EnsDTI-kinase: Web-server for Predicting
Kinase-Inhibitor Interactions with Ensemble
Computational Methods and Its Applications

Yijingxiu Lu\textsuperscript{a,1}, Sangsoo Lim\textsuperscript{b,1}, Sungjoon Park\textsuperscript{a}, MinGyu Choi\textsuperscript{c,d},
Changyun Cho\textsuperscript{e}, Soosung Kang\textsuperscript{f}, Sun Kim\textsuperscript{a,c,g,*}

\textsuperscript{a}Department of Computer Science and Engineering, Seoul National
University, Gwanak-ro 1, Gwanak-gu, Seoul, 08826, South Korea
\textsuperscript{b}Bioinformatics Institute, Seoul National University, Gwanak-ro 1,
Gwanak-gu, Seoul, 08826, South Korea
\textsuperscript{c}AIGENDRUG, Co., Ltd., Gwanak-ro 1, Gwanak-gu, Seoul, 08826, South Korea
\textsuperscript{d}Department of Chemistry, Seoul National University, Gwanak-ro 1,
Gwanak-gu, Seoul, 08826, South Korea
\textsuperscript{e}Interdisciplinary Program in Bioinformatics, Seoul National University, Gwanak-ro 1,
Gwanak-gu, Seoul, 08826, South Korea
\textsuperscript{f}College of Pharmacy, Ewha Womans University, 52 Ewhayeodae-gil,
Seodaemun-gu, Seoul, 03760, South Korea
\textsuperscript{g}Interdisciplinary Program in Artificial Intelligence, Seoul National
University, Gwanak-ro 1, Gwanak-gu, Seoul, 08826, South Korea

Abstract

Motivation: Kinase inhibitors are a major category of drugs. Experimental
panel assay protocols are routinely used as a standard procedure to evaluate
the efficiency and selectivity of a drug candidate to target kinase. However,
current kinase panel assays are time-consuming and expensive. In addition,
the panel assay protocols neither provide insights on binding sites nor allow
experiments on mutated sequences or newly-characterized kinases. Existing
virtual screening or docking simulation technologies require extensive computa-
tional resources, thus it is not practical to use them for the panel of kinases.
With rapid advances in machine learning and deep learning technologies, a
number of DTI tools have been developed over the years. However, these
methods are yet to achieve prediction accuracies at the level of practical use.

*Corresponding Author
\textsuperscript{1}Equal contribution
In addition, the performances of current DTI tools vary significantly depending on test sets. In this case, an ensemble model can be used to improve and stabilize DTI prediction accuracies.

**Results:** In this work, we propose an ensemble model, EnsDTI-kinase, that integrates eight existing machine learning and deep learning models into a unified model deployed as a web-server. Upon submission of a compound SMILES string, potential target kinases are automatically predicted and evaluated on the web-server. Importantly, EnsDTI-kinase is a computational platform where newly developed DTI tools can be easily incorporated without modifying core components so that its DTI prediction quality can improve over time. Besides, many useful functionalities are provided on our platform for users to further investigate predicted DTI: it allows confidence experiments by changing the amino acid (AA) at a specific position in a kinase sequence, named *in silico* mutagenesis, to investigate the effect of AA changes in binding affinity; it predicts kinase sequential regions where the query compound likely binds to by slidingly masking the sequence of selected kinases so that confidence in the predicted binding sites can be evaluated. Our model was evaluated in three experimental settings using four independent datasets, and showed accuracy of 0.82 compared to the average accuracy of 0.69 from five deep learning methods on the ChEMBL dataset. It achieved average selectivity of 0.95 within kinase families such as TK, CAMK and STE. For 8 out of 17 recent drugs, our model successfully predicted their interactions with 404 proteins at average accuracy of 0.82.

**Availability:** http://biohealth.snu.ac.kr/software/ensdti

**Contact:** sunkim.bioinfo@snu.ac.kr

**Keywords:** drug target interaction, kinase inhibitor, ensemble machine learning, small molecule drug, web server

---

1. **Introduction**

Protein kinases (PKs) play crucial roles in controlling a majority of signaling processes such as cell growth, differentiation, and apoptosis, which is often related to many human diseases [1][2][3]. Since the development of Imatinib as the first strategic small-molecule kinase inhibitor to treat chronic myelogenous leukemia (CML), small-molecule inhibitors are designed to target specific mutagenesis profiles on PKs. Examples of such inhibitors include erlotinib, sorafenib and gefitinib [4]. *In vitro* kinase assay has been widely
used as a standard protocol of experimental evaluation for both efficiency and selectivity for the set of pre-defined PKs [5, 6]. However, the assay requires a lot of cost for testing different concentrations of an inhibitor, various inhibitor candidates or different sequence compositions of kinase proteins. Thus, machine learning (ML) based in silico drug-target interaction (DTI) prediction methods naturally gained interest as one of pre-screening strategies to narrow down the candidate PK targets for experimental validation [7, 8].

Deep learning methods demonstrated a good predictive power leveraging recent data accumulation by encoding data types such as strings, images, or graphs in many application areas [9, 10, 11]. Deep learning technologies have also been applied for DTI prediction [12, 13, 14]. However, deep learning-based DTI prediction methods are yet to achieve high prediction accuracy at the level of practical use and their performance highly depends on the selection of benchmark datasets in kinase inhibitor prediction (See Section Results). In this situation, ensemble methods have been successful to improve the overall prediction performance by integrating the base classifier results [15, 11]. EnsemKRR and EnsemDT are introduced to address the data imbalance problem by Ezzat et al. [16, 17] using decision tree and kernel ridge regression as base classifiers. DrugE-Rank combines six similarity-based methods to prioritize drugs by utilizing features and similarities of drugs and targets [18]. DeepFusionDTA generates a feature map from structure and sequence information via neural network and applies bagging-based lightGBM to predict binding affinity [19].

2. Motivation and Contributions

The widely used panel assays are laborious and time-consuming. Only a limited number of kinases can be tested in a single assay. Thus, powerful and accurate computational methods can be helpful to narrow down the list of kinases to be tested extensively. However, predictions of existing ML DTI methods are not at the level of such prescreening use. One critical limitation is that each method takes single encoding schemes of compounds and proteins as input. This constrains ML model to consider data only from a single point of view, which probably makes an upper bound on prediction performance and also limited power for generalizability. In addition, there is no DTI prediction tool that provides a web interface in which users can easily perform in silico experiments only with SMILES information before synthesizing the actual compound.
In this study, we present an ensemble method for DTI prediction for kinases, EnsDTI-kinase, where anyone can perform DTI prediction for kinases with SMILES information only. EnsDTI-kinase combines five deep learning models and three machine learning techniques to provide improved predictive power. Our contributions in this study are summarized as follows. First, EnsDTI-kinase allows a flexible combination of additional models by taking prediction outcomes from each of the base classifiers. In this architecture of EnsDTI, newly developed methods can be easily included over time so that DTI prediction accuracy can be improved always. Second, EnsDTI-kinase allows confidence experiments by changing the amino acid (AA) at a specific position in a kinase sequence, named \textit{in silico} mutagenesis, to investigate the effect of AA changes in binding affinity. Third, EnsDTI-kinase predicts where the inhibitor binds to the kinase by sliding the masking sequences of selected PKs so that confidence in the predicted binding sites can be evaluated. We evaluated our model rigorously in three different performance evaluation settings to demonstrate the effectiveness of our ensemble strategy. EnsDTI-kinase is publicly available at \url{http://biohealth.snu.ac.kr/software/ensdti}.

3. Materials and Methods

3.1. Data

3.1.1. Datasets

We used four different DTI datasets: davis, kiba, kinome, and human. Below are brief descriptions of each of the datasets. The data sizes and characteristics of each dataset are summarized in Table 1.

\textit{davis} dataset [20] is a well-known benchmark dataset for binding affinity prediction of kinase inhibitors. It includes experimentally determined dissociation constant ($K_d$) values of 68 kinase inhibitors against 442 kinase proteins, resulting in a total of 30,056 DTIs.

\textit{kiba} dataset is a kinase-inhibitor bioactivity dataset [21] that integrated a total of 246,088 DTIs for 52,498 chemical compounds and 467 kinases. In this study, we used a reduced version of the kiba dataset retrieved from He et al. [22] to make sure that each target kinase has at least 10 DTIs. The filtered kiba dataset contains 118,254 DTIs, 229 kinases and 2,111 compounds.

\textit{kinome} dataset is used in MDeePred [23]. The dataset is originally collected from ChEMBL (version: 25) on selected compounds from clusters by
ECFP4 fingerprints with Tanimoto similarity over 0.7. The final dataset consists of 346 proteins, 28,860 compounds, and 48,547 DTIs.

*human* dataset was used in Tsubaki et al. [24], originally curated by Liu et al. [25]. The dataset contains 3,364 positively labeled DTIs on 1,052 compounds and 2,504 proteins. According to the authors, dissimilar proteins are less likely to be targeted by similar ligands. Based on the assumption, they screened 2,848 highly credible negative DTI pairs.

### 3.1.2. Data Pre-processing

We removed DTIs whose compounds contain any non-covalent bonds. This resulted in 64, 2111, 28860 and 1052 compounds for davis, kiba, kinome and human datasets, respectively. For datasets that provide continuous affinity values, we binarized them so that data can be used for a classification problem. Specifically, we used $pK_d$ of 7.0 and 12.1 as binarization cutoff values for davis and kiba datasets, respectively, as suggested by Tang et al. [21]. Detailed information (e.g. the positive-negative class label ratio, and the number of samples) about the datasets are in the Table 1.

### 3.2. Encoding

We used eight base classifiers (five deep learning and three machine learning models) in this study (See Section 3.3). Among these base classifiers, DL models have their own data encoding schemes while ML models lack benchmark encoding schemes. Thus, we chose data encoding schemes for the three ML models independently.

For DL models, compound information is used in SMILES format for CPI-Prediction [24], DeepDTA [26], and DeepPurpose [27], while as fingerprints for MDeePred [23] and DeepConv-DTI [28]. Protein information is used for DeepDTA, DeepConvDTI and DeepPurpose as AA sequence, for CPI-Prediction as 3-gram AA sequence, and for MDeePred as five fixed-size feature matrices. For ML models, compounds were encoded by Morgan fingerprints (256 bits) for all three ML models. We used three different protein descriptors considering structural and physico-chemical characteristics: AA composition (AAC) [29], pseudo-AA composition (PAAC) [30], and dipeptide deviation from expected mean (DDE) [31] for MLP, SVM and RF models, respectively. AAC is a 20-dimensional vector to reflect the occurrence frequency of 20 native AAs in a sequence. PAAC introduces a set of discrete numbers to approximate the effect of AA sequence order to generate an 80-dimensional vector. It combines both AA components and different sequence
Table 1: The summary of DTI databases used in this study. Detailed information on data pre-processing are described in Section 3.1.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>DTIs</th>
<th>P/N Ratio</th>
<th>Compounds</th>
<th>Kinases</th>
</tr>
</thead>
<tbody>
<tr>
<td>davis</td>
<td>28,288</td>
<td>1:11</td>
<td>64</td>
<td>442</td>
</tr>
<tr>
<td>kiba</td>
<td>118,254</td>
<td>4:1</td>
<td>2,111</td>
<td>229</td>
</tr>
<tr>
<td>kinome</td>
<td>48,547</td>
<td>1:1.6</td>
<td>28,860</td>
<td>346</td>
</tr>
<tr>
<td>human</td>
<td>6,212</td>
<td>1.2:1</td>
<td>1,052</td>
<td>2,504</td>
</tr>
</tbody>
</table>

correlation factors while maintaining the same mathematical framework as AAC. DDE is an amino acid composition-based feature vector differentiating exact epitopes from non-epitopes. DDE is known to be able to capture both the relative positional information in a protein sequence and evolutionary information when detecting DNA-binding proteins.

Both RDKit [32] and iFeature [33] libraries were used to generate the aforementioned chemical and protein representations.

3.3. Model Architecture

The basic idea of EnsDTI-kinase is that data encoding, model architecture, and final prediction modules can be independently designed and built as separate components in a pipeline so that newly developed DTI models can be easily combined to the training module and independently participate in ensemble prediction for future improvements. In EnsDTI-kinase, we integrated eight base classifiers: five DL models such as DeepDTA [26], DeepConv-DTI [28], CPI-Prediction [24], DeepPurpose [27] and MDeePred [23]; three ML models such as SVM, RF, and MLP. Our ensemble model consists of three parts: encoding, training and ensemble (Figure 1). The encoding module transforms both compound and protein data into computable formats (See Section 3.2). The training module consists of the eight base classifiers where the models are independent of each other in training the encoded data. The classifiers were trained and evaluated using the datasets prepared in this study (Details in Section 3.3.1). The ensemble module combines predictions from the training module and generates the final prediction using a series of fully connected layers. In the ensemble module, we combined the eight pairs of positive and negative probability values from the eight base classifiers to make the final prediction. Instead of using a voting scheme, we trained a fully-connected layer by applying GridSearchCV of scikit-learn package [34] to exhaustively search model parameters through a number of
Figure 1: Framework of EnsDTI-kinase: (a) Given a SMILES string of query compound and the built-in kinase panel data as input, we first (b) encode both proteins and compound into model-learnable vectors considering their structural and functional characteristics (encoding module). (c) Then, we send these vectors into our ensemble framework to predict the interaction label of each pair of compound and protein (with training and ensemble module). (d) We also design two components, in silico mutagenesis analysis and binding site exploration, to provide model explainability. (e) Both of the two components are implemented with useful functions in our web-server.
hidden neurons among \{3, 4, 5\}.

3.3.1. Base Classifiers

Deep Learning Classifiers. We chose five DL classifiers: MDeePred [23], CPI-prediction [24], DeepDTA [26], DeepPurpose [27], and DeepConv-DTI [28]. These methods generate latent patterns of protein and compound separately and combine them to make the final prediction. The output size for all the models was set to two: positive and negative binding probabilities. Detailed information on each model will be explained in the following paragraphs.

DeepDTA and MDeePred were designed for regression problems. For the classification task, we modified the activation and loss functions and model parameter settings slightly so that the model could be applied to classification task. To be specific, we change the activation function for the regression task to a sigmoid function for the binary classification prediction \(1\) and calculate loss with binary cross entropy \(2\). The output size is set to two to represent positive and negative binding probabilities.

\[
sigmoid(z) = \frac{1}{1 + e^{-z}}
\]

\[
BinaryCrossEntropy = -\sum_i^m (y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i))
\]

DeepDTA. DeepDTA [26] uses sequential information for proteins and compounds. By representing each character of SMILES strings and amino acid in one-hot encoding, DeepDTA uses two blocks of convolution neural networks (CNNs) to capture compound and protein features separately, followed by a concatenation of both representations for final prediction through an MLP layer.

DeepConv-DTI. DeepConv-DTI [28] extracts latent information separately for protein and compound. It employs CNNs on various lengths of protein subsequences and a fully connected layer (FC layer) on Morgan fingerprints to capture the local sequential patterns of DTI.
CPI-Prediction. CPI-prediction [24] learns chemical features from 2D molecular graphs and protein sequences with GCN and CNN, respectively. Protein sequences are encoded as a sequence of words that consist of 3-gram amino acid fed to 1D-CNN for latent representation. An attention mechanism is then used to calculate the relationship between each compound and these 3-gram amino acids.

DeepPurpose. DeepPurpose is an encoder-decoder framework-based model for DTI prediction [27]. It learns two embeddings for SMILES and AA sequences and feeds them into a CNN decoder to predict DTI binding affinity. Among various options they provide to build architecture, we chose message passing neural network (MPNN) to obtain the latent vector of a molecular graph and CNN to encode protein features from decomposing target sequence for binding AAs finding. Both features are sent to the decoder for DTI prediction.

MDeePred. MDeePred [23] uses protein encoding matrix and four different AA feature matrices, such as amino acid substitution matrix, physicochemical property difference matrix, distance-dependent statistical potential matrix, and residue contact energy matrix, to represent protein. A multi-channel CNNs was used to combine these 5 features. Compounds are represented as fingerprints and fed to an independent feed-forward network in parallel.

Three ML Base Classifiers. We implemented ML base classifiers using sklearn Python package. According to the parameter settings for MLP, the number of hidden layers is set to one and the number of neurons is 100. We chose ReLU as the activation function, and learning rate of 1e-3. We implemented SVM with regularization parameter equal to 1.0, with radial basis function (RBF). We implemented RF with the number of trees equal to 100, and the Gini impurity as the function to measure the quality of split.

3.4. Evaluation
To rigorously evaluate the performance of EnsDTI, we designed three different experimental settings.

Internal Cross-Validation. For each dataset in Table 1, DTI data were divided into 5-fold to build training, validation, and testing sets with a ratio of 3:1:1.
External Validation. As we have prepared four independent datasets from different sources, EnsDTI-kinase was evaluated by training an ensemble model on one of the datasets and using the model for prediction on the other independent dataset. This experiment evaluates how much EnsDTI-kinase can be generalized to previously unseen data.

Ensemble Consensus Validation. We next conducted an experiment to resample data by considering two aspects from a machine-learning perspective. First, we redefined the train and test sets by majority voting of the five DL models with threshold of 3. DTIs classified by three or more DL models are considered as training data and the rest as test sets. In other words, we can evaluate how well the classifiers trained on easily predictable samples (i.e. distant from decision boundary) can predict difficult samples (close to decision boundary). Second, as most DTI datasets contain much more negatively-labeled data points than positively-labeled ones, the five DL models were high in consensus prediction on the negative samples. We removed negatively-labeled DTIs that were correctly predicted by all models from training set to prevent the model from being negatively biased. Our intention is to remove examples that any of models can easily predict negatively. Inclusion of such easy examples guides ML models to set decision boundary between positive and negative samples is not as tight as it should be. We sampled training data class balanced. As a result, we obtained 594 and 1,912 positives and negatives in training data, while 1,699 positives and 862 negatives in test sets.

3.5. Experiments for Prediction Confidence Evaluation

To provide explainability on our model, we introduce two components below:

- **in silico point mutagenesis** - investigation of the effect of sequential variation on binding affinity by *silico* mutagenesis to specific AAs of target kinases.

- **binding site prediction** - prediction of the potential binding sites by sliding a window of sequential perturbation.

We employed the perturbation-based interpretation method 35 to build an AA sequence-wise importance map to investigate the contribution of each
Figure 2: Three validation schemes used in this study: a) Internal cross-validation: randomly split each dataset into 5 folds and use the 3 folds for training, one of the other folds for validating, and the remaining fold for testing in turn. b) External validation: given several databases from different sources, use each of them for training and validating, and each of the rest databases for testing. c) Ensemble consensus validation: for Davis dataset, we split the database into train and test datasets according to the consensus results of credible SOTA deep-learning models.

Specifically, perturbed interaction value is calculated by comparing the occlusion result and the original predicted value through the function presented in Equation 3. The measure of masked region importance $K_{ij}$ is defined as the division of the masked prediction deviation by the prediction error, where $v_i$ represents the actual binding label of protein $i$, and predicted value before and after occlusion is represented as $p_i$ and $p_{ij}$ respectively.

$$K_{ij} = \frac{|p_{ij} - v_i|}{|p_i - v_i| - \epsilon} \quad (3)$$

One major hurdle for displaying interpretation results is that majority of panel kinases have not been crystallized. Therefore, many protein binding
site prediction models showed a limited application focusing on the heatmap
generation, rather than projecting on protein structures. Fortunately, all
kinase structures could be obtained by using AlphaFold2. Google Deepmind
recently released an [AlphaFold Protein Structure Database] which contains
the human proteome and majority of manually curated UniProt entries [38,
39]. Interpretation results and motif annotation references from the kinases-
ligand interaction fingerprints and structure database (KLIFS) were both
projected on these numerically predicted kinase structures.

3.6. Web-server Implementation

EnsDTI-kinase is implemented using Django 2.0 framework and SQLite
3 database for persistent data storing and requesting. For the web page
design and web development, the open source toolkit (Bootstrap 4.6) was
utilized. EnsDTI-kinase is tested on Chrome, Firefox, Microsoft Edge, and
Safari across different operating systems. Predicted target kinase table is pre-
sented with the [jQuery javascript library DataTables] Interactive PCA and
projection figures are generated using [Chart.js] while mutagenesis influence
map and sequential importance distribution are plotted using [D3] library.

Providing a SMILES string of a query compound on the EnsDTI-kinase
platform can quickly start searching for potential target proteins on 426
kinase proteins. On the getting-start page, users can also select one protein
for residual mutagenesis significance prediction and up to ten representative
proteins for probable binding sites prediction.

4. Results

4.1. Performance Evaluation

We evaluated our model using the following schemes as mentioned in
Section 3.4 and illustrated in Figure 3.4:

- Internal Cross-Validation
- External Validation
- Ensemble Consensus Validation
Figure 3: Internal validation performances under (a) Accuracy and (b) AUC metrics on all three datasets: kinome, kiba and human.
4.1.1. Internal Cross-Validation

We first evaluate our model and the base classifiers under the 5-fold CV scheme on davis, kiba, kinome and human datasets (Figure 4.1.1 Table S1). Overall, EnsDTI-kinase achieved the best performance in terms of accuracy and AUC for the three datasets. On kiba data, models such as CPI-Prediction and MDeePred achieved low recall values (CPI-Prediction: 0.149 and MDeePred: 0.0), while EnsDTI-kinase achieved the best recall value (0.728). We note here that EnsDTI-kinase made balanced predictions on positive and negative labels, with the best F1-score (0.773), while the second-best performing model (DeepPurpose) achieved 0.531. AUPR of our model is the best at 0.840 while that of the runner-up model is 0.808 by DeepDTA. This means that our model is superior to other base classifiers in identifying true kinase inhibitor targets.

4.1.2. External Validation

We performed the external validation scheme to compare the generalization power of the models on independent data sources. We obtained the results in Table 2 that EnsDTI-kinase has more reliable predictive power for independent data sources when compared with the DL base classifiers. External validation performances of models trained on human dataset are generally poor, which is likely due to its small data size of 6,212 DTIs. Overall, EnsDTI-kinase showed stable prediction power and best-achieved performance on the external validation experiments.

Table 2: External validation performance of EnsDTI-kinase and base classifiers under balanced accuracy.

<table>
<thead>
<tr>
<th>Test Data</th>
<th>Train Data</th>
<th>EnsDTI-kinase</th>
<th>DeepDTA</th>
<th>DeepConv-DTI</th>
<th>CPI-Prediction</th>
<th>MDeePred</th>
<th>DeepPurpose</th>
</tr>
</thead>
<tbody>
<tr>
<td>davis</td>
<td>human</td>
<td>0.489</td>
<td>0.219</td>
<td>0.301</td>
<td>0.094</td>
<td>0.481</td>
<td>0.451</td>
</tr>
<tr>
<td>kiba</td>
<td>human</td>
<td>0.895</td>
<td>0.198</td>
<td>0.199</td>
<td>0.165</td>
<td>0.215</td>
<td>0.193</td>
</tr>
<tr>
<td>kinome</td>
<td>human</td>
<td>0.930</td>
<td>0.579</td>
<td>0.083</td>
<td>0.878</td>
<td>0.711</td>
<td>0.866</td>
</tr>
<tr>
<td>human</td>
<td>kiba</td>
<td>0.761</td>
<td>0.462</td>
<td>0.493</td>
<td>0.536</td>
<td>0.521</td>
<td>0.505</td>
</tr>
<tr>
<td>davis</td>
<td>kiba</td>
<td>0.841</td>
<td>0.219</td>
<td>0.211</td>
<td>0.208</td>
<td>0.210</td>
<td>0.225</td>
</tr>
<tr>
<td>kiba</td>
<td>human</td>
<td>0.805</td>
<td>0.708</td>
<td>0.653</td>
<td>0.763</td>
<td>0.506</td>
<td>0.539</td>
</tr>
<tr>
<td>kinome</td>
<td>human</td>
<td>0.865</td>
<td>0.387</td>
<td>0.753</td>
<td>0.294</td>
<td>0.351</td>
<td>0.285</td>
</tr>
<tr>
<td>kiba</td>
<td>kinome</td>
<td>0.713</td>
<td>0.389</td>
<td>0.624</td>
<td>0.607</td>
<td>0.607</td>
<td>0.611</td>
</tr>
<tr>
<td>kinome</td>
<td>human</td>
<td>0.597</td>
<td>0.606</td>
<td>0.472</td>
<td>0.417</td>
<td>0.499</td>
<td>0.503</td>
</tr>
<tr>
<td>kiba</td>
<td>kinome</td>
<td>0.604</td>
<td>0.403</td>
<td>0.405</td>
<td>0.419</td>
<td>0.408</td>
<td>0.417</td>
</tr>
</tbody>
</table>
Table 3: Performance comparison of EnsDTI-kinase with the five DL classifiers on the re-sampled davis dataset. Resampling of the original davis dataset produces highly confident labels as training, while the others as test sets.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>EnsDTI-kinase</td>
<td>0.706</td>
<td>0.791</td>
<td>0.607</td>
<td>0.756</td>
<td>0.682</td>
<td>0.753</td>
<td>0.856</td>
<td>0.773</td>
</tr>
<tr>
<td>CPI-Prediction</td>
<td>0.429</td>
<td>0.732</td>
<td>0.842</td>
<td>0.219</td>
<td>0.531</td>
<td>0.556</td>
<td>0.718</td>
<td>0.337</td>
</tr>
<tr>
<td>DeepDTA</td>
<td>0.404</td>
<td>0.765</td>
<td>0.912</td>
<td>0.146</td>
<td>0.529</td>
<td>0.533</td>
<td>0.704</td>
<td>0.245</td>
</tr>
<tr>
<td>DeepConv-DTI</td>
<td>0.454</td>
<td>0.791</td>
<td>0.875</td>
<td>0.241</td>
<td>0.558</td>
<td>0.569</td>
<td>0.729</td>
<td>0.369</td>
</tr>
<tr>
<td>MDeePred</td>
<td>0.342</td>
<td>0.710</td>
<td>0.990</td>
<td>0.013</td>
<td>0.501</td>
<td>0.533</td>
<td>0.684</td>
<td>0.025</td>
</tr>
<tr>
<td>DeepPurpose</td>
<td>0.538</td>
<td>0.813</td>
<td>0.821</td>
<td>0.394</td>
<td>0.608</td>
<td>0.570</td>
<td>0.765</td>
<td>0.531</td>
</tr>
<tr>
<td>SVM</td>
<td>0.342</td>
<td>0.607</td>
<td>0.022</td>
<td>0.497</td>
<td>0.042</td>
<td>0.42</td>
<td>0.497</td>
<td>0.662</td>
</tr>
<tr>
<td>RF</td>
<td>0.408</td>
<td>0.670</td>
<td>0.211</td>
<td>0.795</td>
<td>0.503</td>
<td>0.321</td>
<td>0.503</td>
<td>0.665</td>
</tr>
<tr>
<td>MLP</td>
<td>0.376</td>
<td>0.748</td>
<td>0.089</td>
<td>0.941</td>
<td>0.515</td>
<td>0.159</td>
<td>0.515</td>
<td>0.671</td>
</tr>
</tbody>
</table>

4.1.3. Ensemble Consensus Validation

The performance of EnsDTI-kinase and DL models were compared on the same re-defined dataset (Table 3). By re-sampling the davis data, we obtained 594 and 1,912 positives and negatives in training data, while 1,699 positives and 862 negatives in test set (See Section 3.4). We observed that EnsDTI-kinase achieved better performances compared to the other state-of-the-art (SOTA) models. EnsDTI-kinase achieved 0.706 in accuracy, while the best-performed base classifier (DeepPurpose) achieved 0.538. It is also notable that combining the results from individual classifiers significantly improved recall. EnsDTI-kinase achieved 0.756 in recall, while the next best-performing model (DeepPurpose) achieved 0.394. Overall, EnsDTI-kinase produces much less false positive DTIs than the five DL models do.

4.2. Ablation Analysis

EnsDTI-kinase combines predictions from eight models, thus the ablation analysis is on how many models contribute to the final prediction. Adding up each base classifier in chronological order for the ablation analysis was performed to investigate how much gain in performance our ensemble model achieved when integrating more base classifiers. Integrating more DL models as an ensemble showed balanced or better performance in different metrics. Moreover, the set of three ML models also provided 10.8% of gain on average AUC, 6.7% of gain on average balanced accuracy, and 6.0% of gain on average AUPR to the ensemble of DL models. Overall, EnsDTI-kinase was able to make improved predictions by avoiding trends that predict all data to the same label.
4.3. Enrichment of Specific Data Labels

To further demonstrate our model, we investigated how much our model can better predict DTIs to specific kinase families or chemical features. To do so, PubChem fingerprints and EC number information were used to investigate the enrichment of correct DTI prediction for compounds and kinases, respectively. We found from the results that there were two kinase families (receptor and non-specific protein-tyrosine kinases) and 39 PubChem fingerprints that our model showed statistically significantly more predictive power than accuracy of 0.706 (Table 5 and Table 6).

4.4. Case Studies

In this section, we demonstrate our results using case studies of target prediction on kinase panel, in silico mutagenesis and binding site prediction. How to use EnsDTI-kinase web-server is also demonstrated by using Imatinib as a query compound.

4.4.1. Target Prediction on Kinase Panel

Figure 4.4A shows ‘Target Prediction’, ‘Result Projection’, and ‘PCA Exploration’ sub-pages of EnsDTI-kinase website. ‘Target Prediction’ sub-page includes a series of bar charts (left) and a summary table (right). Each bar chart represents the predicted binding probability of a kinase group, and the summary table containing detailed information is user-interactive and supports functions like filtering, sorting, searching and downloading.

‘Result Projection’ sub-page presents two types of result visualization, compound cluster plot (left) and odds distribution plot (right), so that users can compare results with training set. Compound cluster plot shows the prediction distribution of training compound clusters and marks the belonging of query compound. Specifically, by calculating the ECFP similarity, we clustered training compounds into eight groups and mapped the query compound to the closest one. From the plot we can investigate to what extent compound structures affect prediction results. Odds distribution plot illustrates the log-odds distribution of both training data and user query. Log-odds are calculated with Equation 4. The higher the log odds, the more selective the query compound is to the specific target. ‘PCA Exploration’ sub-page visualizes model latent vectors with PCA figures. These latent vectors are obtained from the final layer of EnsDTI-kinase and implies how model makes decision on training and query data. PCA figures are colored in three schemes (interaction label, compound ATC code, and kinase EC...
Table 4: Ablation study performance of EnsDTI-kinase.

<table>
<thead>
<tr>
<th>Model</th>
<th>ML Models</th>
<th>DL Models</th>
<th>Metric</th>
<th>Accuracy</th>
<th>Precision</th>
<th>Recall</th>
<th>Specificity</th>
<th>Balanced Accuracy</th>
<th>AUC</th>
<th>AUPR</th>
</tr>
</thead>
<tbody>
<tr>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>(A, B) + C + D + E</td>
<td>0.455</td>
<td>0.700</td>
<td>0.311</td>
<td>0.737</td>
<td>0.524</td>
<td>0.479</td>
<td>0.680</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>0.609</td>
<td>0.651</td>
<td>0.883</td>
<td>0.068</td>
<td>0.476</td>
<td>0.426</td>
<td>0.606</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>0.660</td>
<td>0.663</td>
<td>0.991</td>
<td>0.009</td>
<td>0.500</td>
<td>0.565</td>
<td>0.708</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>0.651</td>
<td>0.662</td>
<td><strong>0.968</strong></td>
<td>0.027</td>
<td>0.497</td>
<td>0.480</td>
<td>0.660</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>0.546</td>
<td><strong>0.825</strong></td>
<td>0.401</td>
<td><strong>0.832</strong></td>
<td>0.617</td>
<td>0.661</td>
<td>0.799</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>0.639</td>
<td>0.763</td>
<td>0.662</td>
<td>0.594</td>
<td>0.628</td>
<td>0.696</td>
<td>0.824</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>0.675</td>
<td>0.729</td>
<td>0.813</td>
<td>0.404</td>
<td>0.608</td>
<td>0.685</td>
<td>0.821</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td><strong>0.706</strong></td>
<td>0.791</td>
<td>0.756</td>
<td>0.607</td>
<td><strong>0.681</strong></td>
<td><strong>0.753</strong></td>
<td><strong>0.856</strong></td>
</tr>
</tbody>
</table>

A: CPI-Prediction, B: DeepDTA, C: DeepConv-DTI, D: MDeePred, E: DeepPurpose

Table 5: EC Numbers that show greater performance than overall EnsDTI-kinase accuracy (0.706).

<table>
<thead>
<tr>
<th>EC Number</th>
<th>Description</th>
<th>Accuracy</th>
<th>f1-score</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.7.10.1</td>
<td>receptor protein-tyrosine kinase</td>
<td>0.774</td>
<td>0.858</td>
<td>5.18e-10</td>
</tr>
<tr>
<td>2.7.10.2</td>
<td>non-specific protein-tyrosine kinase</td>
<td>0.729</td>
<td>0.806</td>
<td>1.29e-2</td>
</tr>
</tbody>
</table>

Table 6: Top 10 PubChem fingerprints that show greater performance than overall EnsDTI-kinase accuracy (0.706).

<table>
<thead>
<tr>
<th>PubChem FP ID</th>
<th>Description</th>
<th>Accuracy</th>
<th>f1-score</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>397</td>
<td>N(~C)(:C):(C)</td>
<td>0.767</td>
<td>0.772</td>
<td>2.18e-5</td>
</tr>
<tr>
<td>662</td>
<td>O-C-C-O-C</td>
<td>0.765</td>
<td>0.754</td>
<td>4.21e-6</td>
</tr>
<tr>
<td>567</td>
<td>O-C-C-O</td>
<td>0.764</td>
<td>0.754</td>
<td>5.95e-6</td>
</tr>
<tr>
<td>213</td>
<td>&gt;= 1 any ring size 7</td>
<td>0.762</td>
<td>0.738</td>
<td>3.79e-5</td>
</tr>
<tr>
<td>218</td>
<td>&gt;= 1 unsaturated non-aromatic nitrogen-containing ring size 7</td>
<td>0.762</td>
<td>0.738</td>
<td>3.79e-5</td>
</tr>
<tr>
<td>219</td>
<td>&gt;= 1 unsaturated non-aromatic heteroatom-containing ring size 7</td>
<td>0.762</td>
<td>0.738</td>
<td>3.79e-5</td>
</tr>
<tr>
<td>737</td>
<td>Cc1cc(N)c1c1</td>
<td>0.760</td>
<td>0.746</td>
<td>3.95e-5</td>
</tr>
<tr>
<td>800</td>
<td>CC1CC(N)CCC1</td>
<td>0.760</td>
<td>0.746</td>
<td>3.95e-5</td>
</tr>
<tr>
<td>157</td>
<td>&gt;= 3 any ring size 5</td>
<td>0.760</td>
<td>0.751</td>
<td>3.73e-5</td>
</tr>
<tr>
<td>612</td>
<td>N-C-O-C-C</td>
<td>0.759</td>
<td>0.741</td>
<td>1.04e-4</td>
</tr>
<tr>
<td>EnsDTI-kinase</td>
<td></td>
<td><strong>0.706</strong></td>
<td><strong>0.773</strong></td>
<td>-</td>
</tr>
</tbody>
</table>
Figure 4: EnsDTI-kinase web-server results on case study: 

a) **In-silico** prediction on query compounds and its interaction probabilities with kinases. 

b) Exploration of query compound and predicted embedding space given known interaction relationships. 

c) **In-silico** interpretation of mutation’s influence on binding and binding site prediction.
number) so that we can infer how well the model distinguish positive and negative data, and whether different chemical characteristics and corresponding enzyme-catalyzed reaction features are captured by the model. All three PCA plots are explorable by clicking the points on the figure.

\[
\text{Log Odds Ratio} = \log\left(\frac{p_{\text{binding}}}{1 - p_{\text{binding}}}\right)
\]

(4)

4.4.2. In-silico Mutagenesis Analysis

For the kinase and specific amino acid position of interest submitted by user, EnsDTI-kinase replaces the corresponding amino acid by applying all possible amino acid residues \((\text{in silico} \text{ point mutagenesis})\) and presented the results in the ‘Mutagenesis Analysis’ sub-page as shown in Figure 4.4B. The degree of perturbation in binding affinity is measured by deviations from that of the original amino acid sequence. Take the binding between Imatinib and kinase ABL1 as example, the binding affinity for ABL1(M351T) is predicted as 0.717 while the binding prediction between original ABL1 and Imatinib is 0.710. The results are in agreement with the experimental results from Apperley [40] that Imatinib is 5% – 10% much more sensitive to ABL1(M351T) than ABL1.

4.4.3. Binding Site Prediction

Figure 5: Binding site prediction heatmap for Imatinib to selected kinases: ABL(h), AURK(h), PIK3CA(p110α&p85α)(h), and STK10(h). Yellow-colored regions indicate highly-likely binding sites, while blue-colored regions are less likely to be binding sites.

We conducted case studies on Imatinib and four selected kinases: ABL(h), AURK(h), PIK3CA(p110α&p85α)(h) and STK10(h). The predicted binding sites are shown in Figure 5. For each amino acid sequence, top 10% of loci that are important to inhibition are marked in yellow in the heatmap.
These include L300-Q306, K904-L910, and A999-G1003 for ABL(h), G10-A14, Y196-A202, A330-K338 for AURK(h), L698, R764, T1030-F1038 for PIK3CA(p110a&p85a)(h), I92-D102, S251-P253 for STK10(h). Binding site prediction results are shown in the ‘Pocket Prediction’ sub-page on the web-server.

Our web-server also provides user-interactive 3D visualization for comparing interpretation results with KLIFS kinase motif annotation [41, 42]. Referring to the KLIFS annotation, our web-server also displays important protein residues with different colors.

Figure 6: 3D Visualization of pocket prediction result on KIT and Imatinib (PDB ID: 1T46). Red: Significant substructure of KIT predicted by EnsDTI-kinase. Gray: The rest of the KIT. Yellow: Imatinib.

Users can be informed about inhibitor types of their compounds whether they bind directly near hinge or allosteric sites. Figure 6 is the 3D visualization of the case study on KIT and Imatinib (PDB ID: 1T46). Residues that are predicted possibly for the pocket by the model are colored in red, while others are colored in gray. Imatinib is highlighted in yellow. We can see that EnsDTI-kinase gave higher weight to the locations near where binding occurs.

4.4.4. Analysis on Recent Drugs

In order to demonstrate our model’s performance in recent experiments, EnsDTI-kinase was evaluated on experimental results obtained from recent publications. Elkins et al. [43] tested 367 lead compounds on 224 kinases.
We filtered out compounds and proteins whose sequences are already used for training, and tested the remaining DTI pairs for evaluation.

On the high-throughput profiling of available kinase inhibitors against a panel of 300 recombinant protein kinases from [5], we applied activity threshold of $pK_d = 7$ and acquired 17 cold drugs and 9,798 interactions with 404 protein kinases. Our model predicted these interactions with an overall accuracy of 0.69. On close examination of the best-predicted 8 out of 17 recent drugs, our model successfully predicted their interactions with kinases at accuracy of 0.82. We provide the complete prediction performance for 17 approved inhibitors in Supplement table S2.

5. Conclusion

We developed an ensemble model, EnsDTI-kinase, that integrates three machine learning and five deep learning-based methods in a single computational framework, to predict kinase-inhibitor interactions. In extensive experiments, we notice that existing DTI models tend to produce high false negative rates and have difficulty in training on databases that are imbalanced in label distribution. Significance of EnsDTI-kinase is to reduce false negatives of kinase targets while improving the overall prediction power. More importantly, EnsDTI-kinase is a computational platform where newly developed DTI tools can be easily incorporated without modifying component tools so that the DTI prediction quality can improve over time. In addition, EnsDTI-kinase provides useful functionalities for users to further investigate predicted DTI so that confidence in predicted kinase targets can be evaluated on the web. In future work, we will continue to improve EnsDTI-kinase by updating the base classifiers with state-of-the-art methods and developing more useful functions for accurate DTI prediction. In particular, we plan to incorporate protein structure information by Alphafold2 [38] and pocket information such as DUD-E [44] for kinase target prediction.

6. Acknowledgements

This research was supported by the Collaborative Genome Program for Fostering New Post-Genome Industry of the National Research Foundation (NRF) funded by the Ministry of Science and ICT (MSIT) (No. NRF-2014M3C9A3063541), by the Bio & Medical Technology Development Program of the National Research Foundation (NRF) & funded by the Korean...
government (MSIT) (No. 2022M3E5F3085677), by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) [NO. 2021-0-01343, Artificial Intelligence Graduate School Program (Seoul National University)], and by a grant (No. DY0002258224) from Ministry of Food and Drug Safety in 2020.
References


[29] X. Chen, J.-D. Qiu, S.-P. Shi, S.-B. Suo, S.-Y. Huang, R.-P. Liang, Incorporating key position and amino acid residue features to identify


Supplementary Information

Table S1: Five-fold CV performance of EnsDTI-kinase and the eight base classifiers on the three data sets: human, kiba, and kinome.

Table S2 shows the internal validation performances of EnsDTI-kinase and all base classifiers.

Figure S1 shows the external validation performances of EnsDTI-kinase and five DL-based base classifiers. Our model shows the best and the most stable performances under different metrics (e.g. accuracy, balanced accuracy, F1, AUC).

Table S3 shows the prediction performance on recent 17 approved inhibitors. To evaluate our model on real drug data, we filtered out the compounds and proteins whose sequences are already used in the training process, and test the remaining DTI pairs from Elkins et al. [13].

Table S3 shows the confusion matrix of ensemble consensus validation experiment. Not only well predict most of the binding interactions, our model also presents the least false negative number which shows its ability to rescue potential targets for further experimental procedures.
Figure S1: Comparison of external validation performance
Table S2: Prediction performance of EnsDTI-kinase on 17 approved kinase inhibitors (CID: 10288191, 11315474, 11511120, 16063245, 22024915, 44480399, 46220502, 50905713, 51039030, 51039094, 58557659, 66599893, 67462786, 71584930, 91667513, 118023034, 134436906).

<table>
<thead>
<tr>
<th>Inhibitor CID</th>
<th>Acc</th>
<th>TP</th>
<th>TN</th>
<th>FP</th>
<th>FN</th>
</tr>
</thead>
<tbody>
<tr>
<td>10288191</td>
<td>0.991</td>
<td>0</td>
<td>422</td>
<td>0</td>
<td>4.0</td>
</tr>
<tr>
<td>11315474</td>
<td>0.751</td>
<td>25</td>
<td>295</td>
<td>29</td>
<td>77</td>
</tr>
<tr>
<td>11511120</td>
<td>0.761</td>
<td>0</td>
<td>324</td>
<td>0</td>
<td>102</td>
</tr>
<tr>
<td>16063245</td>
<td>0.714</td>
<td>2</td>
<td>302</td>
<td>22</td>
<td>100</td>
</tr>
<tr>
<td>22024915</td>
<td>0.681</td>
<td>5</td>
<td>285</td>
<td>39</td>
<td>97</td>
</tr>
<tr>
<td>44480399</td>
<td>0.761</td>
<td>0</td>
<td>324</td>
<td>0</td>
<td>102</td>
</tr>
<tr>
<td>46220502</td>
<td>0.953</td>
<td>0</td>
<td>406</td>
<td>0</td>
<td>20</td>
</tr>
<tr>
<td>50905713</td>
<td>0.934</td>
<td>0</td>
<td>398</td>
<td>11</td>
<td>17</td>
</tr>
<tr>
<td>51039030</td>
<td>0.723</td>
<td>16</td>
<td>292</td>
<td>32</td>
<td>86</td>
</tr>
<tr>
<td>51039094</td>
<td>0.599</td>
<td>35</td>
<td>220</td>
<td>104</td>
<td>67</td>
</tr>
<tr>
<td>58557659</td>
<td>0.669</td>
<td>25</td>
<td>260</td>
<td>64</td>
<td>77</td>
</tr>
<tr>
<td>66599893</td>
<td>0.993</td>
<td>0</td>
<td>423</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>67462786</td>
<td>0.627</td>
<td>40</td>
<td>227</td>
<td>97</td>
<td>62</td>
</tr>
<tr>
<td>71584930</td>
<td>0.671</td>
<td>5</td>
<td>281</td>
<td>43</td>
<td>97</td>
</tr>
<tr>
<td>91667513</td>
<td>0.761</td>
<td>0</td>
<td>324</td>
<td>0</td>
<td>102</td>
</tr>
<tr>
<td>118023034</td>
<td>0.709</td>
<td>3</td>
<td>299</td>
<td>25</td>
<td>99</td>
</tr>
<tr>
<td>134436906</td>
<td>0.737</td>
<td>2</td>
<td>312</td>
<td>12</td>
<td>100</td>
</tr>
</tbody>
</table>

Table S3: Confusion matrix of ensemble consensus validation experiment.

<table>
<thead>
<tr>
<th>Model</th>
<th>TP</th>
<th>TN</th>
<th>FP</th>
<th>FN</th>
</tr>
</thead>
<tbody>
<tr>
<td>EnsDTI-kinase</td>
<td>1285</td>
<td>523</td>
<td>339</td>
<td>414</td>
</tr>
<tr>
<td>CPI-Prediction</td>
<td>372</td>
<td>726</td>
<td>136</td>
<td>1327</td>
</tr>
<tr>
<td>DeepDTA</td>
<td>248</td>
<td>786</td>
<td>76</td>
<td>1451</td>
</tr>
<tr>
<td>DeepConvDTI</td>
<td>409</td>
<td>754</td>
<td>108</td>
<td>1290</td>
</tr>
<tr>
<td>MDeePred</td>
<td>22</td>
<td>853</td>
<td>9</td>
<td>1677</td>
</tr>
<tr>
<td>DeepPurpose</td>
<td>669</td>
<td>708</td>
<td>154</td>
<td>1030</td>
</tr>
<tr>
<td>SVM</td>
<td>55</td>
<td>856</td>
<td>6</td>
<td>1644</td>
</tr>
<tr>
<td>RF</td>
<td>481</td>
<td>807</td>
<td>55</td>
<td>1218</td>
</tr>
<tr>
<td>MLP</td>
<td>158</td>
<td>818</td>
<td>44</td>
<td>1541</td>
</tr>
</tbody>
</table>