity assay data sets deposited in public databases such as ChEMBL and PubChem Bioassay.[1] The use of multiple data sets is, however, limited to only a few cases.[2−4] One of the main reasons lies in *incompatibility* of these data sets. Even if biological activities are represented in the same unit such as $IC_{50}$, the combination of these data sets may not lead to improvement of prediction accuracy in machine learning due to the differences in experimental methods and conditions.

For example, let us take the two bioactive assay data sets, ChEMBL968695 and ChEMBL3885775. In both assays, the target protein is factor Xa, a protease involved in the blood coagulation pathway.[5] It acts by cleaving prothrombin in two

**Received:** August 12, 2021
**Accepted:** December 27, 2021
**Published:** December 29, 2021







https://doi.org/10.1021/acsmedchemlett.1c00439
*ACS Med. Chem. Lett.* 2022, 13, 70−75

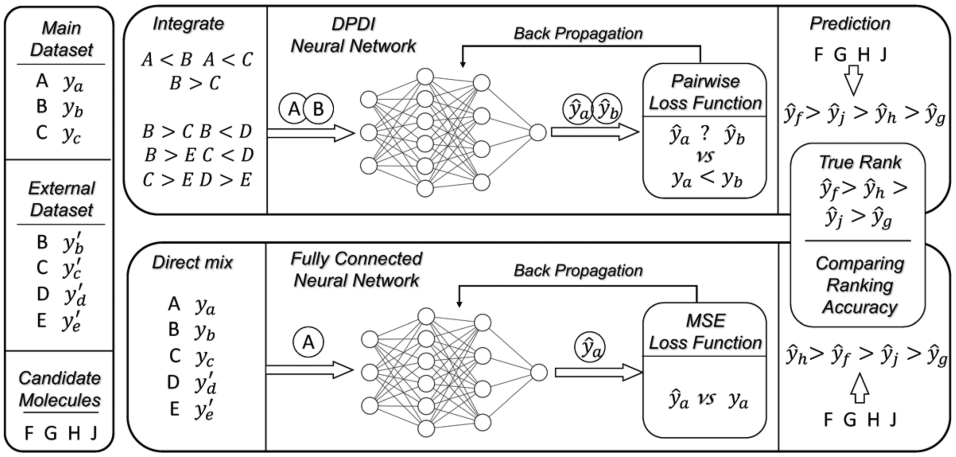



**Figure 2.** Experimental details. In *learning with integrated data set* (shown as integrate), the main data set and external data set are independently converted to preferences. After DPDI is trained with the preferences, the candidate molecules can be converted to surrogate values. After converting the surrogate values to preferences, it is compared with the true ranking. Normalized discounted cumulative gain (NDCG) is used as the accuracy measure. In *direct mix*, the main data set and external data set are used as they are. A fully connected network is trained by minimizing the mean squared loss (MSE) with both data sets, and the activity values of the candidate molecules are induced. After they are converted to preferences, NDCG is used to measure the accuracy.

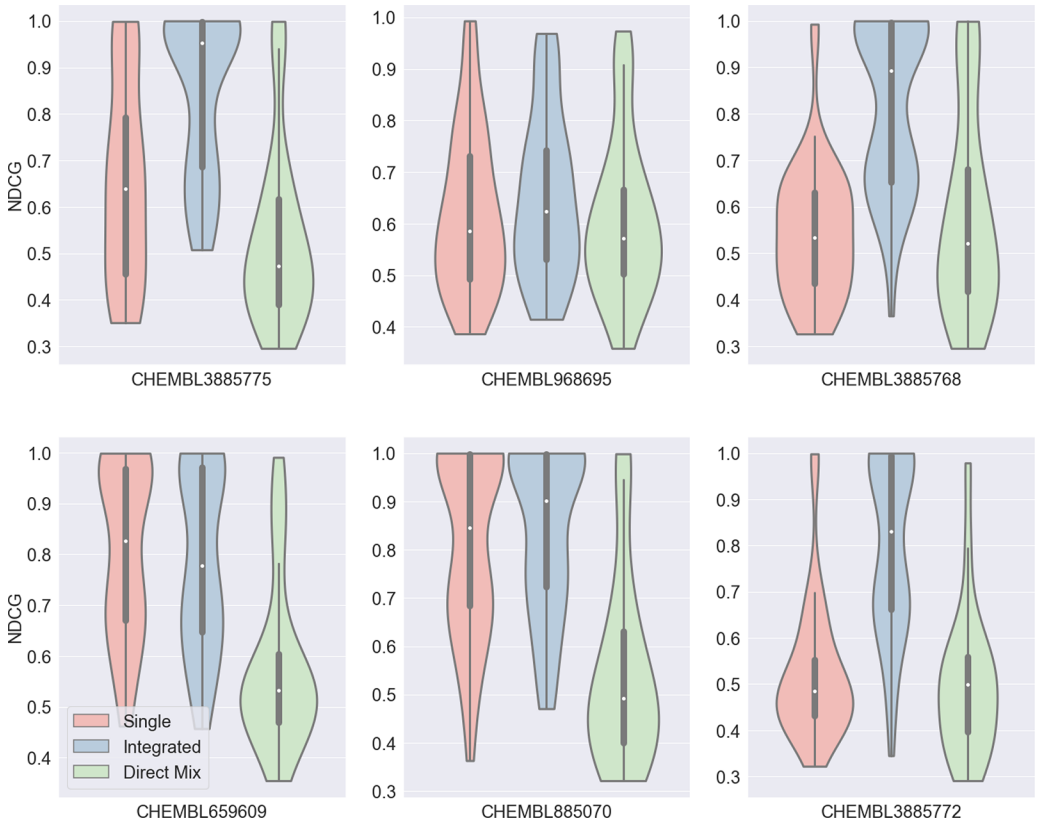



**Figure 3.** Results of interpolation experiments.

places, which yields active thrombin. The first assay data set is obtained by the human plasma-based thrombin generation test, where the activity is measured by the amount of thrombin, the product of factor Xa, in human plasma.[6] The second data set is obtained by the biochemical assay using fluorogenic peptide substrate.[7] A fluorogenic peptide substrate consists of a peptide that factor Xa can cleave and a fluorophore. The substrate is normally not fluorescent, but fluorescence is restored, when factor Xa cleaves off the fluorophore. Using this method, one can measure the activity of factor Xa by measuring the fluorescence intensity. Measurements from completely different assay types, as exemplified above, cannot be compared directly, and mixing such data without any treatment may be harmful for machine learning.

Assume that $n$ ligands are represented as $d$-dimensional fingerprints $x_1,\ldots,x_n \in \{0,1\}^d$. Denote by $y_{ji} \in \mathbb{R}$ the outcome of ligand $i$ for assay type $j$. Typically, some of the values of the outcome variables are not available (Figure 1a). One possible way to integrate such data sets is multitask learning,[8] where a


https://doi.org/10.1021/acsmedchemlett.1c00439
ACS Med. Chem. Lett. 2022, 13, 70−75

**Table 1. List of ChEMBL Assay Data Sets Used as the Main Data Set**[a]

| | | | NDCG (mean ± STD) | | | | |
|---|---|---|---|---|---|---|---|
| | | | interpolation | | | extrapolation | |
| main data set | size | source (document year) | single | integrated | direct mix | single | integrated |
| CHEMBL3885775 | 56 | K4DD project | 0.66 ± 0.21 | 0.85 ± 0.17 | 0.63 ± 0.23 | 0.41 ± 0.14 | 0.36 ± 0.12 |
| CHEMBL968695 | 55 | scientific literature (2009) | 0.62 ± 0.15 | 0.65 ± 0.16 | 0.61 ± 0.15 | 0.35 ± 0.15 | 0.43 ± 0.17 |
| CHEMBL3885768 | 55 | K4DD project | 0.54 ± 0.14 | 0.82 ± 0.18 | 0.59 ± 0.22 | 0.37 ± 0.09 | 0.41 ± 0.08 |
| CHEMBL659609 | 62 | scientific literature (2004) | 0.81 ± 0.17 | 0.78 ± 0.18 | 0.57 ± 0.16 | 0.24 ± 0.06 | 0.46 ± 0.20 |
| CHEMBL885070 | 46 | scientific literature (2002) | 0.81 ± 0.19 | 0.84 ± 0.17 | 0.54 ± 0.19 | 0.33 ± 0.23 | 0.42 ± 0.20 |
| CHEMBL3885772 | 55 | K4DD project | 0.53 ± 0.15 | 0.80 ± 0.19 | 0.50 ± 0.15 | 0.30 ± 0.09 | 0.46 ± 0.08 |

[a]Test accuracies in different experimental settings are summarized. For information about the K4DD project, see Schuetz et al.[18] The sources of CHEMBL968695, CHEMBL959609, and CHEMBL885070 are Zhang et al.,[6] Jia et al.,[19] and Zhang et al.,[20] respectively.

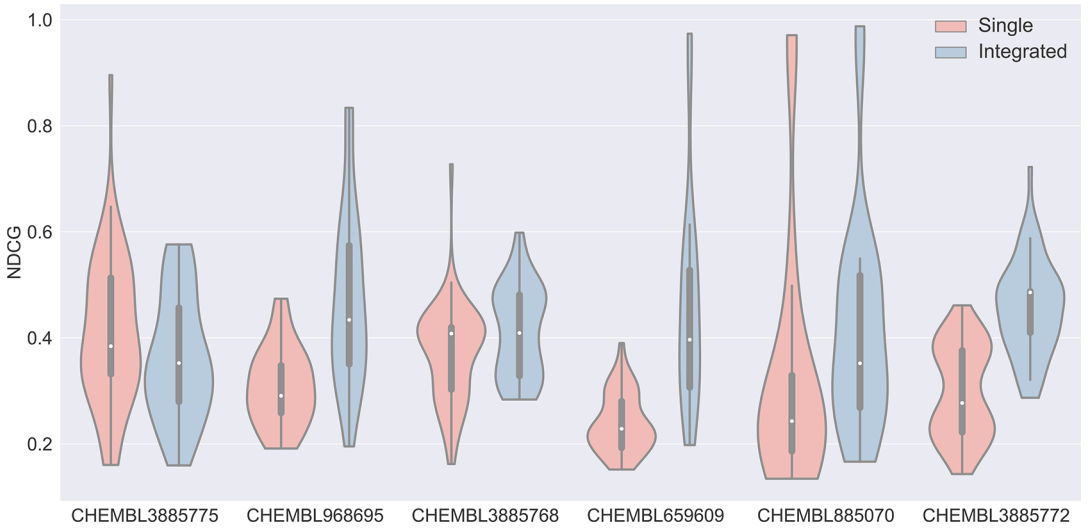



**Figure 4.** Results of extrapolation experiments.

machine learning model is trained to predict all outcomes from a fingerprint. However, the number of available ligands for an assay type can be extremely small (e.g., 2 or 3), hence accurate prediction of all the outcome variables would not be feasible. In this paper, we consider a virtual outcome variable $\hat{y}_i$ and call it the surrogate variable. A neural network model is trained to predict the surrogate variable from a fingerprint such that the total order induced by the surrogate variable conforms to all available data. To this aim, each data set is represented as a set of pairwise preferences (i.e., larger-than relationship, $\succ$). For example, assay type 1 in Figure 1b is represented as $C \succ A$, $B \succ C$, and $B \succ A$. The neural network is trained to minimize the number of preferences contradicting the total order by the surrogate variable. When a new ligand is given, the neural network predicts its surrogate value. A user can place the new ligand in the ranking of any assay type to understand how promising it is.

In the literature,[9,10] it is reported that the accuracy of machine learning models depends on the domain of applicability, i.e., the outcome range that encloses training examples. Machine learning is powerful in interpolation (i.e., prediction for test examples within the domain) but poor in extrapolation (i.e., prediction for those outside the domain). Notably, DPDI often expands the domain of applicability. For assay type 1 in Figure 1b, the domain of applicability is from A to B. When both assay types are integrated into the surrogate values, the domain is expanded to from A to G. Since a ligand better than the known ones is always wanted in virtual screening, extrapolation is more important than interpolation. In our computational experiments with 129 ChEMBL assay data sets, we observed strong improvement of extrapolation accuracy as a result of data integration. On the other hand, simple mixing of assay data sets resulted in accuracy deterioration. This result demonstrates that DPDI can overcome differences in assay types and enables the effective use of public data for better virtual screening. In addition, DPDI was shown to be more scalable in comparison to an alternative Gaussian-process-based preference learning model,[11] indicating that DPDI can be applied to large-scale projects without difficulty.

In DPDI, a fully connected neural network[12] is employed to predict the surrogate value from a fingerprint. Throughout this paper, 300 dimensional Mol2vec fingerprints[13] are used due to high expression ability. Shibayama et al.[21] reported better prediction performances of Mol2vec in comparison to existing fingerprints. The hyperparameters of the network are adjusted using a black-box optimization software, Optuna.[14] The hyperparameters and their ranges are as follows: the number of hidden layers (1−5), the number of units in each layer (4−1024), learning rate (0.0001−0.1), dropout rate (0−0.4), and optimizer type (Adam or stochastic gradient descent).

Each assay data set is converted to pairwise preferences and summarized into one training set $D = \{u_m \succ v_m\}_{m=1}^{M}$, where $u_m$ and $v_m$ are indices of ligands and $M$ is the total number of preferences. Since there are multiple assays, it is possible that the same pair of ligands appears multiple times. Let $\hat{y}_i$ denote


https://doi.org/10.1021/acsmedchemlett.1c00439
ACS Med. Chem. Lett. 2022, 13, 70−75

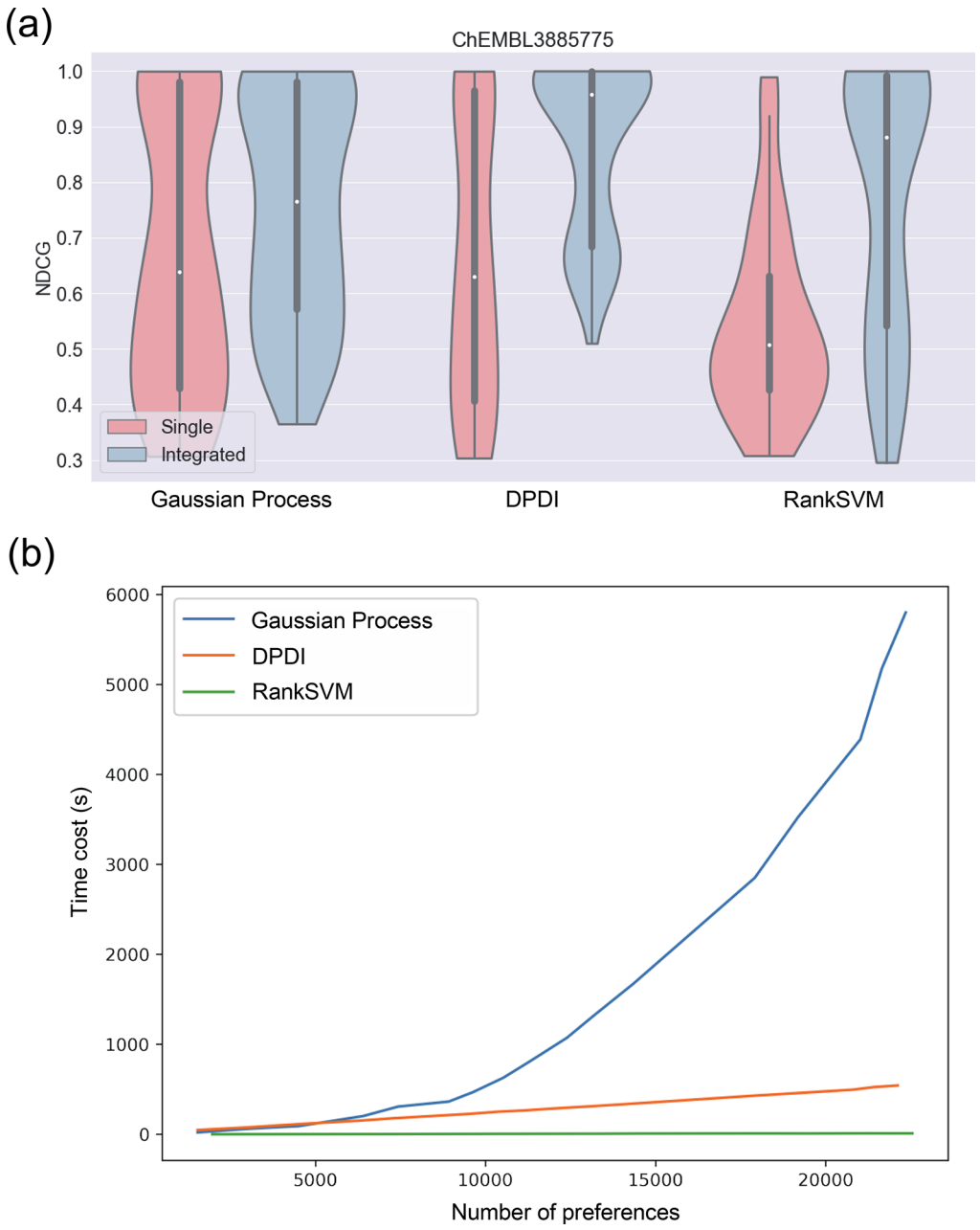



**Figure 5.** (a) Accuracy of Gaussian process, DPDI, and rankSVM in interpolation experiments for ChEMBL3885765. (b) Computational time of Gaussian process, DPDI, and rankSVM.

the surrogate value of ligand *i*. We would like to train the network to minimize a loss function that represents the number of training examples contradicting the order induced by the surrogate variable. To make the neural network trainable, however, a loss function has to be differentiable. To this aim, the number of contradicting examples is approximated by the following cross entropy function,

$$L = -\sum_{m=1}^{M} [I(v_m \succ u_m)\log P(v_m \succ u_m) + I(u_m \succ v_m) \times \log P(u_m \succ v_m)] \quad (1)$$

where $P(u \succ (v)$ is defined via surrogate values as

$$P(u \succ v) = \frac{\exp(\hat{y}_u)}{\exp(\hat{y}_u) + \exp(\hat{y}_v)} \quad (2)$$

and $I(\cdot)$ is the indicator function that returns 1 if the condition inside the parentheses is satisfied and otherwise 0.

We collected 129 bioactivity assay data sets about factor Xa from ChEMBL database (Supporting Information). We chose factor Xa, because of its clinical importance and availability of quite a few data sets in public databases. Factor Xa is a target for the development of new anticoagulants for the treatment of pathologic arterial and venous thrombosis.[22] Each data set contains from 2 to 85 ligands, and the total number of ligands is 5929. A data set is selected as *main data set*, which is then divided into training, validation, and test sets in the fraction of 3:1:1. The validation set is kept aside to monitor the loss


https://doi.org/10.1021/acsmedchemlett.1c00439
ACS Med. Chem. Lett. 2022, 13, 70−75

during the neural network training and hyperparameter tuning. In this section, the data set is divided randomly to test DPDI's interpolation performance. We compared the following three scenarios: In one scenario called *learning with single data set*, only the training set taken from the main data set is used. In the second scenario called *learning with integrated data set*, the training set from the main data set is integrated with all the other 128 assay data sets via DPDI. In the third scenario called *direct mix*, the training set from the main data set is simply mixed with all the other data sets without any treatment. A fully connected neural network is trained with the squared loss function. Hyperparameter tuning is performed in the same way as DPDI. See Figure 2 for experimental details.

In all scenarios, the test accuracy is computed by comparing the ranking due to predicted surrogate values against the ground-truth ranking. As the accuracy measure, we employed normalized discounted cumulative gain (NDCG).[15,16] Let us assume that the entity ranked at $i$th position in the ground-truth ranking is ranked at $R(i)$th position in the predicted ranking. Discounted cumulative gain (DCG) is defined as

$$\sum_{i=1}^{c} \frac{c - R(i)}{\log(i+1)} \tag{3}$$

where $c$ is the number of all entities. NDCG is the ratio of DCG to its maximum possible value. It is one if the two rankings match completely, and a lower value indicates poorer match.

Computational experiments are performed with each of six assay data sets listed in Table 1 designated as the main data set. These are the largest ones among all the data sets. Figure 3 shows the distribution of test accuracy for 50 different data divisions. Their summary statistics are shown in Table 1 as well. Notably, the test accuracy of the direct mix scenario was worse than that of the single data scenario in most cases. This result illustrates the difficulty of data integration due to differences in assay types. Comparing single and integrated data sets, the accuracy improved in five out of six cases, indicating the DPDI makes the effective use of additional information included in other data sets.

Next, we tested extrapolation performance of DPDI. To simulate extrapolation, the main data set is divided as follows. First, the test set is designated as the ligands with top 20% outcome. The rest is randomly divided into training and validation data sets in the fraction of 3:1. Figure 4 and Table 1 show the distribution and summary statistics of test accuracy, respectively. First of all, the test accuracy is significantly lower than that in interpolation. It indicates that extrapolation is a much more difficult task than interpolation. In five out of six cases, the integrated data scenario by DPDI outperformed the single data scenario. For CHEMBL659609, the improvement is dramatic; i.e., the average test accuracy is almost doubled. This result implies that DPDI can help extrapolation by expanding the domain of applicability.

We compare DPDI with two existing methods for preference-based data integration. One is the Gaussian process-based approach by Sun et al.,[11] and the other is the linear support vector machine (SVM)-based approach (rankSVM) by Matsumoto et al.[4] Due to the high computational cost of the Gaussian process model, we conducted a scaled-down interpolation experiment of integrating ChEMBL3885775 with five other data sets listed in Table 1. Figure 5a shows the accuracy of the three methods. The accuracy of DPDI was highest, indicating superior modeling ability of deep neural networks. Computational time for training from preference data is summarized in Figure 5b. RankSVM is a linear model and most efficient to train. As in most deep learning models, DPDI showed linear growth as the number of preferences increases. Gaussian process was particularly slow, showing superlinear growth. Among the three methods, DPDI achieved high standards both in accuracy and scalability.

We presented a deep learning approach, DPDI, for integrating multiple bioassay data sets. The significance of our method is that public data sets, otherwise useless, can be used to our advantage. DPDI converts data sets to a set of preferences. A favorable point of using preferences is that one can use bioassay data sets without any preprocessing. To derive clinically useful information from bioactivity data, researchers search for molecular substructures related to the bioactivity by statistical analysis.[17] By integrating multiple data sets into surrogate values, the number of samples used in statistical analysis is increased, leading to more conclusive results. A possible drawback of DPDI is that the user receives prediction results in the form of ranking, not an exact outcome value. We anticipate that this point does not affect scientists' decision making, because assay outcomes are always error prone and small changes may not be critically important. To serve the community, we made our PyTorch-based code publicly available at https://github.com/tsudalab/PrefIntNN.

## ■ ASSOCIATED CONTENT

### Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acsmedchemlett.1c00439.

Assay ChEMBL IDs, number of molecules in each assay, and detailed description of each assay (XLSX)

## ■ AUTHOR INFORMATION


### Corresponding Author

**Koji Tsuda** − *Graduate School of Frontier Sciences, The University of Tokyo, Kashiwa, Chiba 277-8561, Japan; Research and Services Division of Materials Data and Integrated System, National Institute for Materials Science, Tsukuba, Ibaraki 305-0047, Japan; RIKEN Center for Advanced Intelligence Project, Tokyo 103-0027, Japan;* orcid.org/0000-0002-4288-1606; Email: tsuda@k.u-tokyo.ac.jp

### Authors

**Xiaolin Sun** − *Graduate School of Frontier Sciences, The University of Tokyo, Kashiwa, Chiba 277-8561, Japan*

**Ryo Tamura** − *Graduate School of Frontier Sciences, The University of Tokyo, Kashiwa, Chiba 277-8561, Japan; Research and Services Division of Materials Data and Integrated System and International Center for Materials Nanoarchitectonics (WPI-MANA), National Institute for Materials Science, Tsukuba, Ibaraki 305-0047, Japan; RIKEN Center for Advanced Intelligence Project, Tokyo 103-0027, Japan;* orcid.org/0000-0002-0349-358X

**Masato Sumita** − *International Center for Materials Nanoarchitectonics (WPI-MANA), National Institute for Materials Science, Tsukuba, Ibaraki 305-0047, Japan; RIKEN Center for Advanced Intelligence Project, Tokyo 103-0027, Japan*




https://doi.org/10.1021/acsmedchemlett.1c00439
ACS Med. Chem. Lett. 2022, 13, 70−75

**Kenichi Mori** − *Astellas Pharma Inc., Tsukuba, Ibaraki 305-8585, Japan*

**Kei Terayama** − *RIKEN Center for Advanced Intelligence Project, Tokyo 103-0027, Japan; Graduate School of Medical Life Science, Yokohama City University, Yokohama 230-0045, Japan;* orcid.org/0000-0003-3914-248X

Complete contact information is available at: https://pubs.acs.org/10.1021/acsmedchemlett.1c00439

### Author Contributions

K. Tsuda and K. Terayama conceived the idea and designed the study. K.M. prepared the assay data sets. X.S, R.T., and M.S. implemented the algorithm and performed experiments. All authors have given approval to the final version of the manuscript.

### Funding

This study was partly funded by Astellas Pharma Inc. X.S. would like to gratefully acknowledge the financial support from the China Scholarship Council (CSC No. 201809120018).

### Notes

The authors declare the following competing financial interest(s): K.M. is an employee of Astellas Pharma Inc. The study is partly funded by Astellas Pharma Inc.

## ■ ABBREVIATIONS

QSAR, quantitative structure−activity relationship; DPDI, deep preference data integration; NDCG, normalized discounted cumulative gain; SVM, support vector machine

## ■ REFERENCES

(1) Muresan, S.; Petrov, P.; Southan, C.; Kjellberg, M. J.; Kogej, T.; Tyrchan, 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 Discovery Today* **2011**, *16* (23−24), 1019−1030.

(2) 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* (7), 1388−1399.

(3) 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.

(4) Matsumoto, K.; Miyao, T.; Funatsu, K. Ranking-Oriented Quantitative Structure−Activity Relationship Modeling Combined with Assay-Wise Data Integration. *ACS Omega* **2021**, *6*, 11964.

(5) Pinto, D. J.; Smallheer, J. M.; Cheney, D. L.; Knabb, R. M.; Wexler, R. R. Factor Xa Inhibitors: Next-Generation Antithrombotic Agents. *J. Med. Chem.* **2010**, *53* (17), 6243−6274.

(6) Zhang, P.; Huang, W.; Wang, L.; Bao, L.; Jia, Z. J.; Bauer, S. M.; Goldman, E. A.; Probst, G. D.; Song, Y.; Su, T.; et al. Discovery of Betrixaban (PRT054021), N-(5-Chloropyridin-2-Yl)-2-(4-(N, N-Dimethylcarbamimidoyl) Benzamido)-5-Methoxybenzamide, a Highly Potent, Selective, and Orally Efficacious Factor Xa Inhibitor. *Bioorg. Med. Chem. Lett.* **2009**, *19* (8), 2179−2185.

(7) Grimm, J. B.; Heckman, L. M.; Lavis, L. D. The Chemistry of Small-Molecule Fluorogenic Probes. *Prog. Mol. Biol. Transl. Sci.* **2013**, *113*, 1−34.

(8) Zhang, Y.; Yang, Q. A Survey on Multi-Task Learning. *IEEE Trans. Knowl. Data Eng.* **2021**, 1.

(9) Meredig, B.; Antono, E.; Church, C.; Hutchinson, M.; Ling, J.; Paradiso, S.; Blaiszik, B.; Foster, I.; Gibbons, B.; Hattrick-Simpers, J.; et al. Can Machine Learning Identify the next High-Temperature Superconductor? Examining Extrapolation Performance for Materials Discovery. *Mol. Syst. Des. Eng.* **2018**, *3* (5), 819−825.

(10) Sutton, C.; Boley, M.; Ghiringhelli, L. M.; Rupp, M.; Vreeken, J.; Scheffler, M. Identifying Domains of Applicability of Machine Learning Models for Materials Science. *Nat. Commun.* **2020**, *11* (1), 4428.

(11) Sun, X.; Hou, Z.; Sumita, M.; Ishihara, S.; Tamura, R.; Tsuda, K. Data Integration for Accelerated Materials Design via Preference Learning. *New J. Phys.* **2020**, *22* (5), 055001.

(12) Goodfellow, I.; Bengio, Y.; Courville, A. *Deep Learning*; MIT Press, 2016.

(13) Jaeger, S.; Fulle, S.; Turk, S. Mol2vec: Unsupervised Machine Learning Approach with Chemical Intuition. *J. Chem. Inf. Model.* **2018**, *58* (1), 27−35.

(14) Akiba, T.; Sano, S.; Yanase, T.; Ohta, T.; Koyama, M. Optuna: A next-Generation Hyperparameter Optimization Framework. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*; 2019; pp 2623−2631.

(15) Jarvelin, K.; Kekäläinen, J. IR Evaluation Methods for Retrieving Highly Relevant Documents. *ACM SIGIR Forum* **2017**, *51* (2), 243−250.

(16) Scholkopf, B.; Tsuda, K.; Vert, J.-P. *Kernel Methods in Computational Biology*; MIT Press, 2004.

(17) Capecchi, A.; Probst, D.; Reymond, J.-L. One molecular fingerprint to rule them all: drugs, biomolecules, and the metabolome. *J. Cheminform.* **2020**, *12*, 43.

(18) Schuetz, D. A.; de Witte, W. E. A.; Wong, Y. C.; Knasmueller, B.; Richter, L.; Kokh, D. B.; Sadiq, S. K.; Bosma, R.; Nederpelt, I.; Heitman, L. H.; Segala, E.; Amaral, M.; Guo, D.; Andres, D.; Georgi, V.; Stoddart, L. A.; Hill, S.; Cooke, R. M.; De Graaf, C.; Leurs, R.; Frech, M.; Wade, R. C.; de Lange, E. C. M.; IJzerman, A. P.; Müller-Fahrnow, A.; Ecker, G. F. Kinetics for Drug Discovery: An Industry-Driven Effort to Target Drug Residence Time. *Drug Discovery Today* **2017**, *22* (6), 896−911.

(19) Jia, Z. J.; Su, T.; Zuckett, J. F.; Wu, Y.; Goldman, E. A.; Li, W.; Zhang, P.; Clizbe, L. A.; Song, Y.; Bauer, S. M.; Huang, W.; Woolfrey, J.; Sinha, U.; Arfsten, A. E.; Hutchaleelaha, A.; Hollenbach, S. J.; Lambing, J. L.; Scarborough, R. M.; Zhu, B.-Y. N,N-Dialkylated 4-(4-Arylsulfonylpiperazine-1-Carbonyl)-Benzamidines and 4-((4-Arylsulfonyl)-2-Oxo-Piperazin-1-Ylmethyl)-Benzamidines as Potent Factor Xa Inhibitors. *Bioorg. Med. Chem. Lett.* **2004**, *14* (9), 2073−2078.

(20) Zhang, P.; Zuckett, J. F.; Woolfrey, J.; Tran, K.; Huang, B.; Wong, P.; Sinha, U.; Park, G.; Reed, A.; Malinowski, J.; Hollenbach, S.; Scarborough, R. M.; Zhu, B.-Y. Design, Synthesis, and SAR of Monobenzamidines and Aminoisoquinolines as Factor Xa Inhibitors. *Bioorg. Med. Chem. Lett.* **2002**, *12* (12), 1657−1661.

(21) Shibayama, S.; Marcou, G.; Horvath, D.; Baskin, I. I.; Funatsu, K.; Varnek, A. Application of the Mol2vec Technology to Large-size Data Visualization and Analysis. *Mol. Inf.* **2020**, *39* (6), 1900170.

(22) Alexander, J. H.; Singh, K. P. Inhibition of Factor Xa: a potential target for the development of new anticoagulants. *Am. J. Cardiovasc Drugs.* **2005**, *5* (5), 279−90.