A quantified discrimination of a query compound between target classes under 3D-similarity metric

Sanghyeok Lee
Gachon University College of Pharmacy https://orcid.org/0000-0002-8830-7060
Sangjin Ahn
Ajou University https://orcid.org/0000-0003-2398-2249
Mi-hyun Kim (✉ kmh0515@gachon.ac.kr)
Gachon University College of Pharmacy https://orcid.org/0000-0002-2718-5637

Research article

Keywords: Kullback-Leibler (K-L) divergence, Chemocentric similarity, Tanimoto-Jaccard coefficient, Gaussian mixture model (GMM), expectation-maximization (EM) algorithm, Maximum likelihood (ML) estimation, Machine learning

Posted Date: February 19th, 2020

DOI: https://doi.org/10.21203/rs.2.23985/v1

A quantified discrimination of a query compound between target classes under 3D-similarity metric

Sanghyeok Lee,[1] Sangjin Ahn,[2] Mi-hyun Kim[1]*

[1]Gachon Institute of Pharmaceutical Science and Department of Pharmacy, College of Pharmacy, Gachon University, Yeonsu-gu, Incheon, 21936, Republic of Korea, [2]Department of Financial Engineering, College of Business, Ajou University, Suwon, 16499, Republic of Korea

[*] Author for correspondence

E-mail: kmh0515@gachon.ac.kr

Abstract

Background: 3D similarity is useful to predict the profiles of unprecedented molecular frameworks, 2D dissimilar to known compounds. Basically, when comparing compound pairs, 3D similarity of the pairs depends on conformational sampling of compounds, alignment method, chosen descriptors, and metric to show limited discriminative power. In addition to four factors, 3D chemocentric target prediction of an unknown compound requires compound - target associations. The associations for the target prediction replace compound-to-compound comparison with compound-to-target comparison.

Results: Quantitative comparison of query compounds to target classes (one-to-group) could be acquired using two type similarity distributions: one is from maximum likelihood (ML) estimation of queries and another is from Gaussian mixture model (GMM) of target classes. While Jaccard-Tanimoto similarity of query-to-ligand pairs could be transformed into query distribution through ML estimation, the similarity of ligand pairs within each target class could be transformed into the representative distribution of a target class through GMM, hyperparameterized through expectation-maximization (EM) algorithm. To quantify the discriminativeness of a query ligand against target classes, Kullback-Leibler (K-L) divergence was calculated between two distributions.

Conclusions: Stratified sampled 14K ligands from four target classes, estrogen receptor alpha (ESR), vitamin D receptor (VDR), cyclooxygenase-2 (COX2), and cathepsin D (CTSD) presented whether or not each query can be a representative ligand of each target through compared K-L divergence value. The feasibility index, F_m and the probability, $\mathbb{P}(\nu(l_m) = i)$ from K-L divergence could summarize 3D chemocentric relationship between target classes.

***Keywords*:** Kullback-Leibler (K-L) divergence; Chemocentric similarity; Tanimoto-Jaccard coefficient; Gaussian mixture model (GMM); expectation-maximization (EM) algorithm; Maximum likelihood (ML) estimation; Machine learning

■ Introduction

In early drug discovery, 3D-similarity between chemicals have been used for virtual screening (VS) to find desirable ligands for a chosen therapeutic target [1-2]. To our knowledge, chemical similarity has been a coarse predictor for filtering out less promising chemicals rather than selecting the most desirable one. The chemical similarity also has contributed to target screening, retro-VS, under the chemocentric assumption, two similar molecules are likely to have similar properties so that they can share biological targets or can show similar pharmacological profile [3-4]. Remarkably, Jain's group conducted on-target and off-target prediction through the comparison of 2D and 3D chemical similarity [5]. Based on the comparison, although dual 2D and 3D similarity-based prediction could show superiority for either 2D or 3D prediction, 3D prediction could not show dramatic improvement over 2D prediction. Therefore, when considering cost and speed for the prediction and feature generation, with the increase of data points according to the chosen conformer number, 3D similarity may not be more efficient than any feature. However, despite its less cost-effectiveness, 3D similarity can be the best feature for in silico target screening of unprecedented drug scaffolds, new druglike molecular frameworks [6], for the following reasons. (1) Novel, unprecedented drug scaffolds can have very low 2D similarity to any known bioactive molecule [7-9], (2) a novel pharmacological profile of a drug can be more frequently found through 3D similar off-target prediction [5], and (3) drug properties close to reality can be generated from their realistic and flexible 3D structure [10-12].

The internalization of Michelangelo Buonarroti's quote, 'every block of stone (chemical) has a statue (utility) inside it, and it is the task of the sculptor (chemist) to discover it' gave us the research inspiration for 'chemistry-oriented synthesis' of an unprecedented drug scaffold [7-9] and chemocentric target profiling of the scaffold [7]. For the purpose, we have intensively studied 3D-similarity of unprecedented drug scaffolds (query) with known compounds (reference). When comparing compound pairs of the queries and references, 3D similarity of the pairs depends on (1) conformational sampling of compounds, (2) alignment method, (3) chosen descriptors, and (4) distance metric (eg. Jaccard-Tanimoto). In addition to the four factors of 3D VS, the retro-VS of unprecedented drug scaffolds (query) requires compound - target associations (target class information) in Figure 1. The associations can make substantial difference between VS and retro-VS in the view of problem solving in data science: (1) one-to-one comparison for VS in Figure 1(a), (2) one-to-group (class) comparison for retro-VS in Figure 1(b), and (3) group-to-group comparison for typical parametric statistics such as ANOVA, t-test. When we calculate similarity of compound pairs in retro-VS, we don't want to find the most similar compound of our query but ultimately hope to know the most reliable target of the query through transformation of chemical similarity. To get the goal, one-to-group comparison should be essentially quantified. However, to our knowledge, such measurement doesn't have been reported in cheminformatics.

Herein, we tried to quantitatively compare a query compound with a target class (one-to-group) using two type similarity distributions: one is from maximum likelihood (ML) estimation of queries and another is from Gaussian mixture model (GMM) of target classes. As raw data of the distributions, Jaccard-Tanimoto similarity coefficient were calculated in (1) query-to-ligand pairs and (2) the ligand pairs within each target class. The query-to-ligand similarity tried to be transformed into query distribution through ML estimation. The ligand pair similarity also tried to be transformed into the representative distribution of a target class through GMM. The difference between two distributions could be quantified by Kullback-Leibler (K-L) divergence. In other words, K-L divergence of a query-to-target means quantified comparison between a query to a target class. In order to evaluate whether the K-L divergence is right one-to-group comparison or not, we try to test whether any query chosen from known ligands within a target can discriminate the original target from other targets or not. In sequence, target profile of an unprecedented drug scaffold was explained by the K-L divergence.

■ Method

Kullback-Leibler divergence. The K-L divergence measures the difference between two statistical or probabilistic distributions. In particular, K-L divergence has been employed in various machine learning & deep learning algorithm for statistical inference [17, 18]. Since K-L divergence implies relative entropy, which is important concept to understand statistical phenomena, it has applied to statistical physics, chemistry and social science.

Let us define two probability spaces $(\Omega, \mathcal{F}, P)$ and $(\Omega, \mathcal{F}, Q)$, where Ω is the sample space, $\mathcal{F}$ is $\sigma -$algebra, and P and Q are probability distributions. Then, to define Kullback-Leibler divergence, there should be a unique measurable function, $\frac{dQ}{dP} : \Omega \rightarrow \mathbb{R}^{+}$, which is called as the Radon-Nykodym derivative, such that

$$Q(\mathrm{E}) = \int_{E} \frac{dQ}{dP} dP \qquad (2.1)$$

for any measurable set $\mathrm{E} \in \Omega$ [19] using the measurable function, $\frac{dQ}{dP}$. We define the Kullback-Leibler divergence, $\mathrm{D(P||Q)}$, as either

$$D(P\|Q) := \int_{\Omega} -\ln\left(\frac{dP}{dQ}\right) dP \quad (2.2)$$

$$\text{or } D(P\|Q) := \int_{-\infty}^{\infty} \ln\left(\frac{p(x)}{q(x)}\right) p(x)dx \quad (2.3)$$

where the probability density functions p(x) and q(x) are defined as

$$P(x) := \int_{-\infty}^{x} p(x)dx \text{ and } Q(x) := \int_{-\infty}^{x} q(x)dx. (2.4)$$

The Kullback-Leibler divergence means the information for comparing distributions, P(x) and Q(x) [19]. Hence, implication of the Kullback-Leibler divergence depends on the definitions of P(x) and Q(x). For example,

- (Model Inference) If P(x) is the testing distribution based on the model, and Q(x) is the distribution from the raw data, the difference is the error between the model and reality [20].
- (Informatics) If P(x) and Q(x) are information extracted from two objectives, the divergence is a measurement for the discrimination between two objectives [21, 22].
- (Bayesian Statistics) If P(x) is a prior distribution and Q(x) is a posterior distribution, the divergence means information gained through updating [23].

In sequence, let us consider a special example. Assume the probability distributions P(x) and Q(x) replace the Gaussian distributions $\mathrm{G}(x;m_i,\sigma_i)$ and $\mathrm{G}(x;m_j,\sigma_j)$, where

$$\mathrm{G}(x;m_i,\sigma_i) \coloneqq \int_{-\infty}^{x} g(s;m_i,\sigma_i)ds \text{ and } \mathrm{G}(x;m_j,\sigma_j) \coloneqq \int_{-\infty}^{x} g(s;m_j,\sigma_j)ds \quad (2.5)$$

and the probability density functions $g(s;m_i,\sigma_i)$ and $g(s;m_j,\sigma_j)$ are follows,

$$\begin{cases} g(s;m_i,\sigma_i) = \frac{1}{\sigma_i\sqrt{2\pi}}\exp\left(-\frac{(s-m_i)^2}{2(\sigma_i)^2}\right) \\ g(s;m_j,\sigma_j) = \frac{1}{\sigma_j\sqrt{2\pi}}\exp\left(-\frac{(s-m_j)^2}{2(\sigma_j)^2}\right) \end{cases} . \quad (2.6)$$

Using (2.3) and (2.5), the Kullback -Leibler divergence between the two Gaussian distributions $\mathrm{G}(x;m_i,\sigma_i)$ and $\mathrm{G}(x;m_j,\sigma_j)$ in (2.5) are as follows.

$$D\left(G(x;m_i,\sigma_i)\middle\|G(x;m_j,\sigma_j)\right)$$

$$=\square_i\left[\ln\left(\frac{G(x;m_i,\sigma_i)}{G(x;m_j,\sigma_j)}\right)\right]$$

$$= \int_{-\infty}^{\infty} g(s;m_i,\sigma_i)\ln(g(x;m_i,\sigma_i))ds - \int_{-\infty}^{\infty} g(s;m_i,\sigma_i)\ln(g(x;m_j,\sigma_j))ds$$

$$= \int_{-\infty}^{\infty} g(s;m_i,\sigma_i) \ln\left(-\frac{(s-m_i)^2}{2(\sigma_i)^2} - \ln\left(\sigma_i\sqrt{2\pi}\right)\right) ds + \int_{-\infty}^{\infty} g(s;m_i,\sigma_i) \ln(-\frac{(s-m_j)^2}{2(\sigma_j)^2} +$$

$$\ln(\sigma_j\sqrt{2\pi}))\, ds \qquad (2.7)$$

By the following relationships,

$$\int_{-\infty}^{\infty} g(s;m,\sigma)ds = 1,$$

$$\int_{-\infty}^{\infty} sg(s;m,\sigma)ds = m, \qquad (2.8)$$

$$\int_{-\infty}^{\infty} (s-m)^2 g(s;m,\sigma)ds = (\sigma)^2,$$

we obtain

$$= -\frac{1}{2} - \ln(\sigma_i\sqrt{2\pi}) + \frac{(\sigma_i)^2 + \left(m_i - m_j\right)^2}{2\left(\sigma_j\right)^2} + \ln(\sigma_j\sqrt{2\pi})$$

$$= \ln\left(\frac{\sigma_j}{\sigma_i}\right) + \frac{(\sigma_i)^2 + \left(m_i - m_j\right)^2}{2(\sigma_j)^2} - \frac{1}{2} \qquad (2.9)$$

and this Kullback-Leibler divergence between univariate normal distributions (2.9) is extended to multivariate distributions [24].

Gaussian mixture model. The mixture models are methods how to analyze compositional data. Denote Φ as a probabilistic density generated from the unknown compositional data and p as well-known probability density. With $\mathbf{x}$ as a random vector, we define the functional operator $\Xi(\Phi(\mathbf{x})|p,\, K)$ as

$$\Xi(\Phi(\mathbf{x})|p,\omega,\, \lambda,\, K) \coloneqq \sum_{k=1}^{K} \omega_k\, p(\mathbf{x} : \lambda_k) \quad (2.10)$$

where for $\mathrm{k} = 1, 2, \dots, \mathrm{K}$, ω_k, λ_k are weights and vectors of hyperparameters, and p_i is the i_{th} component, which is independently and identically distributed (i.i.d) [25]. In this paper, to obtain a representative distribution, we adopt GMM [26]. Notably, GMM is a model applicable for describing non-Gaussian distributions as well as Gaussian distributions [27]. The probability density $p(\mathrm{x} : \lambda_k)$ is the Gaussian density function $\mathrm{g}(\mathrm{x}; \mathrm{m}_k, \sigma_k)$ in (2.5). In the

Gaussian mixture model, it is essential to estimate the weight(ω_k), mean(m_k), and standard deviation (σ_k). Herein, the two methods, EM algorithm [28] and ML estimation [29], were chosen for the estimation of hyperparameters from sparse and incomplete data. Concretely, EM algorithm for GMM is summarized as follows.

- Start with an initial guess for the parameters of GMM

- E-step: calculate conditional expectation of log-likelihood, $L_h()$ for incomplete data $\Phi(x)$ with respect to $\sum_{k=1}^{K} \widehat{\omega}_k^{(n)}\, g\left(\mathbf{x}\,; \widehat{m}_k^{(n)}, \widehat{\sigma}_k^{(n)}\right)$, $Q_{\boldsymbol{\Phi}}$,

$$\int \sum_{k=1}^{K} \widehat{\omega}_k^{(n)}\, g\left(\mathbf{x}; \widehat{m}_k^{(n)}, \widehat{\sigma}_k^{(n)}\right) L_h(\boldsymbol{\Phi}, \mathbf{x} | \omega_1, \cdots, \omega_K, m_1, \cdots, m_K, \sigma_1, \cdots, \sigma_K\,) d\mathbf{x} =: Q_{\boldsymbol{\Phi}}\left(\boldsymbol{\theta}\,,\ \widehat{\boldsymbol{\theta}}^{(n)}\right)$$

(2.11)

where $\boldsymbol{\theta}$, $\widehat{\boldsymbol{\theta}}^{(n)}$ are vectors of hyperparameters such that

$$\begin{cases} \boldsymbol{\theta} = (\omega_1, \cdots, \omega_K, m_1, \cdots, m_K, \sigma_1, \cdots, \sigma_K)\,, \\ \widehat{\boldsymbol{\theta}}^{(n)} = \left(\widehat{\omega}_1^{(n)}, \cdots, \widehat{\omega}_K^{(n)}, \widehat{m}_1^{(n)}, \cdots, \widehat{m}_K^{(n)}, \widehat{\sigma}_1^{(n)}, \cdots, \widehat{\sigma}_K^{(n)}\right) \ \ for\ \ n \in positive\ integers, \end{cases} \qquad (2.12)$$

and $\mathbf{x}$ is a random variable

- M-step: determine the parameters $\widehat{\boldsymbol{\theta}}^{(n+1)}$ such that

$$\widehat{\boldsymbol{\theta}}^{(n+1)} = \arg\max_{\boldsymbol{\theta}} Q_{\boldsymbol{\Phi}}(\boldsymbol{\theta}\,,\ \widehat{\boldsymbol{\theta}}^{(n)}), \qquad (2.13)$$

Then, in order to find the set of hyperparameters θ, we obtain the following recursive relationship of elements in $\widehat{\boldsymbol{\theta}}^{(n)}$:

$$\widehat{\omega}_i^{(n+1)} = \frac{1}{N}\sum_{j}^{N} z_{ij}^{(n)},$$

$$\widehat{m}_i^{(n+1)} = \frac{\sum_j^N z_{ij}^{(n)} x_j}{\sum_j^N z_{ij}^{(n)}}, \text{ and} \qquad (2.14)$$

$$\left(\widehat{\sigma}_i^{(n+1)}\right)^2 = \frac{\sum_j^N z_{ij}^{(n)}\, (x_j - \widehat{m}_i^{(m+1)})^2}{\sum_j^N z_{ij}^{(n)}}$$

where

$$z_{ij}^{(n)} := \frac{\widehat{\omega}_i^{(n)}\ g(x_j;\widehat{m}_i^{(n)}, \widehat{\sigma}_i^{(n)})}{\sum_{k=1}^{K} \widehat{\omega}_k^{(n)}\ g(x_j;\widehat{m}_k^{(n)}, \widehat{\sigma}_k^{(n)})} \qquad (2.15)$$

- Iterate the E- step and M-steps until the following condition is satisfied: there exists a positive, infinitesimal number ϵ such that

$$\left| \ \widehat{\boldsymbol{\theta}}^{(n)} - \ \widehat{\boldsymbol{\theta}}^{(m)} \right| < \epsilon, \qquad (2.16)$$

for $n > N$, where N is a proper large number [30].

For convenience, when applying the ML estimation, $\Phi(\boldsymbol{x})$ is transformed as the mixture model, $\Xi(\Phi(\boldsymbol{x})|p, \omega, \lambda, K)$ is replaced by $\Xi_{EM}(\Phi(\boldsymbol{x})|p, \omega, \lambda, K)$. When applying the ML estimation, $\Phi(\boldsymbol{x})$ is transformed as the mixture model $\Xi(\Phi(\boldsymbol{x})|p, \omega, \lambda, K)$ is replaced by $\Xi_{ML}(\Phi(\boldsymbol{x})|p, \omega, \lambda, K)$.

■ Results and discussion

In this study, we aim the quantitative method to extract the representative information (for target prediction of a query compound) from chemical similarity and known 'compound - target association' information. For the purpose, 3D-similarity distributions could be acquired from 3D similarity matrix, which is occupied by Jaccard-Tanimoto coefficients of (1) query-to-ligand pairs and (2) the ligand pairs within each target class. Jaccard-Tanimoto coefficients could be calculated from two type features, molecular shape and heteroatom features (pharmacophore features) using Openeye Toolkit. Practically, the discriminativeness between a query and a target class could be quantified according to the next process.

Step 1. EM algorithm based GMM made us obtain representative distribution (Q-distribution) for a target class, which follows Gaussian or non-Gaussian distributions.

Step 2. A query-to-ligand similarity distribution could be fitted into Gaussian distributions through ML estimation.

Step 3. K-L divergence between the two distributions from Step 1 and 2 made us predict target of the query. The higher deviation of K-L divergence values between target classes made the query more representative ligand of a class than other queries. In addition, the probability $\mathbb{P}(\nu(l_m) = i)$ from K-L divergence values and the feasibility index, F_m also made us quantify the discrimination between target classes.

Data set. In detail, let us select four pharmacological targets, four classes, which are estrogen receptor alpha (ESR), vitamin D receptor (VDR), cyclooxygenase-2 (COX2), and cathepsin D (CTSD). We randomly sample 13957 queries in each class. Before performing specific calculations, for convenience, let simple numbers denote the four classes, i.e.,

$$\begin{cases} Estrogen\ receptor\ alpha \ \rightarrow 1, \\ Vitamin\ D\ receptor \ \rightarrow 2, \\ Cyclooxygenase - 2 \rightarrow 3, \\ Cathepsin\ D \rightarrow 4. \end{cases} \tag{3.1}$$

Let either *m* or *n* be called as the class number, which is an integer between 1 and 4 such as (3.1), and $\mathrm{C}_L(m)$ and $\mathrm{C}_L(n) \in \mathbb{R}^N$ denote vectors whose element is the Tanimoto coefficient of a query in the *m*-th class. Let us define $\mathrm{T}_M : \mathbb{R}^{2N} \rightarrow \mathbb{R}^N \times \mathbb{R}^N$ as the Tanimoto matrix operator such that

$$(\mathbf{T_M}[\mathbf{C}_l(m), \boldsymbol{C}_l(n)])_{ij} := \mathrm{T_c}(<\mathbf{e_i} \cdot \ \mathbf{C_l}(m)>, <\mathbf{e_j}, \mathbf{C_l}(n)>) \tag{3.2}$$

where $\mathrm{T_c}(i,j)$ is a scalar operator between two queries *i*-th and *j*-th for the Tanimoto coefficient, and e_i and e_j are unit vectors for the *i*-axis and *j*-axis < , > is the inner product.

Representative distributions, *Q* for target classes. In this section, we will obtain the representative distributions corresponding to each target class through GMM of ligand pair similarity. First, using the similarity matrix $\mathbf{T_M}[\mathbf{C_l}(m), \mathbf{C_l}(n)]_{ij}$ in (3.2), where m = n, we define the following univariate probability densities, $\Phi_n(x_k)$, by

$$\Phi_n(x_i)\delta x := \mathbb{P}\big(x_k \leq X = \mathbf{T_M}[\mathbf{C_l}(m), \mathbf{C_l}(n)]_{ij} \leq x_{k+1}\big), \quad (3.3)$$

where $\mathbb{P}$ is the probability measure, and $0 = x_0$ and $\mathrm{x}_{k+1} = x_k + \delta x$. Therefore, the probability densities, $\Phi_n(x)$, satisfy the following equation:

$$\textstyle\sum_{i=0}^{999} \Phi_n(x_i)\delta x = 1 \quad (3.4)$$

Second, to extract representative distributions from $\Phi_n(x)$, we utilize the Gaussian mixture model. In details, probability densities, $\Phi_n(x)$, are expressed as approximated from $\Xi_{EM}(\Phi_n(x)|\boldsymbol{G}, \omega, \mu, \sigma, K)$, which is the weighted sum of K univariate Gaussian distributions. That is,

$$\Xi_{EM}(\Phi_n(x)|g, \omega, \mu, \sigma, K) = \textstyle\sum_{k=1}^{K} \omega_k \, g(x; m_k, \sigma_k), \quad (3.5)$$

where $\omega_i, m_i,$ and σ_i are shown in Table 1. To estimate hyperparameters ω_i, m_i, *and* σ_i, we use the EM algorithm in the method section. Table 1 shows the mean, standard deviation, and weight corresponding to components in the mixture model. And Figure 2 depicts the difference between probability densities, $\Phi_n(x)$, and $\Xi_{EM}(\Phi_n(x)|g, \omega, \mu, \sigma, K)\ where\ K = 1, 3,$ and 7. When comparing component K, K =3 and 7 showed similarly fitted to histograms of raw data and normal Gaussian showed insufficient fitting in ESR, COX2, and CTSD (Figure 2). Commonly, mean and mode of the representative distributions existed near to 0.5 and every distribution was skewed to the right.

Gaussian distributions for queries. So far, we have built the representative distributions corresponding to ESR, VDR, COX2, and CTSD. To compare them quantitatively with the distributions of queries, let us introduce the Kullback-Leibler divergence. To calculate the K-L divergence, we need to build each distribution for each query.

For the purpose, we also call $\mathbf{T_M}[\mathbf{C_l}(m), \mathbf{C_l}(n)]$ of (3.2) in a similar way to the described method for representative distributions for target classes. When a query is *l*-th ligand of $\mathbf{C_l}(n)$, the *l*-th column's elements in the above matrix, $\mathbf{T_M}[\mathbf{C_l}(m), \mathbf{C_l}(n)]$ can be used for the *l*-th column vector, $\boldsymbol{\tau_m}(m, n, l)$, as

$$\boldsymbol{\tau_m}(m, n, l) := \mathbf{T_M}[\mathbf{C_l}(m), \mathbf{C_l}(n)]\mathbf{E}_l \qquad (3.6)$$

Where the value of E_l for $j = 1, 2, \ldots, \mathrm{N}$ are the $(\mathrm{N} \times \mathrm{N})$ matrices for which the elements $(\mathbf{E}_l)_{ij}$ satisfies

$$(\mathbf{E}_l)_{ij} := \begin{cases} 1, & if \quad i = j \\ 0, & otherwise \end{cases} \qquad (3.7).$$

Using the above vector, $\boldsymbol{\tau_m}(m, n, l)$ in (3.6), we define the following univariate probability densities, $\Phi_{mn}^{(l)}(x_k), as$

$$\Phi_{mn}^{(l)}(x_k)\delta x := \mathbb{P}\left(x_k \leq X = \left(\boldsymbol{\tau_m}(m, n, l)\right)_i \leq x_{k+1}\right) \qquad (3.8)$$

where the probability measure $\mathbb{P}$ is in (3.3).

Before obtaining the probability distribution, we make two assumptions. First, we assume that a distribution from one query is not a weighted sum of Gaussian distributions but a simple Gaussian distribution. It is reasonable that a distribution from one query is simpler than *Q* distribution of a target class having 14K queries. Second, to estimate the parameters in the Gaussian distribution, we choose the ML estimation, which is a general method in which

$$\Xi_{ML}\left(\Phi_{mn}^{(l)}(x_k)\middle| g, \omega, \mu, \sigma, 1\right) = g(x; \mu_1, \sigma_1) \quad (3.9)$$

where μ_1 and σ_1 are hyperparameters and are maximized log-likelihood functions for normal distribution, i.e.,

$$(\mu_1, \sigma_1) := \arg\max_{(\mu,\sigma)} \sum_{k=1}^{100} \frac{(x_k - \mu)^2}{\sigma^2} \qquad (3.10)$$

Using the above definition (3.9), (3.10), each query makes the four distributions corresponding to the 4 classes, ESR, VDR, COX2, and CTSD. For example, when CHEMBL539392 is chosen as a query (*l*) among ligands of ESR (class 1), we can obtain the distributions

$\Phi_{11}^{(l)}(x_k), \Phi_{12}^{(l)}(x_k), \Phi_{13}^{(l)}(x_k)\ and\ \Phi_{14}^{(l)}(x_k)$ under the definitions of (3.1) and (3.8). According to (3.9) and (3.10), four representative Gaussian distributions of the query, CHEMBL539392 were acquired from the column vector between CHEMBL539392 and 14K ligands of each class, which are

$$\begin{cases} \Xi_{ML}\left(\Phi_{11}^{(l)}(x_k)\middle|g,\omega,\mu,\sigma,1\right) = g(x; 0.24055, 0.07472\ ), \\ \Xi_{ML}\left(\Phi_{12}^{(l)}(x_k)\middle|g,\omega,\mu,\sigma,1\right) = g(x; 0.21976{,}0.06466\ ), \\ \Xi_{ML}\left(\Phi_{13}^{(l)}(x_k)\middle|g,\omega,\mu,\sigma,1\right) = g(x; 0.24389{,}0.04857\ ), \\ \Xi_{ML}\left(\Phi_{14}^{(l)}(x_k)\middle|g,\omega,\mu,\sigma,1\right) = g(x; 0.21187{,}0.06631\ \ ), \end{cases} \quad \text{for } k = 0, 1, \dots, 99. \tag{3.11}$$

In the same way, we can obtain univariate normal distributions of all queries in each class. Since the number of classes is four and there are 14K queries in each class, the Gaussian distributions, $\mathrm{G}(\mathrm{x}; \mu_1, \sigma_1)$, from $\Xi_{ML}\left(\Phi_{mn}^{(l)}(x_k)\middle|g,\omega,\mu,\sigma,1\right)$ presented the class number, either *m* or *n,* is an integer between 1 and 4 and the query number, *l,* is an integer from 1 to 14000. As the result, the frequency distributions of the estimates, mean (μ_1) and standard deviation (σ_1) were described in Figure 3, 4, 5, and 6. In the view of frequency, ML estimation can't show any difference between self-query (m = n) and cross-query (m ≠ n). Even though cathepsin D (CTSD) showed slightly lower mean than other classes, the self-comparison also showed low mean in Figure 6. Regardless of a class or a query (self/cross), 3D-similarity of ligand pairs within a class showed the mode near to 0.6 so that the result reminded us why the quantitative comparison between queries is required. Notably, univariate probability distributions of 3D-similarity, themselves cannot discriminate target class at all.

Discrimination and K-L divergence. In sequence, 3D-similarity distributions of target classes and queries are quantitatively compared through K-L divergence calculation. First, the information describing specific Tanimoto-Jaccard coefficients, *x,* can be written as

$$\ln\left(\frac{\Xi_{ML}\left(\Phi_{mn}^{(l)}(x)\middle|g,\omega,\mu,\sigma,1\right)}{\Xi_{EM}\left(\Phi_n(x)\middle|g,\omega,\mu,\sigma,K\right)}\right) \qquad (3.12)$$

from two probability density distributions, $\Xi_{ML}\left(\Phi_{mn}^{(l)}(x)\middle|g,\omega,\mu,\sigma,1\right)$ and $\Xi_{EM}(\Phi_n(x)|g,\omega,\mu,\sigma,K)$, which are generated from a query and a class. Hence, following the expected value from the above information (3.12) with respect to one query, the K-L divergence,

$$D\left(\Xi_{ML}\left(\phi_{mn}^{(l)}(x)\middle|g,\omega,\mu,\sigma,1\right)\middle\|\Xi_{KL}\left(\phi_n(x)\middle|g,\omega,\mu,\sigma,K\right)\right)$$

$$=\int\Xi_{ML}\left(\phi_{mn}^{(l)}(x)\middle|g,\omega,\mu,\sigma,1\right)\ln\left(\frac{\Xi_{ML}\left(\phi_{mn}^{(l)}(x)\middle|g,\omega,\mu,\sigma,1\right)}{\Xi_{EM}\left(\phi_n(x)\middle|g,\omega,\mu,\sigma,K\right)}\right)dx \qquad (3.13)$$

is a measurement for the discrimination.

In the GMM with one component (K = 1), the K-L divergence between Gaussian distributions of every query and Q distributions (Table 1) were calculated and randomly chosen queries were described in Tables 2. To show the calculation process in detail, let us consider the example where a query is the CHEMBL539392. Using the above equation for Kullback-Leibler divergence between normal distributions (2.9),

$$D\left(G(x;m_i,\sigma_i)\middle\|G(x;m_j,\sigma_j)\right)=\ln\left(\frac{\sigma_j}{\sigma_i}\right)+\frac{(\sigma_i)^2+(m_i-m_j)^2}{2(\sigma_j)^2}-\frac{1}{2} \qquad (3.14)$$

where

$$\begin{cases}G(x;m_i,\sigma_i)=\Xi_{ML}(\phi_{1n}^{(1)}(x)|g,\omega,\mu,\sigma,1)\\ G(x;m_j,\sigma_j)=\Xi_{EM}(\boldsymbol{\phi}_n(x)|g,\omega,\mu,\sigma,1)\end{cases} \qquad (3.15)$$

We obtain the four values.

① When n = 1, i.e., the class is ESR, the parameters are as follows

$$m_i = 0.24055,\ \sigma_i = 0.07472,\ m_j = 0.5483,\ \sigma_j = 0.1458 \qquad (3.16)$$

and we then obtain the following K-L divergence:

$$D\big(G(x; m_i, \sigma_i)\big\|G\big(x; m_j, \sigma_j\big)\big) = 2.1493 \qquad (3.17)$$

② When $n = 2$, i.e., the class is VDR, the parameters are as follows

$$m_i = 0.21976,\ \sigma_i = 0.06466,\ m_j = 0.5981,\ \sigma_j = 0.1224 \qquad (3.18)$$

and we then obtain the following K-L divergence:

$$D\big(G(x; m_i, \sigma_i)\big\|G\big(x; m_j, \sigma_j\big)\big) = 4.6939 \qquad (3.19)$$

③ When $n = 3$, i.e., class is Cyclooxygenase-2, the parameters are follows,

$$m_i = 0.24389,\ \sigma_i = 0.04857,\ m_j = 0.5941\ \sigma_j = 0.1758 \qquad (3.20)$$

and then we obtain the following K-L divergence:

$$D\big(G(x; m_i, \sigma_i)\big\|G\big(x; m_j, \sigma_j\big)\big) = 2.0810 \qquad (3.21)$$

④ When n = 4, i.e., the class is CTSD, the parameters are as follows

$$m_i = 0.21187,\ \sigma_i = 0.06631,\ m_j = 0.4560,\ \sigma_j = 0.1320 \qquad (3.22)$$

and we then obtain the following K-L divergence:

$$D\big(G(x; m_i, \sigma_i)\big\|G\big(x; m_j, \sigma_j\big)\big) = 1.6354 \qquad (3.23)$$

As shown in the table 2, K-L divergence of every query could not always be the smallest value in their original targets annotated by ChEMBL DB. Even though considerable number of queries could show that K-L divergence resulting from an original target are smaller than the values from other target classes, CHEMBL539392 of ESR, CHEMBL1163237 of COX2, and CHEMBL263810 of CTSD assigned less difference into another target to give false prediction (Table2). When we counted the queries discriminating between the original targets and other targets from each 14K queries of 4 classes under GMM (K = 1), the right predicted numbers were 6.3K, 5.2K, 4.1K, and 6.4K among each 14K queries respectively (the order: ESR, VDR, COX2, and CTSD). When applying GMM (K = 3) and (K = 7) for Q distributions, the true positive ratio decreased (ESR: 5.1K, VDR: 4.5K, COX2: 3.2K, and CTSD: 4.9K in K

= 3; ESR: 4.9K, VDR: 4.5K, COX2: 3.1K, and CTSD: 4.8K in K = 7).

After the indivisual K-L divergence comparison of each compound, discriminativeness between target classes need to be quntitifyed. In sequence, The K-L divergence between Gaussian distributions of 14K queries and Q distributions (K =1,3, and 7) for the four target classes were presented as a cumulative distribution from the Figure 7 to Figure 10. To investigate the feasibility of the information, let us define following distribution,

$$\mathbb{P}(\nu(l_m) = i) \text{ for } i = 1, 2, 3, 4, \qquad (3.24)$$

where l_m is the query number in class m and the random variable $\nu(l_m)$ is a class number such that

$$\nu(l_m) \coloneqq \arg\min_{n}\{D\{\ \Xi_{ML}\left(\phi_{mn}^{l_m}(x)\middle|g,\omega,\mu,\sigma,1\right)\|\ \Xi_{EM}\left(\Phi_n\ (x)\middle|g,\omega,\mu,\sigma,1\right)\}|\ 1 \leq \mathrm{n} \leq 4, 1 \leq \mathrm{l}_m \leq 14\mathrm{K}\}. \ (3.25)$$

If the K-L divergence (3.13) is an ideal measurement for discrimination between target classes, $(\nu(l_m) = i)$ should satisfy the following conditions

- Necessary condition : $\mathbb{P}(\nu(l_m) = m) \geq \max_{i \neq m} \mathbb{P}(\nu(l_m) = i)$
- Sufficient condition : Let us define the feasibility index, F_m, such that

$$\mathrm{F}_m \coloneqq \sqrt{\frac{\mathbb{P}(\nu(l_m)=m)}{1-\mathbb{P}(\nu(l_m)=m)}} \geq 1 \qquad (3.26)$$

The above conditions imply a quantitative measurement for the discrimination. In particular, F_m in the sufficient condition represents the ratio between two probabilities, which are that a query belongs to the class of itself and the query belongs to other classes. A larger value of F_m indicates better feasibility or resolution for the discrimination. The below Table 3 depicts the probability of K-L divergence $\mathbb{P}(\nu(l_m) = i)$ for $1 \leq \mathrm{i}, m \leq 4$, and shows that, except for the example m=3 where the class is COX2, tested classes meet the necessary condition $\mathbb{P}(\nu(l_m) = m) \geq \max_{i \neq m} \mathbb{P}(\nu(l_m) = i)$ in 3.3. With respect to feasibility index, F_m in 3.3, it is easiest to distinguish a query in the class CTSD, where m =4, from in in every classes except itself (Figure 11). When the feasibility index resulting from the GMM (K = 1) was compared with the index calculated from GMM (K = 3) and (K = 7) for Q distributions, GMM (K = 1) showed superior feasibility for the class discrimination to GMM (K = 3) or (K = 7) as shown in Table 3.

A representative of ligands for better discriminative prediction. According to described results (Figure from 7 to 11 and Table 2 and 3), 3D-similarity based K-L divergence together with $\mathbb{P}(\nu(l_m) = m)$ and $\mathrm{F}_m$showed discriminative power on some 'query – class'. Then, how can we use the 3D-chemocentric approach efficiently under the current discriminative power? Notably, it is applicable for the investigation on the novel pharmacology of an unprecedented compound. For the purpose, K-L divergence of an unprecedented compound should be calculated for the comparison with known ligands and target classes. In detail, a representative of ligands within each target class can be chosen for the comparison. For example, we selected the four type representatives: (1) mean of Q-distribution (GMM, K = 1), (2) an outlier of Q-distribution (Mean ± 2SD), (3) the biggest gap of K-L divergence between two target classes, and (4) the highest similarity with unprecedented compound (Table 4). As an example, BNDS-A, recently reported in-house compound [7], was used as the unprecedented compound due to the absence of ChEMBL DB. The first type query near to mean of Q-distribution could show smaller K-L divergence rather than other queries (Table 4). Initial assumption, initial selection of the target class of BNDS-A (in other words, selection of Q-distribution), made the critical effect on K-L divergence of BNDS-A as a query for the prediction of target class. When ESR was assumed as the initial target of BNDS-A, BNDS-A was more ESR ligand like than CHEMBL558943 (at 'Mean − 2SD' of ESR Q-distribution) and CHEMBL604989 (having biggest K-L divergence gap), and was less ESR like than CHEMBL499809 (at Mean of ESR

Q-distribution) and CHEMBL2 (at Mean + 2SD). Under the assumption (*Q* of ESR), BNDS-A showed lowest KL divergence with VDR ligands (0.0588 of VDR < 0.2116 of ESR) to suggest VDR ligand-like more than ESR ligand-like. When the initial target was transferred to VDR or COX2, BNDS-A showed lowest K-L divergence to satisfy the assumption (chosen *Q*). In particular, BNDS-A was more VDR ligand-like than representative ligands. Experimentally, BNDS-A concentration-dependently regulated the expression level of targets (VDR > CTSD >> ESR) [7]. Based on K-L divergence and acquired experiments, COX2 is enough reasonable target for testing BNDS-A. In addition, the 3D-similarity based K-L divergence (between a query and a class) is superior comparison to (1) 3D-similarity score between a query and a ligand or (2) univariate distribution of 3D-similarity described in the previous study of BNDS-A [7]. Such as the example of BNDS-A, whenever getting relevance between a novel query and a target class, K-L divergence can be called for 3D-chemocentric informatics.

■ Conclusions

We proposed the K-L divergence measurement of 3D-similarity information for discriminative prediction in chemocentric informatics. Since *Q* distributions of target classes could be compared with Gaussian distribution of query pairs, the K-L divergence of any unknown query could be applicable with the squared number of classes to suggest the best class. The feasibility index, F_m from K-L divergence could show us the discriminativeness between target classes. In this study, CTSD could show the most desirable feasibility and COX2 indicated less desirable target for chemocentric informatics. Although the feasibility also depends on fitting model (eg. *K* number of GMM), the order of feasibilities retained regardless of fitting models. The study could contribute to the 3D-chemocentric target deconvolution for unprecedented drug scaffolds. In the recent future, we hope that the quantitative method will apply to the chemical optimization between chemical space and pharmacological space with further study.

Declarations

Availability of data and materials

The method and data is available at https://github.com/college-of-pharmacy-gachon-university/KLD-Pharmacological_Class_Similarity

Competing interests

The authors declare that they have no competing interests.

Funding

This study was supported by the Basic Science Research Program of the National Research Foundation of Korea (NRF), funded by the Ministry of Education, Science, and Technology (No.: 2017R1E1A1A01076642) and by the Agency for Defense Development (Grant ID: PD1806130GD).

Acknowledgements

Authors would like to thank OpenEye Scientific Software providing the academic free license.

Authors' contributions

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

References

[1] Hawkins, P.C.D.; Skillman, A.G.; Nicholls A. Comparison of Shape-Matching and Docking as Virtual Screening Tools. *J. Med. Chem.* **2007**, *50*, 74-82.

[2] Gadhe, C.G.; Lee, E.H.; Kim, M.-h. Finding New Scaffolds of JAK3 Inhibitors in Public Database: 3D-QSAR Models & Shape-Based Screening. *Arch. Pharmacal Res.* **2015**, *38*,

2008-2019.

[3] Keiser, M.J.; Roth, B.L.; Armbruster, B.N.; Ernsberger, P.; Irwin, J.J.; Shoichet, B.K., Relating protein pharmacology by ligand chemistry. *Nat. Biotechnol.* **2007**, *25*, 197-206.

[4] Eckert, H.; Bajorath, J. Molecular Similarity Analysis in Virtual Screening: Foundations, Limitations and Novel Approaches. *Drug Discovery Today* **2007**, *12*, 225-233.

[5] Year, E.R.; Cleves, A.E.; Jain, A.N. Chemical Structural Novelty: On-Targets and Off-Targets. *J. Med. Chem.* **2011**, *54*, 6771-6785.

[6] Taylor, R.D.; MacCoss, M.; Lawson, A.D., Rings in drugs: Miniperspective. *J. Med. Chem.* **2014**, 57(14), pp.5845-5859.

[7] Venkanna, A.; Kwon, O. W.; Afzal, S.; Jang, C.; Cho, K.; Yadav, D. K.; Kim, K.; Park, H.-g.; Chun, K.-H.; Kim, S. Y.; Kim, M.-h. Pharmacological Use of a Novel Scaffold, Anomeric *N,N*-Diarylamino Tetrahydropyran: Molecular Similarity Search, Chemocentric Target Profiling, and Experimental Evidence. *Sci. Rep.* **2017**, *7*, 12535.

[8] Afzal, S.; Venkanna, A.; Park, H.-g.; Kim, M.-h. Metal-Free α-C(sp3)−H Functionalized Oxidative Cyclization of Tertiary N,N-Diarylamino Alcohols: Construction of N,N-Diarylaminotetrahydropyran Scaffolds. *Asian J. Org. Chem.* **2016**, *5*, 232-239.

[9] Venkanna, A.; Cho, K.; Dorma, L. P.; Kumar, D.N.; Hah, J.M.; Park, H.-g.; Kim, S. Y.; Kim, M.-h. Chemistry-Oriented Synthesis (ChOS) and Target Deconvolution on Neuroprotective Effect of a Novel Scaffold, Oxaza Spiroquinone. *Eur. J. Med. Chem.* **2019**, *163*, 453-480.

[10] Hu. G.; Kuang, G.; Xiao, W.; Li, W.; Liu, G.; Tang, Y. Performance Evaluation of 2D Fingerprint and 3D Shape Similarity Methods in Virtual Screening. *J. Chem. Inf. Model.* **2012**, *52*, 1103-1113.

[11] Vilar, S.; Hripcsak, G. Leveraging 3D Chemical Similarity, Target and Phenotypic Data in the Identification of Drug-Protein and Drug-Adverse Effect Associations. *J. Cheminf.* **2016**,

8, 35

[12] Pacureanu, L.; Avram, S.; Bora, A.; Kurunczi, L.; Crisan, L. Portraying the Selectivity of GSK-3 Inhibitors towards CDK-2 by 3D Similarity and Molecular Docking. *Struct. Chem.* **2018**, *13*.

[13] Keiser, M.J.; Roth, B.L.; Armbruster, B.N.; Ernsberger, P.; Irwin, J.J.; Shoichet, B.K. Relating Protein Pharmacology by Ligand Chemistry. *Nat. Biotechnol.* **2007**, *25,* 197.

[14] Pérez-Nueno, V.; Venkatraman, V.; Mavridis, L.; Ritchie, D. Detecting Drug Promiscuity Using Gaussian Ensemble Screening. *J. Chem. Inf. Model.* **2012**, *52*, 1948-1961.

[15] Baldi, P.; Nasr, R. When is Chemical Similarity Significant? The Statistical Distribution of Chemical Similarity Scores and its Extreme Values. *J. Chem. Inf. Model.* **2010**, *50*, 1205-1222.

[16] Kim, H.R.; Jang, C.Y.; Yadav, D.K.; Kim, M.-h. The Comparison of Automated Clustering Algorithms for Resampling Representative Conformer Ensembles with RMSD Matrix. *J. Cheminf.* **2017**, *21,* 9.

[17] Maaten, L.J.P.V.D.; Hinton, G.E. Visualizing Data Using t-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579-2605.

[18] Hershey, J.R.; Olsen, P.A. Approximating the Kullback Leibler Divergence Between Gaussian Mixture Models. *Acoustics, Speech and Signal Processing, IEEE International Conference on, IEEE* **2007**.

[19] Kullback, S.; Leibler, R.A. On Information and Sufficiency. *Ann. Math. Stat.* **1951**, *22*, 79-86.

[20] Burnham, K.P.; Anderson, D.R. Kullback-Leibler Information as a Basis for Strong Inference in Ecological Studies. *Wildl. Res.* **2001**, *28*, 111-119.

[21] Nalewajski, R.F.; Parr, R.G. Information Theory, Atoms in Molecules, and Molecular Similarity. *Proc. Natl. Acad. Sci. U. S. A.* **2000**, *97*, 8879-8882.

[22] Vogt, M., Bajorath, J. Introduction of an Information-Theoretic Method to Predict Recovery Rates of Active Compounds for Bayesian in Silico Screening: Theory and Screening Trials. *J. Chem. Inf. Model.* **2007**, *47*, 337-341.

[23] Koller, D.; Sahami, M. Toward Optimal Feature Selection. *Proceedings of the Thirteenth International Conference on International Conference on Machine Learning*: **1996**, 284-292.

[24] Duchi, J. Derivations for Linear Algebra and Optimization. *Berkeley California*, **2007**, *3*.

[25] McLachlan, G.J.; McGiffin, D.C. On the Role of Finite Mixture Models in Survival Analysis. *Stat. Methods. Med. Res.* **1994**, *3*, 211-226.

[26] Singh, R.; Pal, B.C.; Jabr, R.A. Statistical Representation of Distribution System Loads using Gaussian Mixture Model. *IEEE T. Power Syst.* **2010**, *25*, 29-37.

[27] Duda, R.O.; Hart, P.E.; Stork, D.G. Pattern Classification and Scene Analysis 2nd ed. *WILEY INTERSCIENCE PUBLICATION,* **1995**.

[28] Dempster, A.P.; Laird, N.M.; Rubin, D.B. Maximum Likelihood from Incomplete Data via the EM Algorithm. *J. R. Stat. Soc. Ser. B Methodol.* **1977**, 1-38.

[29] Hartley, H.O. Maximum Likelihood Estimation from Incomplete Data. *Biometrics* **1958**, *14*, 174-194.

[30] McLachlan, G.; Krishnan, T. The EM Algorithm and Extensions (Vol. 382). John Wiley & Sons, Hoboken, New Jersey, **2007**.

[31] Bajusz, D.; Rácz, A.; Héberger, K. Why is Tanimoto Index an Appropriate Choice for Fingerprint-Based Similarity Calculations? *J. Cheminf.* **2015,** *20*, 1-13.

[32] Piaggi, P. M.; Parrinello, M. "Predicting Polymorphism in Molecular Crystals Using Orientational Entropy." *Proc. Natl. Acad. Sci. U. S. A.* **2018,** *115*, 10251-10256.

[33] Lemey, P.; Rambaut, A.; Drummond, A.J.; Suchard, M.A. Bayesian Phylogeography Finds its Roots. *PLOS Comput. Biol.* **2009,** *5*, 1-16.

[34] Kümmerer, M.; Wallis, T.S.A.; Bethge, M. Information-Theoretic Model Comparison

Unifies Saliency Metrics. *Proc. Natl. Acad. Sci. U. S. A.* **2015,** *112*, 16054-16059.

[35] Grant, J.A.; Gallardo, M.A.; Pickup, B.T. A. Fast Method of Molecular Shape Comparison: A Simple Application of a Gaussian Description of Molecular Shape. *J. Comput. Chem.* **1996,** *17*, 1653-1666.

Figure Legend

Figure 1. The problem definition of 3D chemo-centric Retro-VS. (a) The role of chemical similarity in virtual screening, (b) The role of chemical similarity in retro-virtual screening, and (c) Target screening of an unprecedented drug scaffold as a retro-virtual screening.

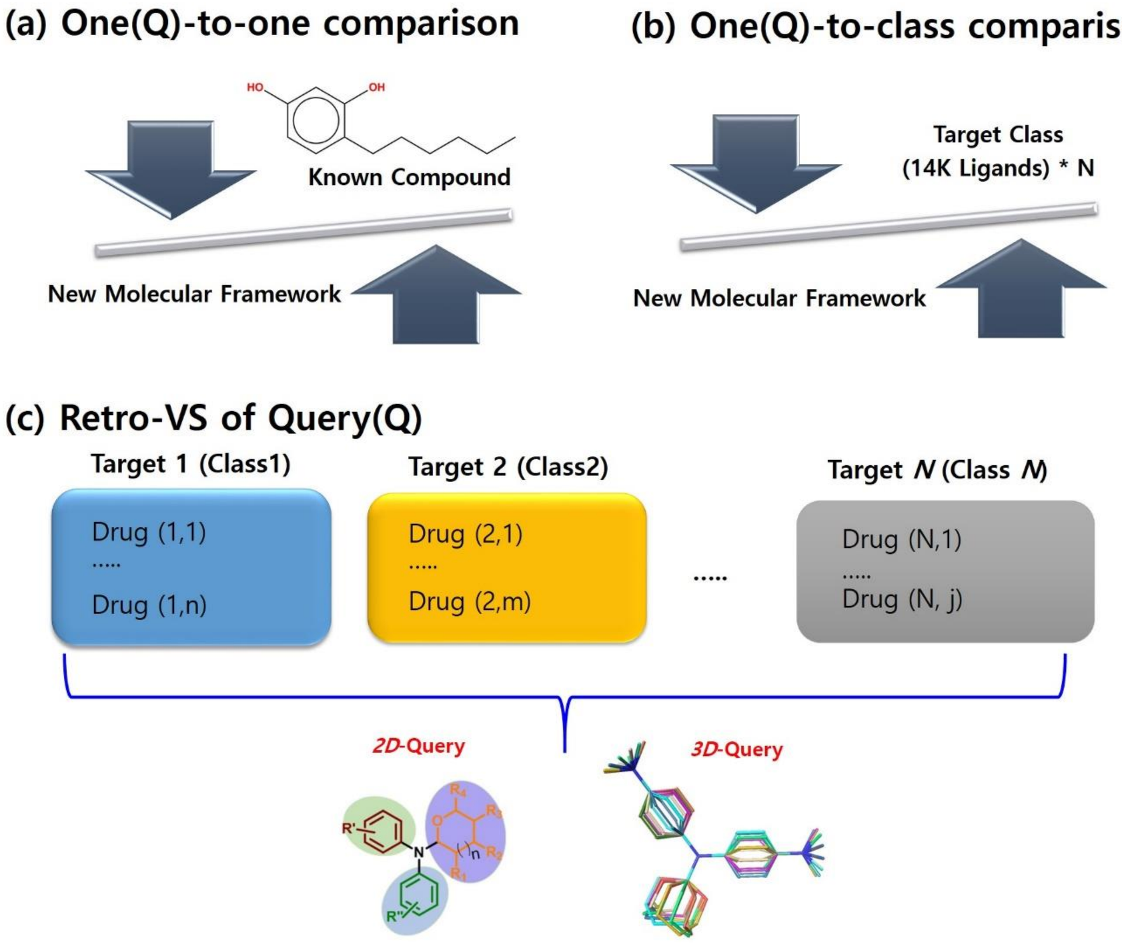
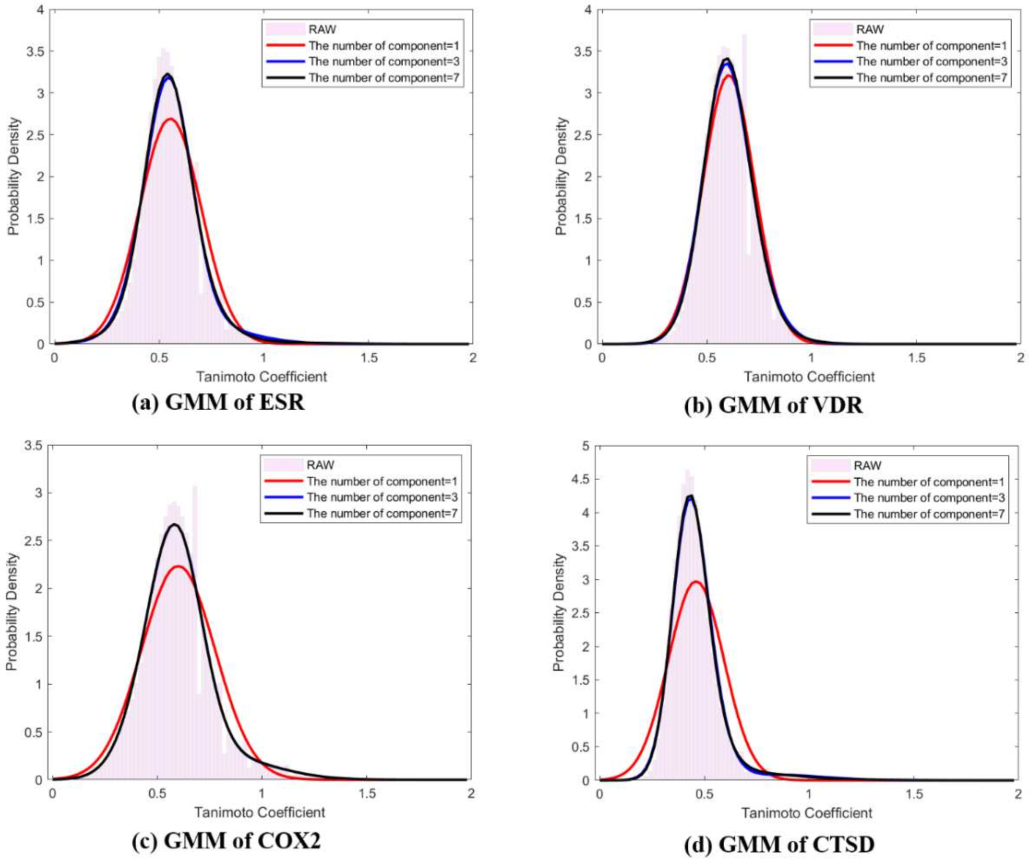


Figure 2. Representative distributions (*Q*-distributions) of target classes using EM based Gaussian Mixture model $(\Xi_{EM}(\Phi_n(x)|g,\omega,\mu,\sigma,K)$ of ligand pair similarity. The red line: GMM *K* = 1, blue line: GMM *K* = 3, black line: GMM *K* = 7, pink bar: histogram of raw data.*



(a) GMM of ESR

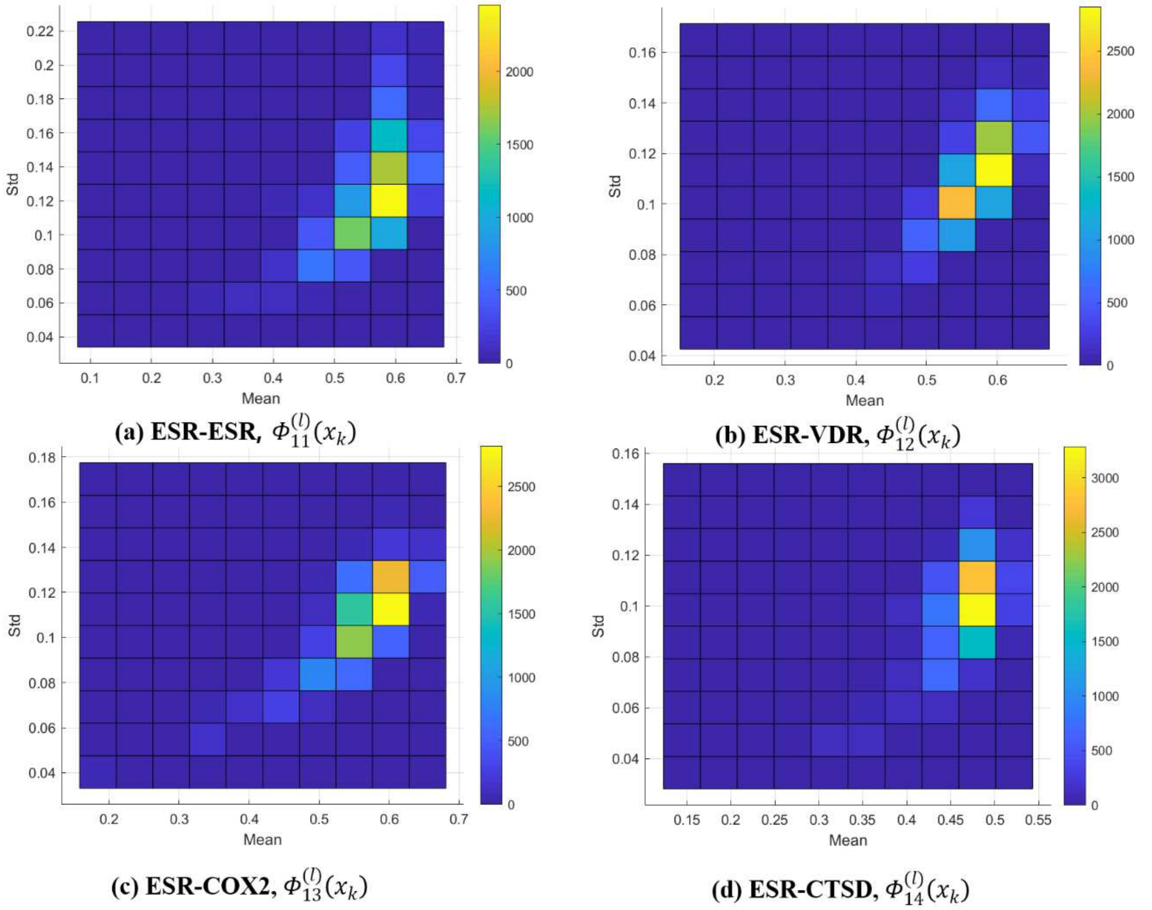
(b) GMM of VDR

(c) GMM of COX2

(d) GMM of CTSD

*Abbreviation: Estrogen receptor alpha (ESR), Vitamin D receptor (VDR), Cyclooxygenase-2 (COX2) and Cathepsin D (CTSD).

Figure3. Frequency distributions of $\Xi_{ML}\left(\Phi_{1n}^{(l)}(x_k)\middle|g,\omega,\mu,\sigma,1\right)$ estimates ($\mu_1$and σ_1). Query (l) ∊ ESR (class = 1).



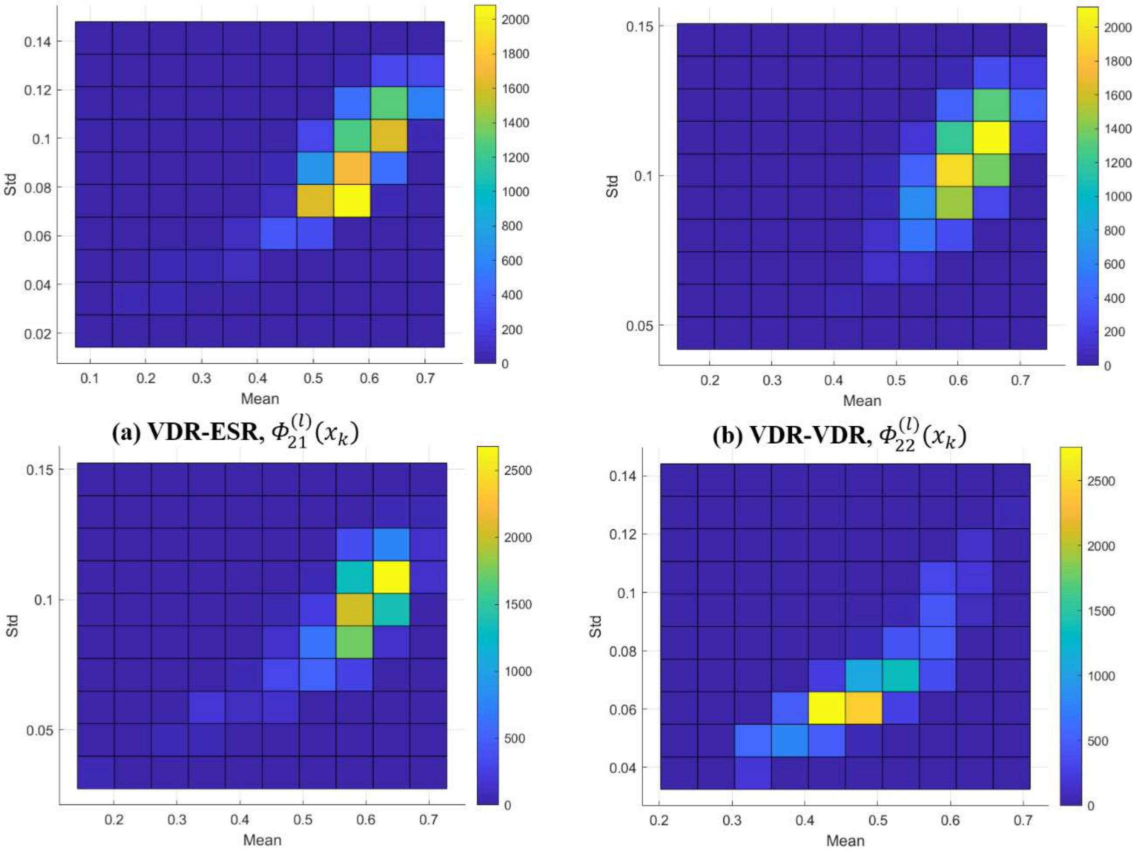
(a) ESR-ESR, $\Phi_{11}^{(l)}(x_k)$

(b) ESR-VDR, $\Phi_{12}^{(l)}(x_k)$

(c) ESR-COX2, $\Phi_{13}^{(l)}(x_k)$

(d) ESR-CTSD, $\Phi_{14}^{(l)}(x_k)$

Figure 4. Frequency distributions of $\Xi_{ML}\left(\Phi_{2n}^{(l)}(x_k)\middle|g,\omega,\mu,\sigma,1\right)$ estimates ($\mu_1$and σ_1). Query (l) ∈ VDR (class = 2).



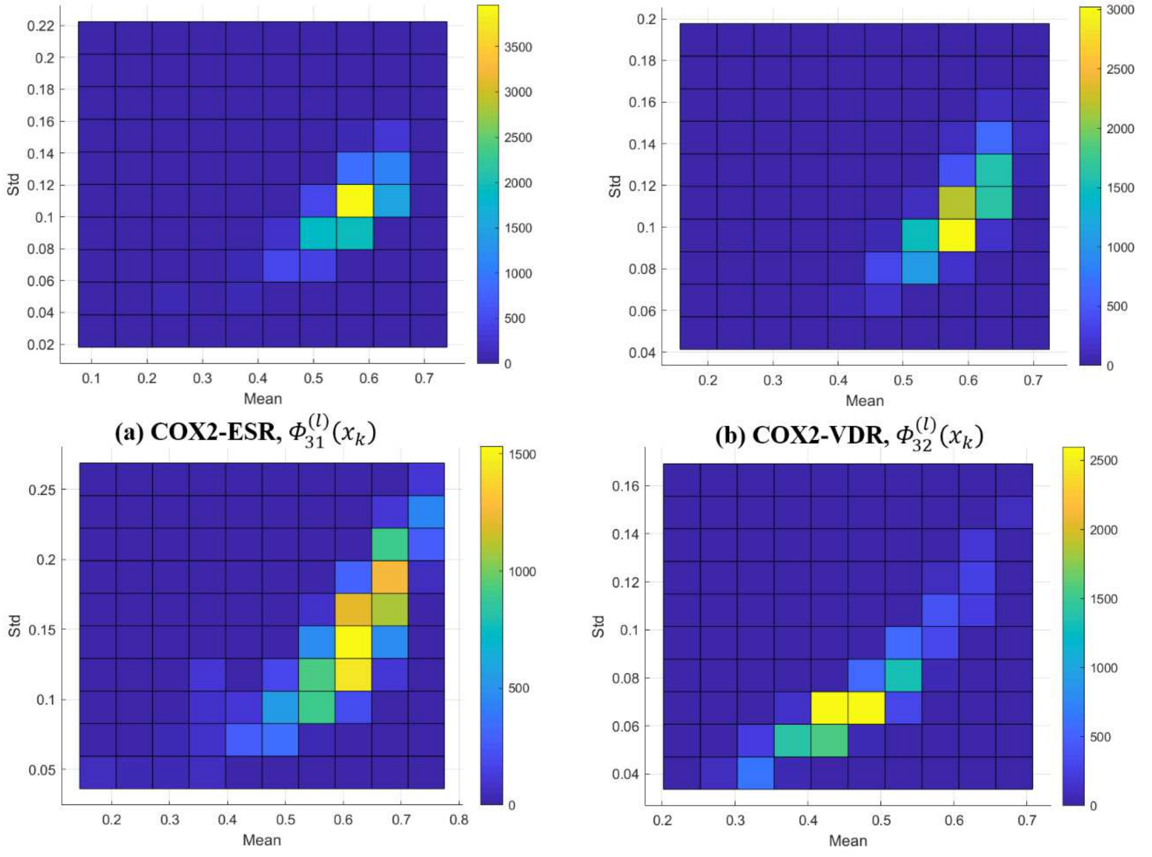
(a) VDR-ESR, $\Phi_{21}^{(l)}(x_k)$

(b) VDR-VDR, $\Phi_{22}^{(l)}(x_k)$

(c) VDR-COX2, $\Phi_{23}^{(l)}(x_k)$

(d) VDR-CTSD, $\Phi_{24}^{(l)}(x_k)$

Figure 5. Frequency distributions of $\Xi_{ML}\left(\Phi_{3n}^{(l)}(x_k)\middle|g,\omega,\mu,\sigma,1\right)$ estimates ($\mu_1$and σ_1). Query (l) ∈ COX2 (class = 3).



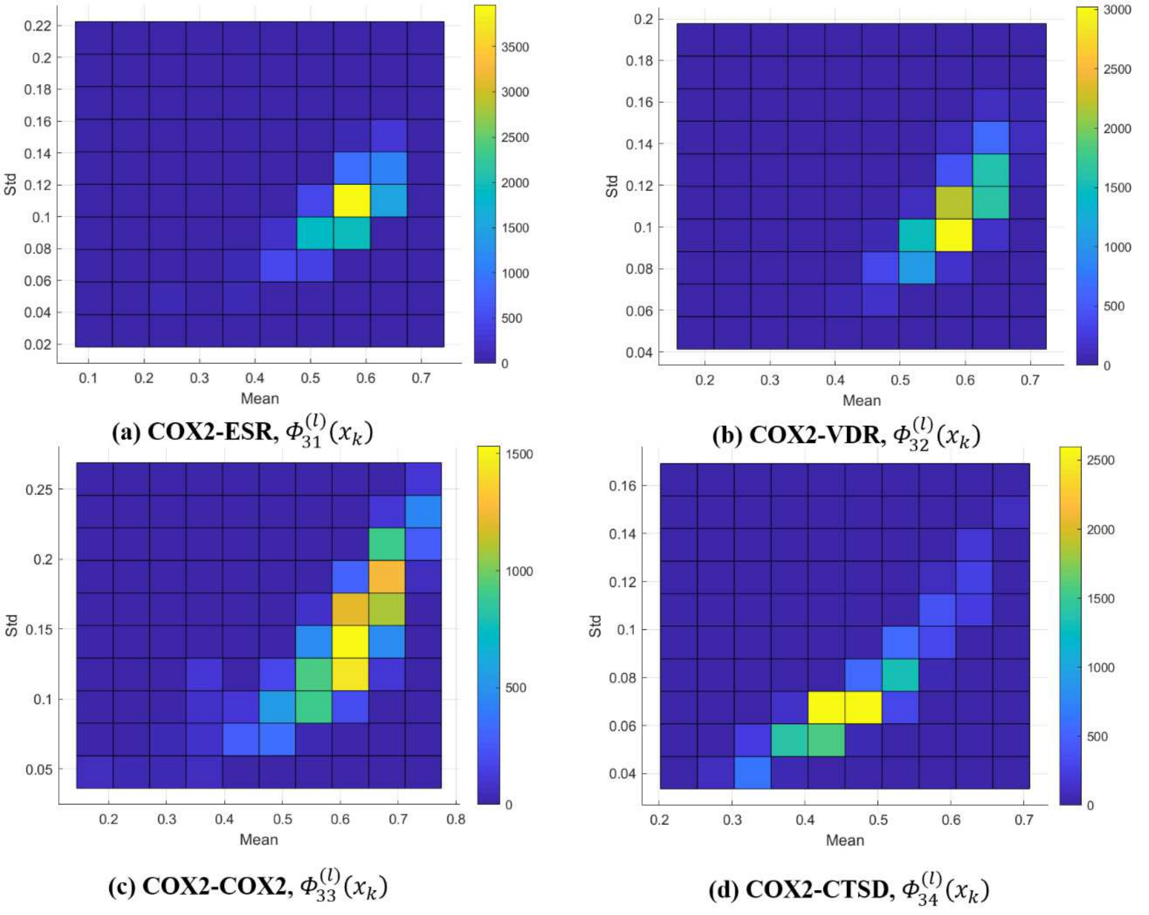
(a) COX2-ESR, $\Phi_{31}^{(l)}(x_k)$

(b) COX2-VDR, $\Phi_{32}^{(l)}(x_k)$

(c) COX2-COX2, $\Phi_{33}^{(l)}(x_k)$

(d) COX2-CTSD, $\Phi_{34}^{(l)}(x_k)$

Figure 6. Frequency distributions of $\Xi_{ML}\left(\Phi_{4n}^{(l)}(x_k)\middle|g,\omega,\mu,\sigma,1\right)$ estimates ($\mu_1$and σ_1). Query (l) ∈ CTSD (class = 4).



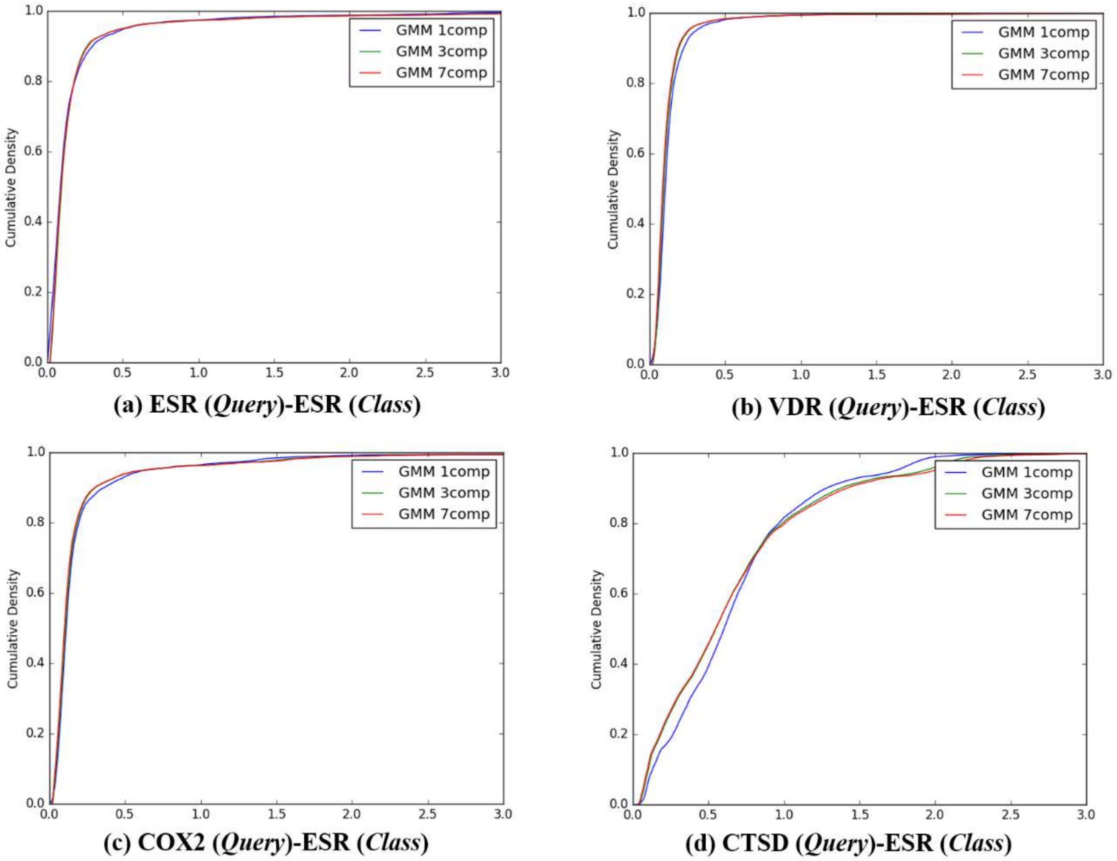
(a) COX2-ESR, $\Phi_{31}^{(l)}(x_k)$

(b) COX2-VDR, $\Phi_{32}^{(l)}(x_k)$

(c) COX2-COX2, $\Phi_{33}^{(l)}(x_k)$

(d) COX2-CTSD, $\Phi_{34}^{(l)}(x_k)$

Figure 7. The cumulative densities of K-L distance between *Q*-distribution (Target class: ESR) and queries.*



(a) ESR (*Query*)-ESR (*Class*)

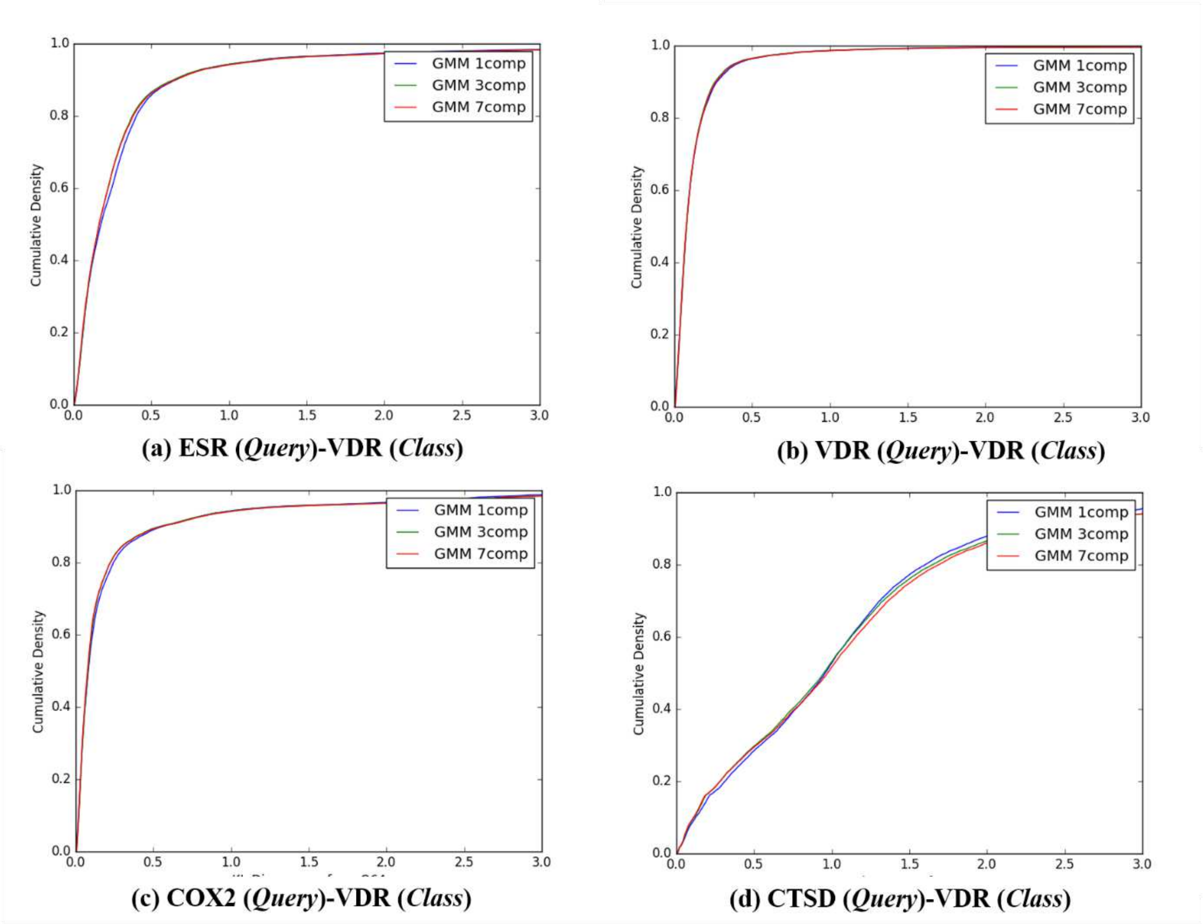
(b) VDR (*Query*)-ESR (*Class*)

(c) COX2 (*Query*)-ESR (*Class*)

(d) CTSD (*Query*)-ESR (*Class*)

*X-axis: K-L divergence, Y-axis: cumulative density; *Q*-distribution of ESR through GMM and distribution of queries were calculated.

Figure 8. The cumulative densities of K-L distance between *Q*-distribution (Target class: VDR) and queries.*



(a) ESR (*Query*)-VDR (*Class*)

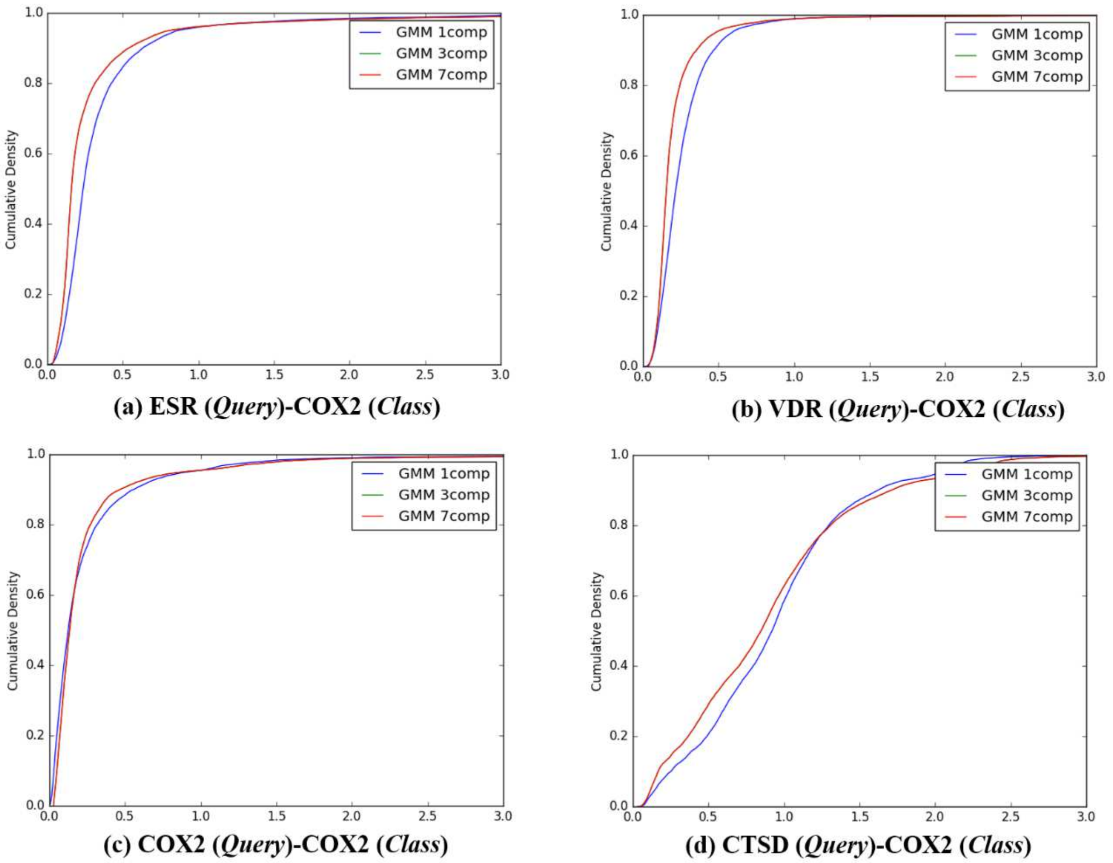
(b) VDR (*Query*)-VDR (*Class*)

(c) COX2 (*Query*)-VDR (*Class*)

(d) CTSD (*Query*)-VDR (*Class*)

*X-axis: K-L divergence, Y-axis: cumulative density; *Q*-distribution of VDR through GMM and distribution of queries were calculated.

Figure 9. The cumulative densities of K-L distance between *Q*-distribution (Target class: COX2) and queries.*



(a) ESR (*Query*)-COX2 (*Class*)

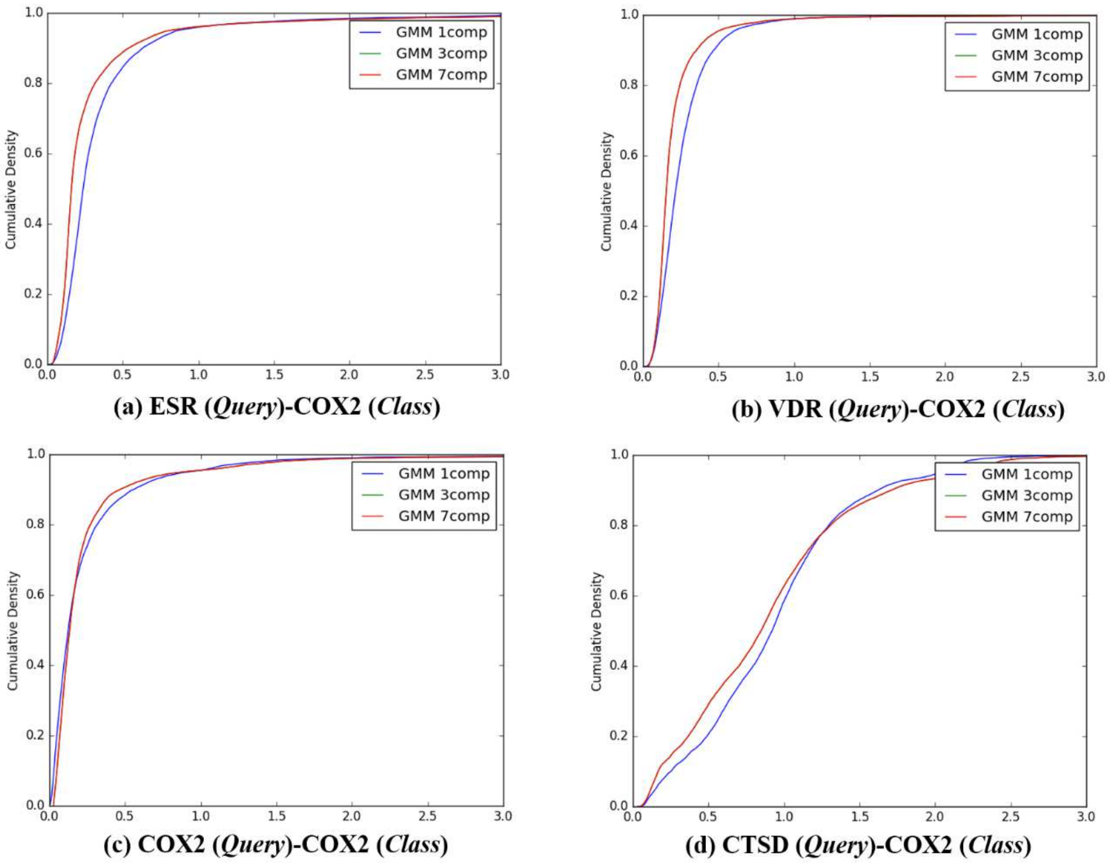
(b) VDR (*Query*)-COX2 (*Class*)

(c) COX2 (*Query*)-COX2 (*Class*)

(d) CTSD (*Query*)-COX2 (*Class*)

*X-axis: K-L divergence, Y-axis: cumulative density; *Q*-distribution of COX2 through GMM and distribution of queries were calculated.

Figure 10. The cumulative densities of K-L distance between *Q*-distribution (Target class:CTSD) and queries.*



(a) ESR (*Query*)-COX2 (*Class*)

(b) VDR (*Query*)-COX2 (*Class*)

(c) COX2 (*Query*)-COX2 (*Class*)

(d) CTSD (*Query*)-COX2 (*Class*)

*X-axis: K-L divergence, Y-axis: cumulative density; *Q*-distribution of CTSD through GMM and distribution of queries were calculated.

Figure 11. Feasibility index according to target class and GMM component (*K*).

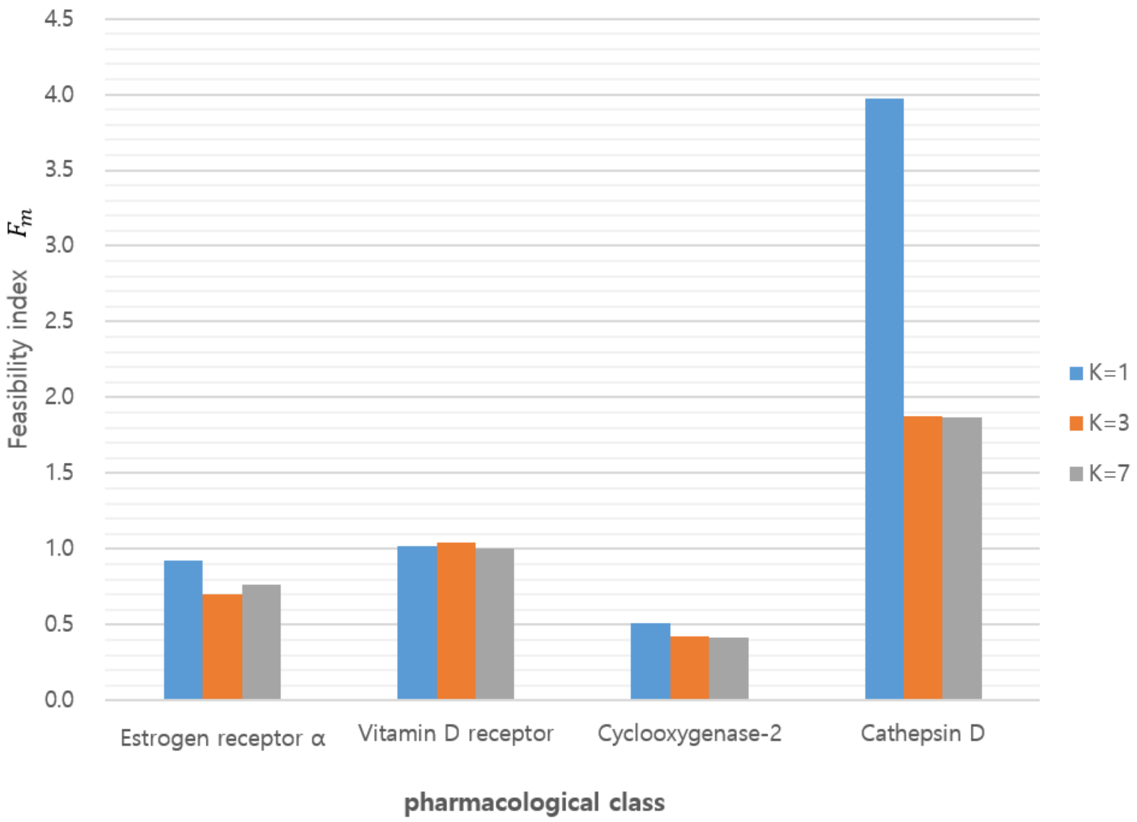
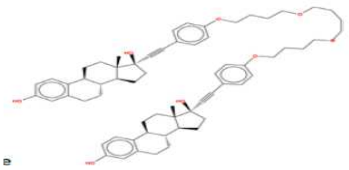


Table legends

Table 1. Hyperparameters of *Q* distributions for target classes

GMM	**ESR**		**VDR**		**COX2**		**CTSD**	
No(i)	m_i	σ_i	m_i	σ_i	m_i	σ_i	m_i	σ_i
1	0.5483	0.1458	0.5981	0.1224	0.5941	0.1758	0.4560	0.1320

Table2. K-L divergence of randomly chosen queries between *Q* distributions and the distributions of queries

Class	Query	Chemical Structure	K-L Divergence			
			ESR	VDR	COX2	CTSD
ESR	CHEMBL 539392		2.6310	5.2420	2.9952	1.9426

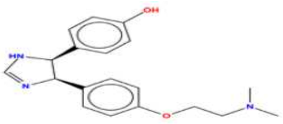
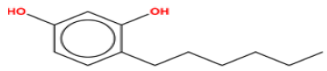
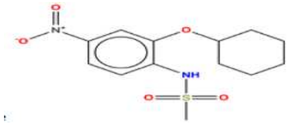
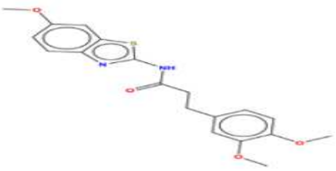
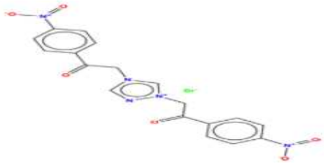
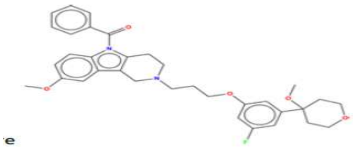
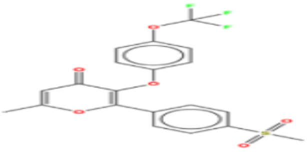
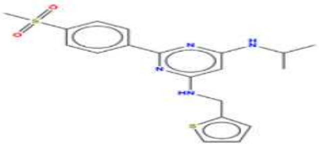
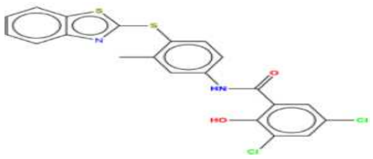
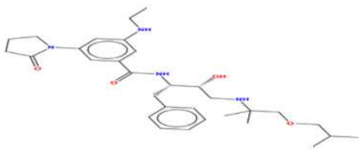
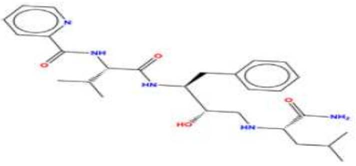
	CHEMBL 193280		0.0223	0.1144	0.0685	0.0363
	CHEMBL 443605		0.0564	0.1847	0.1638	0.2186
VDR	CHEMBL 7162 – cox2&VDR		0.0658	0.0107	0.0795	0.0637
	CHEMBL 1322390		0.0488	0.0420	0.2391	0.0682
	CHEMBL 1452735		0.0983	0.0849	0.3748	0.1003
COX2	CHEMBL 1163237		0.4773	0.7264	0.4693	0.2694
	CHEMBL 127560		0.0811	0.0436	0.0326	0.0490
	CHEMBL 271614		0.0704	0.0417	0.0684	0.0724
CTSD	CHEMBL 263810		0.0889	0.0146	0.2667	0.1014
	CHEMBL 252655		0.6800	1.0065	0.9193	0.1174
	CHEMBL 436438		0.5331	0.8771	0.8109	0.0766

Table 3. The description on $\mathbb{P}(\nu(l_m) = i)$ and F_m according to the number of components of GMM, K and the class $\nu(l_m)$ of queries l_m.[a]

K=1		$\mathbb{P}(\boldsymbol{\nu}(\boldsymbol{l_m}) = \boldsymbol{i})$				$\mathbf{F}_{\boldsymbol{m}}$[b]
		Class of representative distributions (i)				
		ESR	VDR	COX-2	CTSD	
Class $\nu(l_m)$ of	ESR	**0.4623**	0.2172	0.0082	0.3123	0.9272
	VDR	0.1116	**0.5101**	0.0054	0.3729	1.0205
	COX-2	0.0882	0.3216	0.2046	0.3856	0.5071
	CTSD	0.0051	0.0489	0.0057	0.9404	3.9718
K=3		$\mathbb{P}(\boldsymbol{\nu}(\boldsymbol{l_m}) = \boldsymbol{i})$				$\mathbf{F}_{\boldsymbol{m}}$[b]
		Class of representative distributions (i)				
		ESR	VDR	COX-2	CTSD	
Class $\nu(l_m)$ of queries	ESR	0.3289	0.2616	0.0725	0.3370	0.7001
	VDR	0.1653	0.5199	0.0517	0.2631	1.0406
	COX-2	0.1024	0.4922	0.1534	0.2520	0.4257
	CTSD	0.1348	0.0741	0.0128	0.7783	1.8738
K=7		$\mathbb{P}(\boldsymbol{\nu}(\boldsymbol{l_m}) = \boldsymbol{i})$				$\mathbf{F}_{\boldsymbol{m}}$[b]
		Class of representative distributions (i)				
		ESR	VDR	COX-2	CTSD	
Class	ESR	0.3669	0.2553	0.0713	0.3065	0.7613
	VDR	0.2164	0.5005	0.0476	0.2356	1.0009

	COX-2	0.1387	0.4891	0.1477	0.2245	0.4164
	CTSD	0.1437	0.0705	0.0084	0.7775	1.8691

[a]This table represents the feasibility of discrimination depending on the number of components in GMM, K and the class ν(l_m) of queries l_m. [b]The larger F_m, the better performance of discrimination between one class and others. Estrogen receptor alpha = ESR, Vitamin D receptor = VDR, Cyclooxygenase-2 = COX-2, Cathepsin D =CTSD.

Table 4. The comparison between representative queries and unprecedented drug, BNDS-A as a query

CLASS	**QUERY**	**SELECTION TYPE**	**MAX. OF K-L DIVERGENCE**			
			ESR	VDR	COX2	CTSD
ESR	CHEMBL 499809	Mean of *Q*	0.0363	0.1991	0.1611	0.2772
	CHEMBL 2	(Mean +2SD) of *Q*	0.1180	0.1001	0.1547	0.0883
	CHEMBL 558943	(Mean -2SD) of *Q*	2.7919	5.2859	2.9632	2.0501
	CHEMBL 604989	Biggest gap of K-L Divergence	6.2458	10.9899	6.1578	5.4983
	CHEMBL 292033	Highest Similarity with SNDS-A	0.0298	0.2570	0.2096	0.1082
	BNDS-A	Unknown	0.2116	**0.0588**	0.1139	0.9704
VDR	CHEMBL 7463	Mean of *Q*	0.0237	0.0442	0.1446	0.1262
	CHEMBL 603	(Mean +2SD) of *Q*	0.0999	0.2738	0.1257	0.0655
	CHEMBL 1116	(Mean -2SD) of *Q*	1.2883	2.1898	1.6169	0.4702
	CHEMBL 486541	Biggest gap of K-L Divergence	4.2675	7.2936	3.9890	3.3430
	CHEMBL 62136	Highest Similarity with SNDS-A	0.2090	0.1854	0.4785	0.1086

	BNDS-A	Unknown	0.2859	**0.0864**	0.1888	1.0807
COX2	CHEMBL 1201356	Mean of *Q*	0.0963	0.1054	0.2187	0.0948
	CHEMBL 16516	(Mean +2SD) of *Q*	0.1445	0.1172	0.0385	0.1205
	CHEMBL 1171450	(Mean -2SD) of *Q*	3.2143	5.5460	3.1399	2.4262
	CHEMBL 1171454	Biggest gap of K-L Divergence	4.4382	7.8994	4.1848	4.1940
	CHEMBL 942	Highest Similarity with SNDS-A	0.1285	0.0546	0.09018	0.06225
	BNDS-A	Unknown	0.6987	0.65378	**0.2273**	2.0276
CTSD	CHEMBL 263810	Mean of *Q*	0.0850	0.0113	0.2512	0.1038
	CHEMBL 504438	(Mean +2SD) of *Q*	0.6941	1.1751	1.1002	0.3305
	CHEMBL 567893	(Mean -2SD) of *Q*	3.5366	6.1606	3.5399	2.0713
	CHEMBL 567893	Biggest gap of K-L Divergence	3.5684	6.1606	3.5399	2.0713
	CHEMBL 387576	Highest Similarity with SNDS-A	0.0835	0.1467	0.0952	0.0129
	BNDS-A	Unknown	**0.0556**	0.26421	0.2092	0.087

Supplementary Files

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

- Terminologyannotation.docx