Random-Forest Model for Drug Target-Interaction Prediction via Kullbeck–Leibler Divergence

Sangjin Ahn
Department of Financial Engineering, College of Business, Ajou University
Sieun Lee
Gachon Institute of Pharmaceutical Science and Department of Pharmacy, College of Pharmacy, Gachon University
Mi-hyun Kim (kmh0515@gachon.ac.kr)
Gachon Institute of Pharmaceutical Science and Department of Pharmacy, College of Pharmacy, Gachon University

Research Article

Keywords: Chemocentric, 3D Molecular Fingerprint, 3D Similarity, Drug Target Interaction Feature, Nonparametric Density Estimation, Kullbeck–Leibler Divergence, Machine Learning

Posted Date: February 18th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1357588/v1

Abstract

Virtual screening has significantly improved the rate of success of early drug discovery. Recent virtual screening methods have improved along with advances in machine learning (ML) and chemical information. Among these advances, the creative extraction of drug features is important for predicting drug-target interaction (DTI), a large-scale virtual screening of known drugs. Herein, we report on Kullbeck–Leibler divergence (KLD) as a DTI feature and the feature driven classification model applicable to DTI prediction. Thus, E3FP 3D molecular fingerprints of drugs as molecular representation allowed the computation of 3D-similarities between ligands within each target (Q-Q matrix) for uniqueness of pharmacological targets and between query–ligand (Q-L vector) for DTI. The 3D similarity matrices were transformed into probability density functions using kernel density estimation (KDE) as a nonparametric estimation. Each density model could exploit the characteristics of each pharmacological target and measure the quasi-distance between ligands. Furthermore, we developed a random-forest (RF) model from the KLD feature vectors to successfully predict DTI, irrespective of whether each query belongs to a target.

1. Introduction

Several ML-based methods have been widely applied to chemo-informatic-related areas. From the classical Bayesian approach to recent deep-learning (DL) technology, the description of the molecule itself and the derivation of a novel representation of molecules play a vital role in computer-aided drug discovery [1–4]. Studies on drug-target interactions (DTI) and several schemes and algorithms have been conducted using molecular descriptors and similarity scores [5]. Despite the availability of 3D descriptors such as E3FP [6], 3D molecular similarity rarely applied into DTI prediction. Previous studies using both molecular descriptors and similarity scores applied similarity scores into statistical rules [7–10]. These schemes for DTI prediction have not focused on modeling the heterogeneity of similarity scores using 3D similarity distribution [11–15]. This study uses probability density model of 3D similarity vectors to obtain a reliable DTI predictive model. In particular, a nonparametric density model is added to our previous Kullbeck–Leibler divergence (KLD) based quantifying method [16], which relatively observe ligands in the view of candidate targets, so that the multiple KLD measurements were used to describe a drug (query).

[INSERT FIGURE1 HERE]

Feature engineering is an essence of ML-based drug discovery. Recently, ML based DTI detection (descriptive/predictive) and ML-aided drug discovery studies mutually have made an optimistic effect on the feature engineering for molecular data. The performance of such ML approaches relied on molecular representation. One need of these ML approaches is perfect transferability of molecular information during molecular representation, similarity scoring, and learning. For this purpose, herein we try to link our 3D similarity-based quantitative method [16] and ML algorithm to predict whether each query belongs to a candidate target. This study also uses on chemo-centric assumptions along with 3D similarity [16] with the sequential works. First, based on E3FP, 3D-radial molecular fingerprints, pairwise similarities are calculated between ligands within each target (Q-Q matrix) and between query–ligand (Q-L vector) for DTI. Second, the 3D-similarity vectors (Q-L) and matrices (Q-Q) are probabilistically modeled to describe uniqueness of targets (Q-Q) and to quantify molecular-specific information for DTI. Finally, KL-divergence (KLD) works as a ‘quasi-distance’ among the density models and KLD as a novel DTI feature vector successfully is extended to DTI model (Fig. 1).

2. Methods And Materials

2.1 Dataset and data preparation.

We collected biological activity data from the CHEMBL 26 database, which is publicly available. [17] The database consists of more than 200 single-protein targets and their chemical and genomic properties. In this work, we used 17 targets selected by the (benchmark) paper. [18] The downloaded information table contained a list of smiles from CHEMBL26 databases, which describe ‘Molecule name’, ‘SMILES’, ‘IC50’ value for each listed CHEMBL ID. The duplicated item was removed to avoid sampling bias. After removal of the smiles, the conformation failed. We focused on 33,730 ligands and 4,693 K conformers. The overall data handling and algorithm computing was conducted using Python with accompanied its modules. 3D conformers were generated under conditions reported by Openeye Omega. [19–20]

2.2 Three-dimensional fingerprinting and ligand pairwise 3D molecular similarity.

All original ligand spaces from 17 targets were randomly re-sampled using 15,000 ligands. The size of each molecular conformers was limited to 15,000 ligands. Thus, we resolve the dimensionality and data imbalance problems. Ten back-tests were performed to determine the stability of random sampling. We confirm that changes in 'random seed' rarely provide stability of the similarity score density structure. Among the enormous descriptors for molecular representation, E3FP was chosen to effectively describe the 3D structure of the molecules. Each 3D fingerprint depicting a conformer of a ligand was encoded using E3FP in the RDKit library. In other words, the E3FP generated 3D molecular fingerprints in the RDKit library, and each 3D conformer was converted to the RDKit format, so that the similarity scores among ligands could be calculated. The 3D-coordinate of each conformer expressed in the sdf format was converted and encoded to a sequence of bit vectors consisting of 1024 '0' and '1' digits. Subsequently, the similarity score was calculated by comparing them. This bit vector-based similarity score calculation was computationally less expensive than the maximum common substructure (MCS)-based approach or shape-based approach (Openeye Shape Toolkit) was, and it retained the 3D conformation. [21–22]

2.3 Query-query matrix

The query–query matrix contains the pairwise similarity score of all ligands that belong to a candidate target. Their dimensions are up to 15,000 × 15,000. Let M_{Q-Q} be the similarity matrix obtained from 17 independent targets, then its elements $a_{1,1}, \ldots, a_{15000,15000}$ are a set of pairwise similarity scores of ligands that belong to a certain target. Note that these query–query matrices potentially work as a benchmark for measuring target-specific (collective, global) information. The descriptive statistics (density information) of Q-Q elements are expected to differ among targets. However, it preserves the stability of the sampling of ligands.

$$M_{Q-Q} = \begin{bmatrix} a_{1,1} & a_{1,2} \cdots a_{15000} \\ \cdots & \\ a_{15000,1} & \end{bmatrix}$$

2.4 Query-ligand vector

Next, we prepare the query–ligand similarity vector to express and measure the interaction with the candidate target from a (certain) query. These vectors preserve ligand-specific information whose descriptive statistics differ based on the ligand, where each of their dimensions is 1 × 15,000(max). It can be obtained from each column vector of the pairwise similarity matrix among the 17 targets. These can be referred to as a query's 'observation' in terms of the candidate target's view. Each query–ligand vector is comparable to each Q-Q matrix because they share a common ligand.

$$M_{Q-L} = \begin{bmatrix} b_1 \\ b_2 \\ \cdots \\ b_{15000,1} \end{bmatrix}$$

[INSERT FIGURE2 HERE]

2.5 Probability density function of each vector space

Our experiment considers probabilistic information reflecting the target representation and the ligand to target interaction. Generally, the shapes of the Q-Q and Q-L matrix, whose number of ligands depends on the target, are all different. One way to unify and structuralize their information is to use their probability density functions. We determine the distributions of both the Q-Q matrix and Q-L vectors. Each Q-Q matrix density function projects unique representations of each target; specifically, the shape of the tail part, symmetry, bias, and sharpness differ between targets. Similarly, the Q-L vector density reflects the information obtained from the query (ligand)-target interaction. Each probability density function is represented by the function value y = p(x), q(x) for each x-axis point that divides the interval [0, 1] into 100 equal parts. After being combined with the information metric, these probability distributions, p(x) and q(x) are the main components that constitute the feature vector of our classification model.

2.6 Kernel density estimation

KDE as a well-known nonparametric density estimation was chosen to estimate probability density function. [23–26] The process of estimating the probability distribution function for the given data involves maximizing the built-in likelihood function. When the KDE obtains probability distributions, the probability for the points on each x-axis can be obtained as follows:

$$\hat{p}(x) \propto \sum_{i=1}^{100} kernel\left(\frac{x - x_i}{(Bandwidth)}\right)$$

In the process of KDE, when the input matrix and vector are constructed, the estimation is performed using a Python script, and a 1 × 100 vector containing each pdf value is output. The Scipy's Python package [27], which automatically selects the optimal bandwidth of KDE, allows us to apply the Gaussian kernel. We confirm that the density structure in this study rarely depends on the KDE methodology and bandwidth. Both Silverman and Scott's methods yielded satisfactory results. Moreover, such a nonparametric approach not only provides flexible and stable results regardless of the experimental environment but also provides results to researchers with fewer estimation errors in density estimation. The KLD was used to calculate the “difference” between estimated density functions from each Q-Q matrix and Q-L vector. The divergence between Q-Q and Q-L density models is interpreted as a measurement of a query’s interaction with a target.

2.7 Kullbeck–Leibler divergence

KLD is a relative entropy that allows researchers to measure whether the two probability density functions are different or equal. [28–30] Lower KLD values imply higher similarity between two density functions and vice versa. The quantity served as a metric to measure a query among several possible targets. The feasibility of a given query belonging to a target was determined by comparing KLD values of the mapped Gaussian mixture model [16]. KLD can be simply computed from the pdf on [0, 1], which is evaluated by their function values. Let q(x) be the 17 Q-Q densities postulated to be fixed to describe the representative characteristics of certain targets. Our observation of a ligand toward a candidate target, p(x), was obtained from the Q-L vector. In our context, KLD is used to measure the degree to which the Q-L vector density (p(x)) differs from the 17 Q-Q matrix (i.e., candidate target’s density, q(x)). The divergence between the query and query similarity density function and query–ligand density measures the magnitude of the difference between a query and a candidate target and illustrates how the KLD is calculated. To calculate KLD directly, considering the point where q(x) is zero, it was calculated temporarily by adding a small number to the functional value on both p(x) and q(x). Let P = { p_1, …, p_17}, q = {q1, …, Q_k}. The KLD is calculated as follows:

$$\mathrm{KL(p|q)} = \sum_x p(x)\ ln\frac{p(x)}{q(x)} * dx + \sum_x \big(q(x) - p(x)\big) * dx$$

Seventeen KLD values were obtained along the ligands (query) from the candidate targets. The divergence of q from p approximate to a minimum if the Q-L density is similar to a Q-Q density so that the value provides a measurement of the distance between a query and several candidate targets for a given query. Generally, a small KLD value suggests that any query has a high similarity to a target, which corresponds to whether a certain query belongs to a target vice versa. In the RF model, individual KLD values became a feature of each query that describes the measurement from the viewpoint of candidate target. Finally, we obtain a labeled vector of 1 × 17 divergence from each ligand (query) for the RF classifier.

2.8 Random forest classifier

RF models, an ensemble of several decision tree models, were chosen for this nonparametric methodology [31–32]. The famous classification and regression tree (CART) algorithm can be easily extended to large-scale data. [33–37] We consider a 17 × 1-sized feature for each ligand. Each query is labeled with its target number (1, 2, … 17), and each decision process of the RF algorithm is generated by comparing each query’s KLD value. Given the 17 × 1 feature vector, which consists of KLD values, the target prediction is made by combining the decisions from individual features. In other words, by measuring the KLD values, an RF classifier is instructed to determine whether a query is suitable for a target. The RF classifier implicitly facilitates correspondence in the value of KLD between such similarity density and an indirect difference between a query and a candidate target. The optimal parameter for our RF model was automatically tuned using the sk-learn package. In our experiment, the RF serves to predict the most possible target from the KL-divergence measured by each candidate target. Combined with nonparametric density estimation and KLD, we suggest that the RF model provides a possible solution for the DTI prediction problem.

3. Results And Discussion

In this study, the probabilistic modeling of similarity scores was used to describe the features of a given ligand (drug) in the RF model. For this purpose, first, nonparametric density estimation put the similarity information (of Q-Q or Q-L) into KLD equation. Second, the KLD values enabled quantitative comparisons between targets against a ligand (query). Finally, the RF classifier was built using the KLD feature vectors for DTI prediction. At this time, we describe notable results of these serial works including predictive power of the RF classifier and feature analysis.

3.1 Representation of targets and ligands via non-parametric probability distribution model.

Herein we introduce a terminology called target class. Even if Q-Q matrix characterize a target using 3D similarity between its ligands, the matrix also can be considered as a group (a class) of ligands sharing a target. Thus, in order to conveniently call the group of a specific Q-Q matrix, we call each group a target class with its target name. Similarity information of target classes was represented by nonparametric probability distribution model of respective Q-Q matrix. Clearly, while many classes were slightly shewed but similar to Gaussian distribution, some classes were very far from Gaussian distribution such as fibroblast growth factor receptor 1(FGFR1). The Fig. 3 illustrates that the probability density of each target class can be severe asymmetry and skewed shape, making it difficult to assume their structural consistency. Notably, FGFR1 having the number of Q10 cannot be well fitted to Gaussian model. Without structural (e.g., Gaussian, gamma) assumptions on similarity data, nonparametric density estimation provided more flexibility and less information loss than our previous Gaussian mixture models (GMM)[16]. As shown in the Fig. 3 and **Supplementary Fig. 1**, KDE perfectly fitted unique distribution of respective target classes. Clearly, the results of Fig. 3 are quite discrepant from almost studies using chemical similarity, which assumed similarity distribution is Gaussian distribution. Because composition of target classes is diverse according to the existence of orthosteric ligands, allosteric ligands, and non-direct binding regulators, it is natural that their distribution are dissimilar each other not to follow Gaussian distribution. Thus, we appeal the KDE distribution is a better reasonable method than Gaussian distribution to describe chemo-centric prediction.

Moreover, the relationship of a specific ligand (drug) with a target also was represented onto the KDE model of respective Q-L vector. Surely, data dimension was very different between Q-L vector due to different number of ligands within a target class (max size of Q-L vector: 15,000). Fortunately, regardless of the size of the dataset, the KDE provided a stable (sufficiently good) density distribution for further study. The probability density distribution of each Q-L vector not only projects respective pair's (of a ligand or a target class) characteristics, but also allows comparison between target classes based on a chemo-centric assumption. In other words, KDE distribution of Q-L vector indicates DTI. However, the DTI should be generalized and quantified for the comparison between different drugs or targets. For this purpose, we used KLD as an information entropy. Thus, paradoxically, 'extraordinary' density distribution (showing severe asymmetry, skewness, fat tale) is preferred to check utility of this study in which the entropy (KLD) is calculated without considering statistical rules.

[INSERT FIGURE3 HERE]

3.2 KLD as drug target-interaction descriptor

If a new drug tends to be more similar to ligands of a specific target class (e.g. Q1) than them of other classes, DTI prediction of the drug indicates the target. However, probability densities of Q-Q matrices considerably vary between target classes (Fig. 3). Similarity scores with some ligands in a target class do not properly work as a feature for the DTI prediction so that chemical similarity is less preferred as a single feature of DTI prediction. Moreover, importance of the similar ligands also depends on their position in the probability density verified by descriptive statistics (e.g., mean and standard deviation) of respective Q-Q matrix. Meanwhile, KLD calculation include the relationship between all ligands of a target class through q(x) of the Q-Q matrix (target-specific information) and also have the relationship of a query drug with the target class through p(x) of the Q-L vector (ligand-specific information). Thus, the divergence quantifies DTI prediction between the drug and target class through comparing the q(x) and p(x). In detail, when a new drug shows a smaller KLD value for a specific target class than KLD values for other classes, we predict DTI of the drug-target pair. Such information makes the KLD based DTI prediction is distinct from any chemical descriptors or similarity scores (one KLD value; relationship of one drug-target pair *vs* one similarity value: that of one drug-drug pair *vs* one descriptor: information on one drug). Thus, we tried to elicit potential of the distribution divergence as a DTI descriptor. The KDE distribution, which was noted in Section 3.1, showed a suitable proxy representing q(x) of the Q-Q matrix and p(x) of the Q-L vector. Each similarity matrix (Q-Q matrix), which identify 'relevance' between ligands in a target class, originated target-specific information. The 'individual' (ligand-target) density and 'collective' (target-target) density could be compared each other through KLD. In other words, in order to examine target-target density, KLD values between target classes (Q-Q vs Q-Q matrix) were calculated. In addition, a reverse divergence quantity was calculated after putting q(x) and p(x) in reverse position. The dual quantity between target pairs (KLD and reverse KLD) of Table 1 showed a correlation between classes in target spaces. The pair with lower divergence suggests that those target classes have similar distribution inferred to have similar characteristics. Similarly, ligand-target density also was examined through dual quantity (KLD and reverse KLD) between a Q-Q matrix and a Q-L vector (Fig. 4). In detail, if target of a query was identical with target of ligands in Q-Q matrix, the Q-L vector of the query was a part of Q-Q matrix. Else, the Q-L vector of a query is filled through similarity calculation of pairwise query-ligand. The KLD measures the extent to which a query (ligand) is different from a target. We directly apply this notion to quantify whether a query belongs to a target. For example, when considering the two targets A and B, the probability density function of similarity data obtained from the "ANY" query contained in target A (the query belongs to target A), and the KLD from the Q-Q density from target A are both close to zero. On the other hand, when calculating KLD from the counterpart distribution of target class B, the Q-L vector belonging to class A tends to have relatively larger values than Q-Q density of target A. The result suggests that KLD calculated from two KDE distributions make DTI prediction realized.

[INSERT TABLE1 HERE]

[INSERT FIGURE4 HERE]

Table 1
KLD between target pairs (Q-Q matrix vs another Q-Q matrix, 17 x 17 Target).[a]

T	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1	0.00	0.11	0.57	0.12	0.02	0.03	0.38	0.13	0.15	0.47	0.29	0.11	**2.54**	0.09	0.04	0.02	0.03
2	0.11	0.00	0.50	0.22	0.16	0.21	0.13	0.41	0.08	0.42	0.18	0.33	1.85	0.32	0.13	0.16	0.17
3	0.36	0.55	0.00	0.20	0.43	0.41	1.13	0.33	0.89	0.27	1.17	0.37	1.15	0.29	0.40	0.27	0.48
4	0.09	0.24	0.23	0.00	0.13	0.14	0.66	0.10	0.43	0.17	0.66	0.06	1.38	0.06	0.19	0.07	0.14
5	0.02	0.18	0.63	0.17	0.00	0.01	0.40	0.10	0.15	0.66	0.27	0.11	**2.90**	0.08	0.03	0.04	0.01
6	0.04	0.23	0.69	0.21	0.01	0.00	0.46	0.10	0.18	0.77	0.28	0.13	**3.19**	0.09	0.04	0.05	0.02
7	0.35	0.11	0.66	0.53	0.40	0.45	0.00	0.80	0.09	0.79	0.09	0.71	**2.07**	0.68	0.33	0.44	0.41
8	0.18	0.55	0.54	0.14	0.17	0.15	1.18	0.00	0.74	0.49	1.03	0.03	**2.49**	0.01	0.24	0.09	0.17
9	0.13	0.09	0.78	0.38	0.13	0.16	0.10	0.42	0.00	0.86	0.03	0.39	**2.98**	0.36	0.11	0.21	0.14
10	0.40	0.61	0.31	0.16	0.48	0.48	1.28	0.32	1.00	0.00	1.33	0.27	0.87	0.27	0.52	0.28	0.49
11	0.28	0.26	1.08	0.68	0.25	0.25	0.16	0.61	0.04	1.45	0.00	0.64	**4.21**	0.57	0.18	0.38	0.27
12	0.12	0.39	0.47	0.07	0.12	0.12	0.86	0.03	0.53	0.34	0.76	0.00	**2.11**	0.02	0.22	0.07	0.11
13	**3.43**	**3.99**	**2.22**	**2.39**	**3.49**	**3.47**	**4.80**	**2.71**	**4.52**	**1.74**	**4.98**	**2.51**	0.00	**2.72**	**4.08**	**3.30**	**3.45**
14	0.12	0.40	0.45	0.07	0.12	0.11	0.95	0.01	0.58	0.38	0.84	0.02	**2.15**	0.00	0.18	0.05	0.12
15	0.05	0.17	0.72	0.28	0.04	0.04	0.41	0.19	0.15	0.84	0.26	0.23	**3.31**	0.17	0.00	0.07	0.06
16	0.03	0.18	0.45	0.08	0.04	0.05	0.56	0.08	0.30	0.38	0.49	0.07	**2.30**	0.05	0.06	0.00	0.05
17	0.03	0.19	0.75	0.21	0.01	0.02	0.43	0.11	0.15	0.73	0.28	0.10	**3.21**	0.10	0.05	0.05	0.00

[a]The lower KLD values indicate that the pairs has similar distribution.

3.3 DTI Prediction of RF Classifier

A binomial classification model was built using KLD for DTI prediction of individual query drugs. Predictive models from the divergence-coordinated features were investigated in the divided condition of training (80%) and test (20%) datasets. Random forest (RF) algorithm showed desirable classifier for DTI prediction with reliable statistical performance (Table 2, Fig. 5–6). Despite the imbalanced number of ligands between different targets, the ensemble learning showed acceptable precision and recall in test set of every targets (Table 2). The average validation accuracy also was 0.88 in five-fold cross validation. Moreover, we visualized our model by depicting both the receiver operating characteristic (ROC) curve and box plot. As shown in Fig. 5, area under the curve (AUC) values (more than 0.96), which indicated the area under the ROC curve, also showed the predictive performance with successful confusion matrix (see **Suppl. Information**). The ROC curve also illustrates that there is no significance dependence on accuracy among ligands classified by targets. Furthermore, average precision according to percentile rank of KLD features described the distributional information of this predictive model in the boxplot in the box plot (Fig. 6). Patterns in a 'RESPONSE' of the RF classifier are shown in the box plot. The horizontal line (Orange) shows a 'skewed' decision boundary in the RF classifier inherited from the characteristics of our RAW dataset showing an 'irregular' probability density.

[INSERT TABLE2 HERE]

[INSERT FIGURE5 HERE]

[INSERT FIGURE6 HERE]

Table 2
Statistical performance of RF model predicting DTI.

Target No.	Precision	Recall	F1-score
Q1	0.91	0.93	0.92
Q2	0.97	0.95	0.96
Q3	0.99	0.95	0.97
Q4	0.98	0.95	0.97
Q5	0.88	0.92	0.9
Q6	0.84	0.79	0.81
Q7	0.95	0.97	0.96
Q8	0.83	0.87	0.85
Q9	0.87	0.91	0.89
Q10	0.97	0.9	0.94
Q11	0.76	0.78	0.77
Q12	0.85	0.9	0.87
Q13	1	1	1
Q14	0.81	0.9	0.85
Q15	0.84	0.79	0.81
Q16	0.93	0.92	0.93
Q17	0.79	0.67	0.73

3.4 Feature correlation and importance of KLD based classifier

To interpret the DTI model, we conducted feature analysis of the correlation matrix between features (Fig. 7) and pruning of less important features (Fig. 8). In addition to correlation, the relative importance of a feature in an RF model is measurable with respect to the dependent variable. Figure 7 shows pairwise correlation coefficients, which reflect the quantity of dependence among features. Each value directly corresponds to a lower divergence between the Q and Q densities of their target classes. Providing a criterion for variable selection, a high correlation among the subset of features tends to dilute the importance of such features, confusing the prediction accuracy. Fortunately, almost DTI features except for KLD17 vector (generated from ligands of Q17, Epidermal growth factor receptor) showed acceptable correlation coefficient less than 0.7. There are several methods for calculating the feature importance in terms of their influence on the model. The most common metric, the mean decrease in impurity (MDI) defines the mean of impurity reduction as the importance criterion when each feature is deleted in a model. If the corresponding feature value is randomly assigned, the prediction result becomes lower than a benchmark case, which can be considered important and vice versa. The higher the importance of a feature in our study implies the uniqueness of the Q-Q density function, which is more comparable. Figure 8 illustrates the importance of features in the DTI model. Generally, pruning less important features is expected to result in higher classification accuracy. In our DTI model, more than 10 features retained accuracy more than 0.8. The selection of features was important in terms of model stability as well as model accuracy. Focusing on the small numbers of 10 to 15 features is acceptable in terms of dimensionality issues. Because the standard size of training samples was 15,000 for each target, the 10 to 15 sized features are reasonable to avoid overfitting.

[INSERT FIGURE7 HERE]

[INSERT FIGURE8 HERE]

4. Conclusion

In this study, we presented an RF model for identifying drug-target interactions using KLD. Our novel combination of nonparametric density estimation, KLD, and RF model showed an effective chemo-centric method for drug discovery. It has been widely emphasized that similarity vector plays an important role in identifying drug-target interactions. The RF model leverages more specific information than our previous

approaches using an information metric-designed feature vector. Furthermore, pairwise comparison of ligands and their candidate targets explicitly describes a ligand's characteristics, which provide a bridge for an ML classifier. When faced with a computationally limited environment, note that the dimension (size of the feature vector) can be controlled depending on the number of target spaces. Other than the Tanimoto coefficient, another similarity metric (e.g., cosine similarity, sorgel similarity) of the descriptor (fingerprint), also becomes a proxy for describing a ligand in the context of our methodology. In future studies, it will be possible to apply other types of ML algorithms to our model framework, suggesting that our method is applicable to other biomedical contexts.

Declarations

Supplementary `:

Supplementary sheets and files are available.

Availability of data and materials:

Python code, and refined data will be available in GitHub.

https://github.com/college-of-pharmacy-gachon-university/KLD2

Conflict of interests:

The authors confirm that this article content has no conflicts of interest.

Funding:

This study was supported by the Basic Science Research Program of the National Research Foundation of Korea (NRF), which is funded by the Ministry of Education, Science and Technology (No.:2017R1E1A1A01076642).

Acknowledgments:

The authors would like to thank OpenEye Scientific Software for providing an academic free license.

Authors' contributions:

M.-h. K. and S. A. conceived and designed the study. S. A. carried out all modeling & data work. M.-h. K. and S. A. analyzed results, wrote the manuscript, and S. L. revised it. M.-h. K. provided every research work facility. All authors read and approved the final manuscript.

References

1. Svava Ósk Jónsdóttir, Flemming Steen Jørgensen, Søren Brunak, Prediction methods and databases within chemoinformatics: emphasis on drugs and drug candidates, Bioinformatics, Volume 21, Issue 10, Pages 2145–2160.
2. Nigsch F, Bender A, Jenkins JL, Mitchell JBO. Ligand-target prediction using Winnow and naive Bayesian algorithms and the implications of overall performance statistics. J Chem Inf Model 2008, 48: 2313–2325.
3. Ma J, Sheridan RP, Liaw A, Dahl GE, Svetnik V (2015) Deep neural nets as a method for quantitative structure–activity relationships. J Chem Inf Model 55:263–274
4. Weininger, D. (1988). SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. Journal of chemical information and computer sciences, *28*(1), 31–36.
5. Willett P. Similarity-based virtual screening using 2D fingerprints. Drug Discov Today. 2006;11(23–24):1046–1053.
6. Axen, S. D., Huang, X. P., Cáceres, E. L., Gendelev, L., Roth, B. L., & Keiser, M. J. (2017). A simple representation of three-dimensional molecular structure. Journal of medicinal chemistry, *60*(17), 7393–7409.
7. Duan J, Dixon SL, Lowrie JF, et al. Analysis and comparison of 2D fingerprints: insights into database screening performance using eight fingerprint methods. J Mol Graph Model. 2010;29(2):157–170.
8. Extended-Connectivity Fingerprints, David Rogers and Mathew Hahn Journal of Chemical Information and Modeling 2010 *50* (5), 742–754.
9. Matter H. Selecting Optimally Diverse Compounds from Structure Databases: A Validation Study of Two-Dimensional and Three-Dimensional Molecular Descriptors. J Med Chem. 1997;40:1219–1229.

10. Schulz-Gasch T, Schärfer C, Guba W, Rarey M. TFD: Torsion Fingerprints as a New Measure to Compare Small Molecule Conformations. J Chem Inf Model. 2012;52:1499–1512.
11. Keiser MJ, Setola V, Irwin JJ, Laggner C, Abbas AI, Hufeisen SJ, Jensen NH, Kuijer MB, Matos RC, Tran TB, Whaley R, Glennon RA, Hert J, Thomas KLH, Edwards DD, Shoichet BK, Roth BL. Predicting New Molecular Targets for Known Drugs. Nature. 2009;462:175–181
12. van Laarhoven, T., Nabuurs, S. B., & Marchiori, E. (2011). Gaussian interaction profile kernels for predicting drug–target interaction. Bioinformatics, *27*(21), 3036–3043.
13. He, Z., Zhang, J., Shi, X. H., Hu, L. L., Kong, X., Cai, Y. D., & Chou, K. C. (2010). Predicting drug-target interaction networks based on functional groups and biological features. PloS one, *5*(3), e9603.
14. Fakhraei, S., Raschid, L., & Getoor, L. (2013, August). Drug-target interaction prediction for drug repurposing with probabilistic similarity logic. In *Proceedings of the 12th International Workshop on Data Mining in Bioinformatics* (pp. 10–17).
15. Hao M, Wang Y, Bryant SH. Improved prediction of drug-target interactions using regularized least squares integrating with kernel fusion technique. Anal Chim Acta. 2016 Feb 25;909:41–50. Epub 2016 Jan 14. PMID: 26851083; PMCID: PMC4744621.
16. Lee, S. H., Ahn, S., & Kim, M. H. (2020). Comparing a Query Compound with Drug Target Classes Using 3D-Chemical Similarity. International Journal of Molecular Sciences, *21*(12), 4208.
17. David Mendez, Anna Gaulton, A Patrícia Bento, Jon Chambers, Marleen De Veij, Eloy Félix, María Paula Magariños, Juan F Mosquera, Prudence Mutowo, Michał Nowotka, María Gordillo-Marañón, Fiona Hunter, Laura Junco, Grace Mugumbate, Milagros Rodriguez-Lopez, Francis Atkinson, Nicolas Bosc, Chris J Radoux, Aldo Segura-Cabrera, Anne Hersey, Andrew R Leach, ChEMBL: towards direct deposition of bioassay data, *Nucleic Acids Research*, Volume 47, Issue D1, 08 January 2019, Pages D930–D940.
18. Montaruli, M., Alberga, D., Ciriaco, F., Trisciuzzi, D., Tondo, A. R., Mangiatordi, G. F., & Nicolotti, O. (2019). Accelerating Drug Discovery by Early Protein Drug Target Prediction Based on a Multi-Fingerprint Similarity Search. Molecules (Basel, Switzerland), *24*(12), 2233. https://doi.org/10.3390/molecules24122233
19. OMEGA 4.0.0.4: OpenEye Scientific Software, Santa Fe, NM. http://www.eyesopen.com.
20. Hawkins, P.C.D.; Skillman, A.G.; Warren, G.L.; Ellingson, B.A.; Stahl, M.T. Conformer Generation with OMEGA: Algorithm and Validation Using High Quality Structures from the Protein Databank and the Cambridge Structural Database J. Chem. Inf. Model. 2010, 50, 572–584.
21. Shape Toolkit
22. Hawkins, P.C.D.; Skillman, A.G.; Warren, G.L.; Ellingson, B.A.; Stahl, M.T. Conformer Generation with OMEGA: Algorithm and Validation Using High Quality Structures from the Protein Databank and the Cambridge Structural Database J. Chem. Inf. Model. 2010, 50, 572–584.
23. J. Beirlant, E. Dudewicz, L. Gyorfi, and E. van der Meulen. Nonparametric entropy estimation: An overview. International Journal of the Mathematical Statistics Sciences, pages 17–39, 1997.
24. Chang, D.TH., Wang, CC. & Chen, JW. Using a kernel density estimation based classifier to predict species-specific microRNA precursors. BMC Bioinformatics 9, S2 (2008).
25. Hsieh, CH., Chang, D.TH., Hsueh, CH. *et al*. Predicting microRNA precursors with a generalized Gaussian components based density estimation algorithm. BMC Bioinformatics 11, S52 (2010).
26. Kausar, S., Falcao, A.O. A visual approach for analysis and inference of molecular activity spaces. J Cheminform 11, 63 (2019). https://doi.org/10.1186/s13321-019-0386-z
27. Virtanen, P. et al., 2020. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. Nature Methods.
28. Lin, J. (1991). Divergence measures based on the Shannon entropy. IEEE Transactions on Information theory, *37*(1), 145–151.
29. S. Kullback and R. A. Leibler. On information and sufficiency. Ann. Math. Statistics, 22(1):79–86, 3 1951.
30. Y. K. Lee and B. U. Park. Estimation of kullback-leibler divergence by local likelihood. Annals of the Institute of Statistical Mathematics, 58(2):327–340, 6 200
31. Breiman L, Friedman JH, Olshen RA, Stone CJ. Classification and Regression Trees. Wadsworth;1984.
32. Breiman, L.: Random forests. Mach. Learn. 45, 5–32 (2001). DOI 10.1023/A:1010933404324
33. Riddick G, Song H, Ahn S,Walling J, Borges-Rivera D, ZhangW, et al. Predicting in vitro drug sensitivity using random forests. Bioinformatics. 2011; 27: 220–224. https://doi.org/10.1093/bioinformatics/btq628 PMID: 21134890
34. Lind, A. P., & Anderson, P. C. (2019). Predicting drug activity against cancer cells by random forest models based on minimal genomic information and chemical properties. PloS one, *14*(7), e0219774.

35. Shi, H., Liu, S., Chen, J., Li, X., Ma, Q., & Yu, B. (2019). Predicting drug-target interactions using Lasso with random forest based on evolutionary information and chemical structure. Genomics, *111*(6), 1839–1852.
36. Cano, G., Garcia-Rodriguez, J., Garcia-Garcia, A., Perez-Sanchez, H., Benediktsson, J. A., Thapa, A., & Barr, A. (2017). Automatic selection of molecular descriptors using random forest: Application to drug discovery. Expert Systems with Applications, *72*, 151–159.
37. Yang, P., Hwa Yang, Y., Zhou, B., Zomaya, Y., et al.: A review of ensemble methods in bioinformatics. Current Bioinformatics 5(4), 296–308 (2010).

Figures

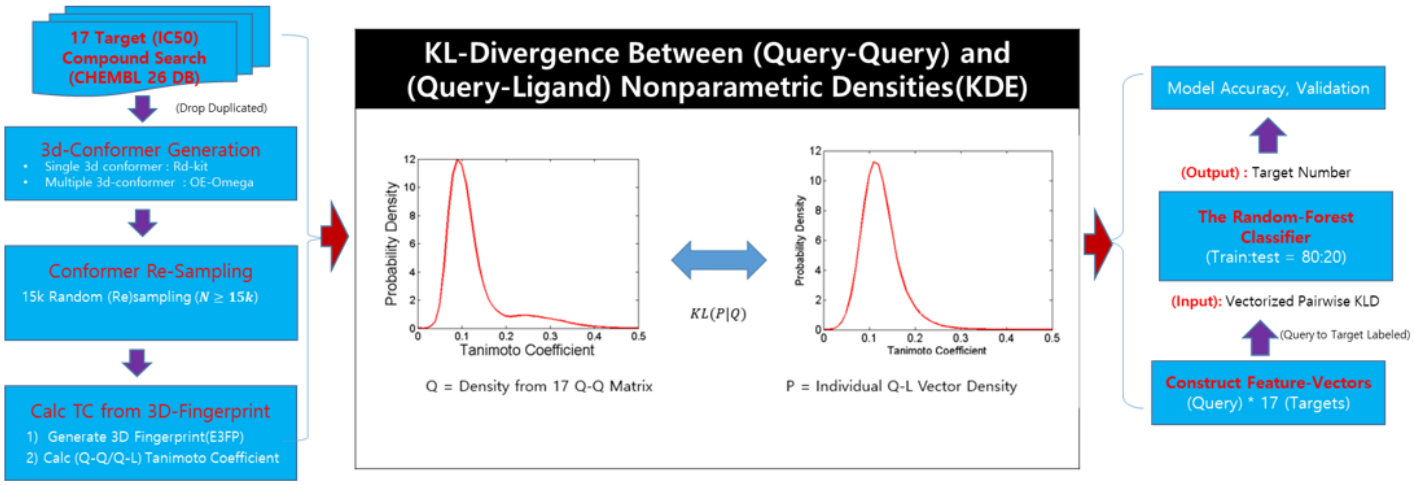

Figure 1

Overview of this study. Kullbeck–Leibler divergence (KLD) between chemical similarity distributions (of Q-Q matrix and Q-L vector) provided feature vectors for drug target interaction (DTI) prediction. The distributions were generated through kernel density estimation (KDE) as a nonparametric density model, which is quite distinct with Gaussian distribution defined with mean and standard deviation of a sample.

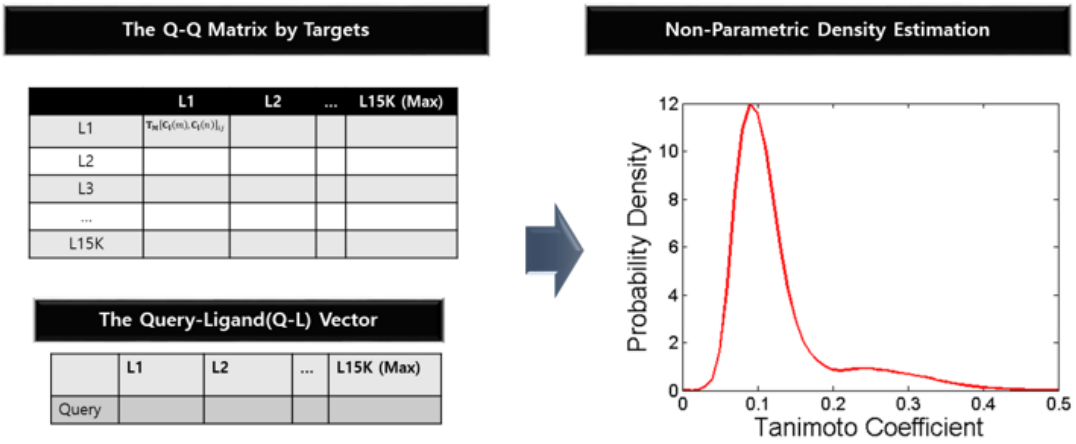

Figure 2

Transformation of a Q-Q matrix or a Q-L vector into KDE distribution.

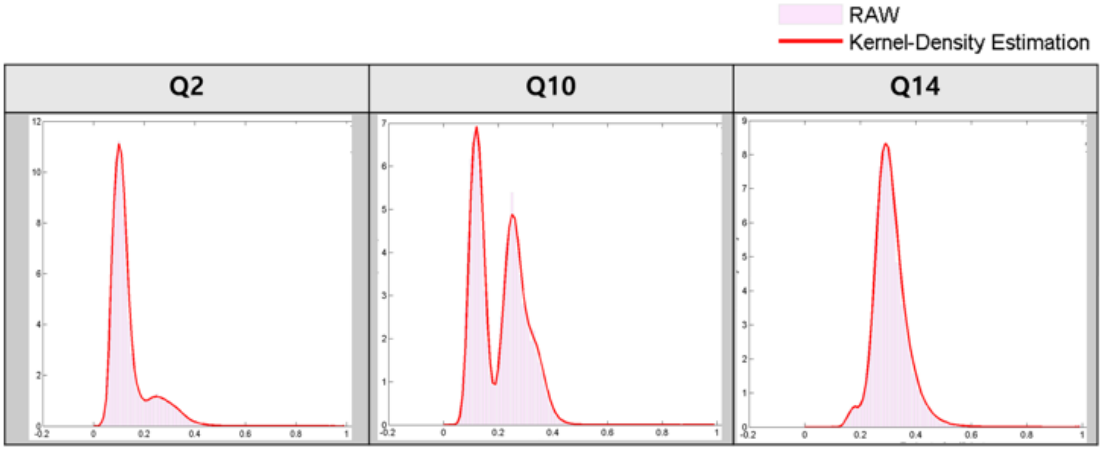

Figure 3

Representative examples of KDE model of three target classes. Q2: Heat shock protein 90 (HSP90), Q10: Fibroblast growth factor receptor 1 (FGFR1), Q14: Neuraminidase - Influenza A virus (HA). X-axis: 3D similarity (Tanimoto coefficient), Y-axis: relative frequency.

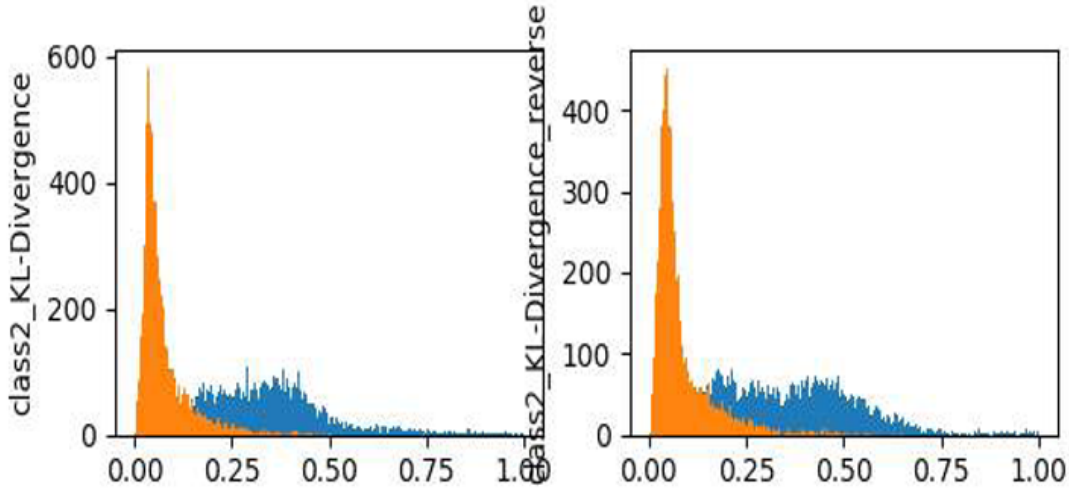

Figure 4

Example dual quantity (KLD and reverse KLD) comparison between a Q-Q matrix and a Q-L vector. Q2 was used as a Q-Q matrix. X-axis: KLD value, Y-axis: relative frequency. Self-comparison (orange),

Figure 5

ROC curves to show DTI prediction performance. X-axis: false positive rate, Y-axis: true positive rate. Each line indicates respective target class with AUC area.

Figure 6

Boxplots to show DTI prediction performance. X-axis: percentile rank of KLD features, Y-axis: average precision.

Figure 7

Correlation map between KLD feature vectors in RF model

Figure 8

Feature pruning result of less important features to show out-of-bag score.

X-axis: the number of KLD feature vectors, Y-axis: accuracy with respect to the number of features.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SuppInfoKLD2v3.docx