International Journal of
Molecular Sciences

Article

Comparing a Query Compound with Drug Target Classes Using 3D-Chemical Similarity

Sang-Hyeok Lee [1,2], Sangjin Ahn [3] and Mi-hyun Kim [1,*]

[1] Gachon Institute of Pharmaceutical Science and Department of Pharmacy, College of Pharmacy, Gachon University, Yeonsu-gu, Incheon 21936, Korea; prizeh83@gmail.com

[2] Innovation Center for Industrial Mathematics, National Institute for Mathematical Science, Yeongtong-gu, Suwon 16229, Korea

[3] Department of Financial Engineering, College of Business, Ajou University, Suwon 16499, Korea; asj92@ajou.ac.kr

[*] Correspondence: kmh0515@gachon.ac.kr

Received: 18 May 2020; Accepted: 11 June 2020; Published: 12 June 2020

Abstract: 3D similarity is useful in predicting the profiles of unprecedented molecular frameworks that are 2D dissimilar to known compounds. When comparing pairs of compounds, 3D similarity of the pairs depends on conformational sampling, the alignment method, the chosen descriptors, and the similarity coefficients. In addition to these four factors, 3D chemocentric target prediction of an unknown compound requires compound–target associations, which replace compound-to-compound comparisons with compound-to-target comparisons. In this study, quantitative comparison of query compounds to target classes (one-to-group) was achieved via two types of 3D similarity distributions for the respective target class with parameter optimization for the fitting models: (1) maximum likelihood (ML) estimation of queries, and (2) the Gaussian mixture model (GMM) of target classes. While Jaccard–Tanimoto similarity of query-to-ligand pairs with 3D structures (sampled multi-conformers) can be transformed into query distribution using ML estimation, the ligand pair similarity within each target class can be transformed into a representative distribution of a target class through GMM, which is hyperparameterized via the expectation–maximization (EM) algorithm. To quantify the discriminativeness of a query ligand against target classes, the Kullback–Leibler (K–L) divergence of each query was calculated and compared between targets. 3D similarity-based K–L divergence together with the probability and the feasibility index, (F_m), showed discriminative power with regard to some query–class associations. The K–L divergence of 3D similarity distributions can be an additional method for (1) the rank of the 3D similarity score or (2) the *p*-value of one 3D similarity distribution to predict the target of unprecedented drug scaffolds.

Keywords: Kullback–Leibler (K–L) divergence; chemocentric similarity; Jaccard–Tanimoto coefficient; Gaussian mixture model (GMM); expectation-maximization (EM) algorithm; maximum likelihood (ML) estimation; machine learning

1. Introduction

An unpresented molecular framework such as that in Figure 1a can be investigated in drug space. In early stages of drug discovery, three-dimensional (3D) similarity between chemicals has been used to find desirable ligands of a chosen therapeutic target in virtual screening (VS; Figure 1b) [1,2]. To our knowledge, chemical similarity is a coarse predictor for filtering out less promising chemicals rather than selecting the most desirable compound. Chemical similarity has also contributed to target screening (in other words, retro-VS) under the chemocentric assumption in Figure 1c. Chemocentric assumption means if two similar molecules are likely to possess similar properties, they can share biological targets or may show similar pharmacological profiles [3,4]. Remarkably, Jain's group conducted on-target and off-target prediction through the comparison of two-dimensional (2D) and 3D chemical similarity [5]. Based on this comparison, while dual 2D and 3D similarity-based predictions showed superiority for either 2D or 3D predictions, 3D predictions did not show dramatic improvement over 2D predictions. In addition, the increase of data points, according to the conformer sampling sizes, makes the computing cost of 3D features increase more rapidly than 2D features. However, despite it being less cost-effective, 3D similarity is the best feature for in silico target screening of unprecedented drug scaffolds and new drug-like molecular frameworks [6] because (1) novel, unprecedented drug scaffolds have very low 2D similarity to known bioactive molecules [7–9], (2) novel pharmacological profiles of drugs are more frequently found using 3D similar off-target predictions [5], and (3) realistic drug properties can be generated from their factual and flexible 3D structures [10–12].

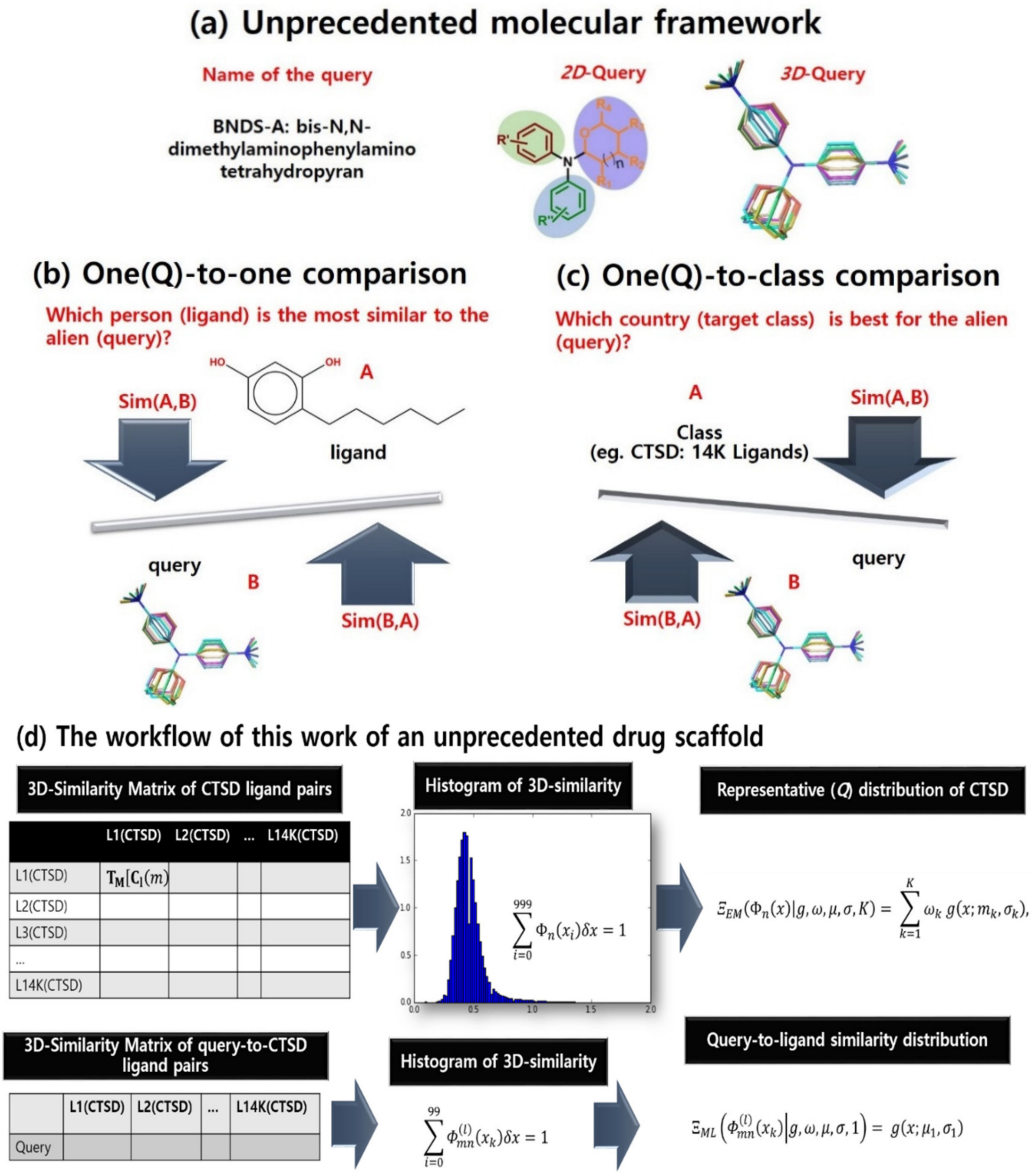

(a) Unprecedented molecular framework
Name of the query
BNDS-A: bis-N,N-dimethylaminophenylamino tetrahydropyran
2D-Query
3D-Query
R'
R''
N
O
n
R₁
R₂
R₃
R₄
(b) One(Q)-to-one comparison
Which person (ligand) is the most similar to the alien (query)?
HO
OH
A
Sim(A,B)
ligand
query
B
Sim(B,A)
(c) One(Q)-to-class comparison
Which country (target class) is best for the alien (query)?
A
Sim(A,B)
Class
(eg. CTSD: 14K Ligands)
query
B
Sim(B,A)
(d) The workflow of this work of an unprecedented drug scaffold
3D-Similarity Matrix of CTSD ligand pairs
L1(CTSD)
L2(CTSD)
...
L14K(CTSD)
L1(CTSD)
$T_M[C_1(m)$
L2(CTSD)
L3(CTSD)
...
L14K(CTSD)
Histogram of 3D-similarity
2.0
1.5
1.0
0.5
0.0
0.0
0.5
1.0
1.5
2.0
$\sum_{i=0}^{999} \Phi_n(x_i)\delta x = 1$
Representative (Q) distribution of CTSD
$\Xi_{EM}(\Phi_n(x)|g,\omega,\mu,\sigma,K) = \sum_{k=1}^{K} \omega_k\, g(x; m_k, \sigma_k),$
3D-Similarity Matrix of query-to-CTSD ligand pairs
L1(CTSD)
L2(CTSD)
...
L14K(CTSD)
Query
Histogram of 3D-similarity
$\sum_{i=0}^{99} \Phi_{mn}^{(l)}(x_k)\delta x = 1$
Query-to-ligand similarity distribution
$\Xi_{ML}\left(\Phi_{mn}^{(l)}(x_k)\middle|g,\omega,\mu,\sigma,1\right) = g(x;\mu_1,\sigma_1)$

Figure 1. The problem definition of 3D chemo-centric screening. (**a**) BNDS-A as a new molecular framework. (**b**) The role of chemical similarity in virtual screening. (**c**) The role of chemical similarity in chemo-centric retro-virtual screening. (**d**) The workflow of this work of an unprecedented drug scaffold.

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", inspired this research for the 'chemistry-oriented synthesis' of an unprecedented drug scaffold [7–9] and the chemocentric target profiling of this scaffold [7]. For this purpose, we have intensively studied the 3D similarity of unprecedented drug scaffolds (the query compounds) with known molecular frameworks (the reference compounds). When comparing query and reference compound pairs, 3D similarity of the pairs depends on (1) conformational sampling of the compounds, (2) the alignment method, (3) the chosen descriptors, and (4) the distance coefficients (e.g., Jaccard–Tanimoto). In addition to the four factors of 3D VS, retro-VS of unprecedented drug scaffolds (query compounds) requires compound–target associations (target class information), as shown in Figure 1. These associations are the source of the substantial difference between VS and retro-VS in problem-solving in data science, specifically, (1) one-to-one comparison for VS, as shown in Figure 1b; (2) one-to-group (class) comparison for retro-VS, as shown in Figure 1c; and (3) group-to-group comparison for typical parametric statistics such as ANOVA and *t*-test. When we calculated the similarity of compound pairs in retro-VS, the hope was to ultimately identify the primary target of the query through calculated chemical similarity rather than finding the most similar compound to the query structure. To achieve this, one-to-group comparison must be essentially quantified. To our knowledge, such measurements have not been properly reported in cheminformatics. Notably, 2D similarity distributions with target annotation have been reported using statistical fitting models such as Shoichet's group [3], Bajorath's group [13], and Nasr's group [14]. However, even though the number of studies using 3D similarity is enormous with review articles by Zhang et al. [15] and Shin et al. [16], 3D similarity distribution is rarely mentioned in the literature. Other than the distribution, network analysis (edge: similarity, node: chemical) such as that by Torres et al. [17] or the machine-learning algorithm-based classifiers have also been used [11,18]. Most classifiers do not only use chemical similarity, but also use other descriptors together [18]. Although several studies have treated 3D similarity distribution such as Jain's group [5], Medina-Franco's group [19], and Pérez-Nueno's group [20], the distribution comprised every compound instead of compounds grouped by target [5,19]. In addition, it was either visualized without a fitting model [19] or its statistical model was chosen without parameter optimization [5]. Exceptionally, although Pérez-Nueno's group reported Gaussian distribution using 3D similarity, the study assumed Gaussian distribution with only one centroid and fitting parameter was also not optimized, despite the small number of ligands [20].

In this study, we quantitatively compared a query compound with a target class (one-to-group) using two types of similarity distributions, namely, maximum likelihood (ML) estimation of queries and a Gaussian mixture model (GMM) of target classes (Figure 1d). As raw data of this study, the Jaccard–Tanimoto similarity coefficients were calculated for (1) query-to-ligand pairs (e.g., the left second row of the Figure 1d) and (2) ligand pairs within each target class (e.g., the left first row of Figure 1d). The query-to-ligand similarity was transformed into query distribution via ML estimation, and the ligand pair similarity was also transformed into a representative distribution of a target class using GMM. The difference between two distributions was quantified by Kullback–Leibler (K–L) divergence, which represented the quantitative comparison between a query and a target class. In order to evaluate whether the K–L divergence accurately achieved one-to-group comparison, a query chosen from a group of known ligands for a target was tested to observe discrimination between the original target and other targets. In sequence, the target profiles of an unprecedented drug scaffold was explained by K–L divergence.

2. Theoretical Background

Kullback–Leibler divergence: K–L divergence measures the difference between two statistical or probabilistic distributions. In particular, K–L divergence is employed in various machine learning and deep learning algorithms for statistical inference [21,22]. Since K–L divergence implies relative entropy, which is an important concept in understanding statistical phenomena, it applies 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 σ–algebra, and P and Q are probability distributions. Then, to define Kullback–Leibler divergence, a unique measurable function is devised, $\frac{dQ}{dP} : \Omega \to \mathbb{R}^{+}$, known as the Radon–Nykodym derivative, so that

$$Q(\mathrm{E}) \; = \; \int\limits_{E} \frac{dQ}{dP} dP \tag{1}$$

For any measurable set, $\mathrm{E} \in \Omega$ [22] when using the measurable function $\frac{dQ}{dP}$. The Kullback–Leibler divergence, $\mathrm{D}(\mathrm{P} \parallel \mathrm{Q})$, is defined as either

$$D(P \parallel Q) := \int_{\Omega} -\ln\left(\frac{dP}{dQ}\right) dP \tag{2}$$

or

$$D(P \parallel Q) := \int_{-\infty}^{\infty} \ln\left(\frac{p(x)}{q(x)}\right) p(x) dx, \tag{3}$$

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

$$P(x) := \int_{-\infty}^{x} p(x) dx \text{ and } Q(x) := \int_{-\infty}^{x} q(x) dx \tag{4}$$

The Kullback–Leibler divergence represents the information for comparing $P(x)$ and $Q(x)$ distributions [23]. Hence, the implication of Kullback–Leibler divergence depends on the definitions of $P(x)$ and $Q(x)$. For example,

Model Inference: If $P(x)$ represents the testing distribution based on the model, and $Q(x)$ represents the distribution from the raw data, the difference is the error between the model and reality [24];

Informatics: If $P(x)$ and $Q(x)$ represent information extracted from two objectives, the divergence is a measurement for the discrimination between two objectives [13,25];

Bayesian Statistics: If $P(x)$ represents a prior distribution and $Q(x)$ represents a posterior distribution, the divergence represents the information gained through updating [26,27].

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 $\mathrm{G}\left(x; m_j, \sigma_j\right)$, where

$$\mathrm{G}(x; m_i, \sigma_i) := \int_{-\infty}^{x} g(s; m_i, \sigma_i) ds \text{ and } \mathrm{G}\left(x; m_j, \sigma_j\right) := \int_{-\infty}^{x} g\left(s; m_j, \sigma_j\right) ds \tag{5}$$

Using Equations (3) and (5), the Kullback–Leibler divergence between the two Gaussian distributions $\mathrm{G}(x; m_i, \sigma_i)$ and $\mathrm{G}\left(x; m_j, \sigma_j\right)$ in Equation (5) are as follows:

$$D\left(G(x; m_i, \sigma_i) \parallel G\left(x; m_j, \sigma_j\right)\right) \; = \; \ln\left(\frac{\sigma_j}{\sigma_i}\right) + \frac{(\sigma_i)^2 + \left(m_i - m_j\right)^2}{2\left(\sigma_j\right)^2} - \frac{1}{2} \tag{6}$$

This Kullback–Leibler divergence between the univariate normal distributions (Equation (6)) therefore extends to multivariate distributions [28].

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

$$\Xi(\Phi(\mathbf{x})\big|p, \omega, \lambda, K) := \sum_{k=1}^{K} \omega_k p(\mathbf{x} : \lambda_k) \tag{7}$$

where for k = 1, 2, ... , K, ω_k, λ_k are the weights and vectors of the hyperparameters and p_i is the i_{th} component, which is independently and identically distributed (iid) [29]. In this work, GMM was adopted to obtain a representative distribution [30]. Notably, GMM is a model that describes non-Gaussian distributions as well as Gaussian distributions [31]. The probability density $p(x : \lambda_k)$ represents the Gaussian density function $g(x; m_k, \sigma_k)$ in Equation (5). In the Gaussian mixture model, estimations of the weight (ω_k), the mean (m_k), and the standard deviation (σ_k) are essential. Herein, the two methods (i.e., the EM algorithm [32] and ML estimation [33]) were chosen to estimate the hyperparameters from sparse and incomplete data. The EM algorithm for GMM consists of an initial guess for the GMM parameters and iterative calculation (E-step)–parameter determination (M-step). The iterative steps continue until the set of hyperparameters, θ, are less than positive, and infinitesimal number, ϵ, as shown in the cccccmathematical elucidation (Supplementary Materials Equations (S1.6)–(S1.12) [34]. For convenience, when applying the ML estimation, $\Phi(x)$ is transformed into the mixture model and $\Xi(\Phi(x)\big|p, \omega, \lambda, K)$ is replaced by $\Xi_{EM}(\Phi(x)\big|p, \omega, \lambda, K)$.

3. Results and Discussion

In this study, a quantitative method was developed to describe discriminative information for target prediction of a query compound only from chemical similarity and known compound–target association information. For this purpose, 3D similarity distributions were acquired from a 3D similarity matrix occupied by Jaccard–Tanimoto coefficients [35] regarding (1) query-to-ligand pairs and (2) ligand pairs within each target class. The Jaccard–Tanimoto coefficients were calculated from two types of features, molecular shape and pharmacophore features, using the Openeye Toolkit. Query compounds and target classes were compared and quantified according to the following process:

Step 1. EM algorithm-based GMM allowed to obtain a representative distribution (Q-distribution) for a target class, following either Gaussian or non-Gaussian distribution;

Step 2. A query-to-ligand similarity distribution was fitted onto a Gaussian distribution using ML estimation;

Step 3. K–L divergence between the two distributions from Step 1 and Step 2 allowed target predictions of the query compound. Greater deviation of K–L divergence values between target classes indicated that the query compound was a more representative ligand of a class than other query compounds. In addition, the probability, $\mathbb{P}(\nu(l_m) = i)$, derived from the K–L divergence values and the feasibility index, F_m, allowed for quantification of discrimination between the target classes.

Dataset: In order to select example target classes for this study, an unprecedented scaffold with structural novelty and its targets were focused. Among our previous studies, bis-*N*,*N*-dimethylaminophenylamino tetrahydropyran (BNDS-A), which was the most potent to regulate in vitro inflammation (IC_{50} of nitric oxide production = 12 μM), was chosen for this quantitative method (Figure 1a). The association of two targets with BNDS-A, estrogen receptor alpha (ESR), and vitamin D receptor (VDR) was proven by the stepwise approach consisting of (1) 2D similarity search, (2) multiplication of 3D similarity coefficients of every ligand within each target, P(Tc)/C(hits), (3) self/cross-similarity, and (4) western blot analysis in our previous work [7]. However, despite low predicted probability, capthesin D (CTSD) and cyclooxygenase-2 (COX2) could also be regulated by BNDS-A in the same study. Neither the most similar compound to BNDS-A (one-to-one comparison) nor ANOVA test between target pairs (group-to-group comparison) could suggest the primary target of BNDS-A. Therefore, to quantitatively compare them with BNDS-A, the four targets, ESR, VDR, COX2, and CTSD, were selected. In addition, an additional four targets, HIV-1 protease (HIV1), heat shock protein 90 (HSP90), transient receptor potential cation channel subfamily V4 (TRPV4), DNA topoisomerase I (TOP1), were randomly selected from the target prediction literature [36] to evaluate our methodology. For convenience, simple numbers denoted the target classes, in other words,

$$\begin{cases} Estrogen\ receptor\ alpha \rightarrow 1, \\ Vitamin\ D\ receptor \rightarrow 2, \\ Cyclooxygenase-2 \rightarrow 3, \\ Cathepsin\ D \rightarrow 4. \end{cases} \tag{8}$$

Either m or n were called the class number, which was an integer between 1 and 4, as in Equation (8), and $\mathrm{C}_L(m)$ and $\mathrm{C}_L(n) \in \mathbb{R}^N$ represent vectors whose elements are the Tanimoto coefficients of query compounds in the mth class. $\mathrm{T}_M : \mathbb{R}^{2N} \rightarrow \mathbb{R}^N \times \mathbb{R}^N$ was defined as the Tanimoto matrix operator, so

$$(\mathbf{T_M}[\mathbf{C_l}(m), \mathbf{C_l}(n)])_{ij} := \mathrm{T_c}\Big(< \mathbf{e_i} \cdot \mathbf{C_l}(m) >, < \mathbf{e_j}, \mathbf{C_l}(n) >\Big) \tag{9}$$

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

Representative distributions *Q* for target classes: The representative distributions corresponding to each target class using GMM of ligand pair similarity were obtained. First, using the similarity matrix $\mathbf{T_M}[\mathbf{C_l}(m), \mathbf{C_l}(n)]_{ij}$ in Equation (9), where m = n, the following univariate probability densities, $\Phi_n(x_k)$, were defined by

$$\Phi_n(x_i)\delta x := \mathbb{P}\Big(x_k \leq X = \mathbf{T_M}[\mathbf{C_l}(m), \mathbf{C_l}(n)]_{ij} \leq x_{k+1}\Big), \tag{10}$$

where $\mathbb{P}$ is the probability measure; x is the Tanimoto–Jaccard coefficients; $0 = x_0$ and the range of x is $[0, 2]$; and $x_{k+1} = x_k + \delta x$. Therefore, the probability densities, $\Phi_n(x)$, satisfy the following equation:

$$\sum_{i=0}^{999} \Phi_n(x_i)\delta x = 1 \tag{11}$$

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

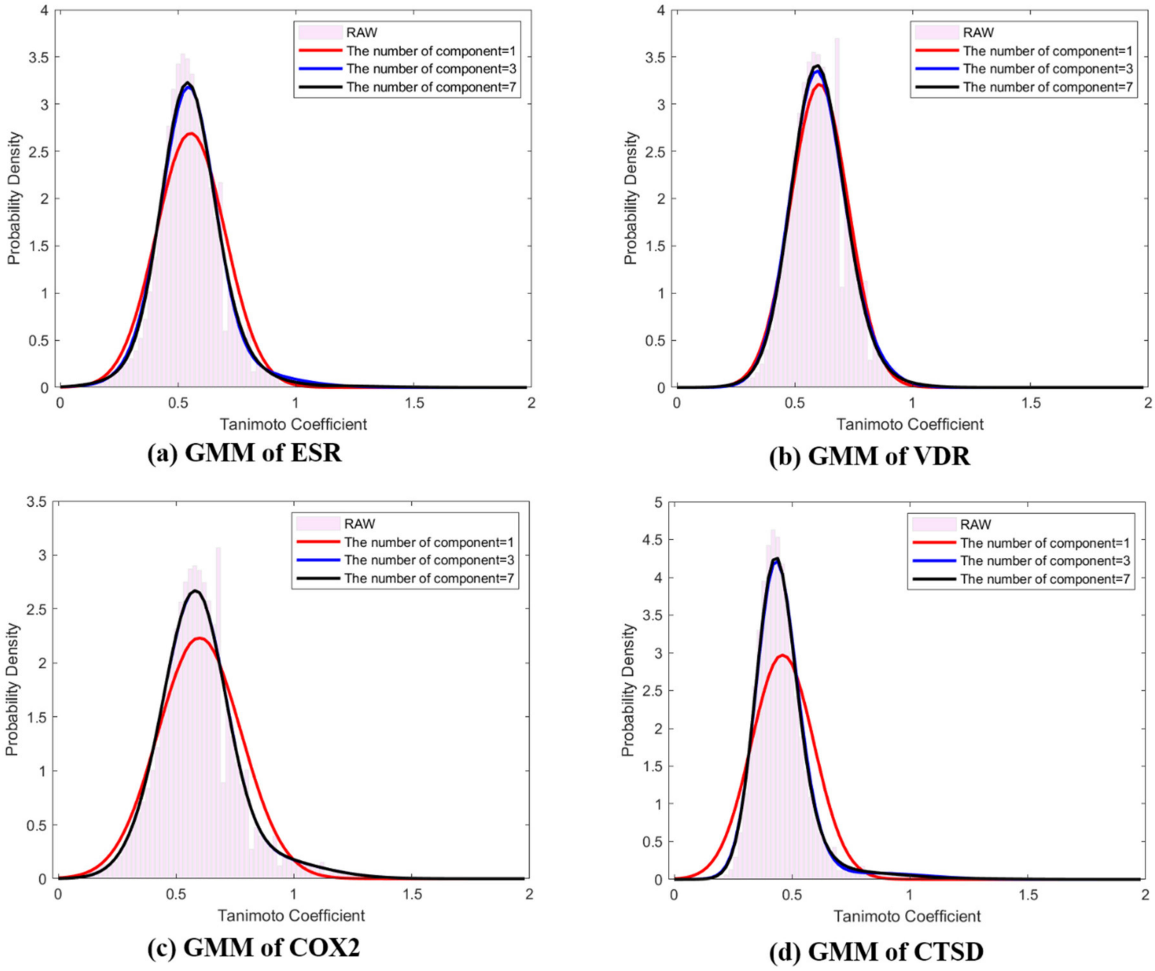
$$\Xi_{EM}(\Phi_n(x) \big| g, \omega, \mu, \sigma, K) = \sum_{k=1}^{K} \omega_k g(x; m_k, \sigma_k), \tag{12}$$

where ω_i, m_i, and σ_i are shown in Table 1. To estimate the hyperparameters ω_i, m_i, and σ_i, the EM algorithm was used as described in Section 4. Table 1 shows the mean, standard deviation, and weight corresponding to the components of the mixture model. Figure 2 depicts the difference between the probability densities, $\Phi_n(x)$, and $\Xi_{EM}(\Phi_n(x) \big| g, \omega, \mu, \sigma, K)$, where $K = 1$, 3, and 7. When comparing component K, raw data were similarly fitted to histograms when K = 3 and K = 7, and normal Gaussian modeling showed insufficient fitting for ESR, COX2, and CTSD (Figure 2). Commonly, the means and modes of the representative distributions existed near 0.5, and every distribution was skewed to the right.

Gaussian distributions for queries: To quantitatively compare the representative distributions corresponding to ESR, VDR, COX2, and CTSD with the query distributions, Kullback–Leibler divergence was introduced and calculated by building each distribution for each query.

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
GMM	**HIV1**		**HSP90**		**TRPV1**		**TOP1**	
No(i)	m_i	σ_i	m_i	σ_i	m_i	σ_i	m_i	σ_i
1	0.419	0.123	0.614	0.206	0.667	0.176	0.510	0.222



(a) GMM of ESR

(b) GMM of VDR

(c) GMM of COX2

(d) GMM of CTSD

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. (**a**) *Q*-distribution of ESR; (**b**) *Q*-distribution of VDR; (**c**) *Q*-distribution of COX2; (**d**) *Q*-distribution of CTSD. The red line: GMM *K* = 1, blue line: GMM *K* = 3, black line: GMM *K* = 7, pink bar: histogram of raw data.

For this purpose, $\mathbf{T_M}[\mathbf{C_l}(m), \mathbf{C_l}(n)]$ of Equation (9) was used in a similar way to the described method for the representative distributions of the target classes. When a query was the *l*th ligand of $\mathbf{C_l}(n)$, the *l*th column's elements in the above matrix were used for the *l*th column vector, $\boldsymbol{\tau}_m(m,n,l)$, as in

$$\boldsymbol{\tau}_m(m,n,l) := \mathbf{T_M}[\mathbf{C_l}(m), \mathbf{C_l}(n)]\mathbf{E}_l \tag{13}$$

where the values of E_l for j = 1, 2, … , N were represented by the N × N matrices, for which the elements $(\mathbf{E}_l)_{ij}$ satisfied

$$(\mathbf{E}_l)_{ij} := \begin{cases} 1, & if \quad i = j \\ 0, & otherwise \end{cases} \tag{14}$$

Using the vector $\boldsymbol{\tau}_m(m,n,l)$ from Equation (13), the following univariate probability densities, $\Phi_{mn}^{(l)}(x_k)$, were defined as

$$\Phi_{mn}^{(l)}(x_k)\delta x := \mathbb{P}(x_k \le X = (\boldsymbol{\tau}_m(m,n,l))_i \le x_{k+1}) \tag{15}$$

where the probability measure $\mathbb{P}$ was derived from Equation (10).

Before obtaining the probability distribution, two assumptions were made. First, it was assumed that a distribution from one query was not a weighted sum of Gaussian distributions, but rather a simple Gaussian distribution. It was reasonable that a distribution from one query was simpler than the *Q*-distribution of a target class with 13,957 queries. Second, to estimate the parameters of the Gaussian distribution, ML estimation was chosen as a general method, in which

$$\Xi_{ML}\left(\Phi_{mn}^{(l)}(x_k)\middle|g,\omega,\mu,\sigma,1\right) = g(x;\mu_1,\sigma_1) \tag{16}$$

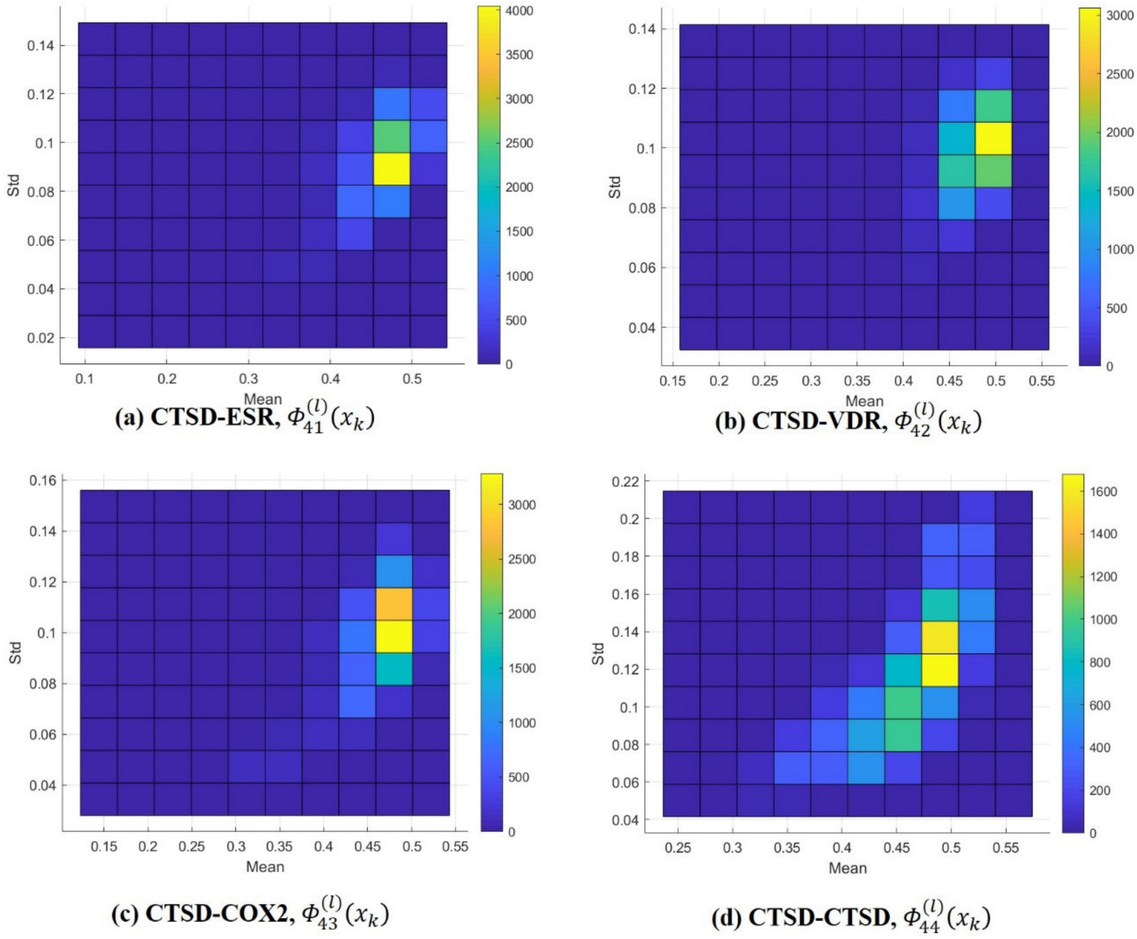
where μ_1 and σ_1 are hyperparameters and are maximized log likelihood functions for normal distribution, in other words,

$$(\mu_1,\sigma_1) := \underset{(\mu,\sigma)}{\arg\max}\sum_{k=1}^{100}\frac{(x_k-\mu)^2}{\sigma^2} \tag{17}$$

Using definitions Equations (16) and (17), each query resulted in four distributions corresponding to the four classes (i.e., ESR, VDR, COX2, and CTSD). For example, when CHEMBL539392 was chosen as a query (*l*) among the ligands of ESR (Class 1), the distributions $\Phi_{11}^{(l)}(x_k)$, $\Phi_{12}^{(l)}(x_k)$, $\Phi_{13}^{(l)}(x_k)$, and $\Phi_{14}^{(l)}(x_k)$ were obtained under the definitions of Equations (8) and (15). According to Equations (16) and (17), four representative Gaussian distributions of the query compound CHEMBL539392 were acquired from the column vector between CHEMBL539392 and 13,957 ligands of each class, which were

$$\left\{\begin{array}{l}\Xi_{ML}\left(\Phi_{11}^{(l)}(x_k)\middle|g,\omega,\mu,\sigma,1\right) = g(x;0.24055,0.07472),\\ \Xi_{ML}\left(\Phi_{12}^{(l)}(x_k)\middle|g,\omega,\mu,\sigma,1\right) = g(x;0.21976,0.06466),\\ \Xi_{ML}\left(\Phi_{13}^{(l)}(x_k)\middle|g,\omega,\mu,\sigma,1\right) = g(x;0.24389,0.04857),\\ \Xi_{ML}\left(\Phi_{14}^{(l)}(x_k)\middle|g,\omega,\mu,\sigma,1\right) = g(x;0.21187,0.06631),\end{array}\right. \quad \text{for } k = 0,1,\ldots,99. \tag{18}$$

In the same way, univariate normal distributions were obtained of all of the query compounds in each class. Since the number of classes was four and there were 13,957 query compounds in each class, the Gaussian distributions G(x; μ_1,σ_1), derived from $\Xi_{ML}\left(\Phi_{mn}^{(l)}(x_k)\middle|g,\omega,\mu,\sigma,1\right)$, presented the class number, either *m* or *n*, which was an integer between 1 and 4, and the query number, *l*, which was an integer from 1 to 13,957. As a result, the frequency distributions of the estimates, alongside the means (μ_1) and standard deviations (σ_1), were described as shown in Figure 3 and Supplementary Figures S5–S7. ML estimation did not show any difference between self-query (m = n) and cross-query (m ≠ n) with regard to frequency. Even though cathepsin D (CTSD) showed a slightly lower mean than the other classes, self-comparison also showed a low mean, as shown in Figure 3. Regardless of whether a class or a query compound was used (self/cross), 3D similarity of ligand pairs within a class showed the mode near 0.6, thereby confirming the need for quantitative comparison between queries. Notably, the univariate probability distributions of 3D similarity did not discriminate between target class at all.



(a) CTSD-ESR, $\Phi_{41}^{(l)}(x_k)$

(b) CTSD-VDR, $\Phi_{42}^{(l)}(x_k)$

(c) CTSD-COX2, $\Phi_{43}^{(l)}(x_k)$

(d) CTSD-CTSD, $\Phi_{44}^{(l)}(x_k)$

Figure 3. Frequency distributions of $\Xi_{ML}\left(\Phi_{4n}^{(l)}(x_k)\middle|g,\omega,\mu,\sigma,1\right)$ estimates (μ_1 and σ_1). Query $(l) \in$ CTSD (class = 4). (**a**) CTSD-ESR, (**b**) CTSD-VDR, (**c**) CTSD-COX2, and (**d**) CTSD-CTSD. * The color bars (right side of the distribution) indicate frequency (e.g., yellow in 3(a) represents 3500 to 4000 queries, the mean of the ML estimates varied from 0.45 to 0.5 and their standard deviation varied from 0.08 to 0.1 in the standard).

Discrimination and K–L divergence: In sequence, 3D similarity distributions of target classes and query compounds were quantitatively compared through K–L divergence calculations. First, the information describing specific Tanimoto–Jaccard coefficients, x, were defined as

$$\ln\left(\frac{\Xi_{ML}\left(\Phi_{mn}^{(l)}(x)\middle|g,\omega,\mu,\sigma,1\right)}{\Xi_{EM}\left(\Phi_{n}(x)\middle|g,\omega,\mu,\sigma,K\right)}\right) \tag{19}$$

from two probability density distributions, $\Xi_{ML}\left(\Phi_{mn}^{(l)}(x)\middle|g,\omega,\mu,\sigma,1\right)$ and $\Xi_{EM}\left(\Phi_{n}(x)\middle|g,\omega,\mu,\sigma,K\right)$, which were generated from a query compound and a class. Hence, following the expected value from the above information in Equation (19) with respect to one query compound, the K–L divergence,

$$\begin{aligned} &D\left(\Xi_{ML}\left(\phi_{mn}^{(l)}(x)\middle|g,\omega,\mu,\sigma,1\right) \parallel \Xi_{EM}\left(\phi_{n}(x)\middle|g,\omega,\mu,\sigma,K\right)\right) \\ &= \int \Xi_{ML}\left(\phi_{mn}^{(l)}(x)\middle|g,\omega,\mu,\sigma,1\right)\ln\left(\frac{\Xi_{ML}\left(\phi_{mn}^{(l)}(x)\middle|g,\omega,\mu,\sigma,1\right)}{\Xi_{EM}\left(\phi_{n}(x)\middle|g,\omega,\mu,\sigma,K\right)}\right)dx \end{aligned} \tag{20}$$

represented a measurement for the discrimination.

In a one-component GMM (K = 1), the K–L divergence between Gaussian distributions of every query and the Q-distributions (Table 1) are calculated; randomly chosen query compounds are described in Table 2. To show the calculation process in detail, CHEMBL539392 was chosen as an example. Using the above equation for Kullback–Leibler divergence between normal distributions,

$$D\Big(G(x; m_i, \sigma_i) \parallel G\big(x; m_j, \sigma_j\big)\Big) = \ln\left(\frac{\sigma_j}{\sigma_i}\right) + \frac{(\sigma_i)^2 + \big(m_i - m_j\big)^2}{2\big(\sigma_j\big)^2} - \frac{1}{2} \tag{21}$$

where

$$\begin{cases} G(x; m_i, \sigma_i) = \Xi_{ML}(\phi_{1n}^{(1)}(x)\big| g, \omega, \mu, \sigma, 1) \\ G\big(x; m_j, \sigma_j\big) = \Xi_{EM}(\phi_n(x)\big| g, \omega, \mu, \sigma, 1) \end{cases} \tag{22}$$

Table 2. K–L divergence of randomly chosen queries between Q distributions and the distributions of queries.

Class	Query	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	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

We obtained four K–L divergences corresponding to the queries of 2.1493, 4.6939, 2.0810, and 1.6354, respectively (see calculation procedure in the Supplementary Materials Equations (S2.1–S2.8). As shown in Table 2 and Supplementary Table S3, the K–L divergence of every query compound was not always the smallest value from their original targets, as annotated by ChEMBL DB. Even though a considerable number of query compounds showed that the K–L divergence resulting from an original target was smaller than values from other target classes, CHEMBL539392 of ESR, CHEMBL1163237 of COX2, and CHEMBL263810 of CTSD were considered to be less different than other targets, therefore giving a false prediction (Table 2). When we counted the query compounds that discriminated between the original targets and other targets from the 13,957 query compounds under the four classes via GMM (K = 1), the correct prediction numbers were 6300, 5200, 4100, and 6400 among each of the 13,957 queries from ESR, VDR, COX2, and CTSD, respectively. When applying GMM (K = 3) and (K = 7) for the Q-distributions, the true positive ratio decreased (ESR: 5100; VDR: 4500; COX2: 3200; CTSD: 4900 (K = 3); ESR: 4900; VDR: 4500; COX2: 3100; CTSD: 4800 (K = 7)).

In order to further evaluate the discriminative power of K–L divergence between target classes, an additional four classes as well as the four classes for BNDS-A were compared with the shared ligands in Table 3 and Supplementary Table S2. In Table 3, ritonavir (CHEMBL163) is a clinically approved drug on the HIV1 (human immunodeficiency virus type 1) protease as its primary target. Notably, ritonavir showed the distinct K–L divergence value to discriminate HIV1 with other targets. In addition, the result can rationalize why ritonavir cannot show a distinct difference between VDR and COX2. In contrast, myricetin (CHEMBL 164) showed very disappointing result with poor discrimination between K–L divergence values. However, when we checked every target of myrcetin, the natural compound did not show target specificity on any single protein to explain the result. The annotated activities were limited to the known targets (VDR: 31–40 μM, COX2: 100 μM, HSP90 13.5 μM in cell-based assay, TOP1: IC50 = 11.9 μg mL^{-1}) in ChEMBL DB. Furthermore, despite the absent data on HIV1 of myrcetin, the flavonoid compound with multiple hydroxyl groups showed experimental activity on ubiquitin-specific protease having functional similarity (peptidase domain) with HIV1 to explain the K–L divergence value of 0.0393. In sequence, because reserpine (CHEMBL772), a clinically approved natural product, has target specificity on vesicular monoamine transporters with trivial activities on the annotated targets (VDR/COX2/TOP1), every target did not show a difference with untested targets (ESR/CPTD/HIV1). In addition, even though CHEMBL1813048 was the ligand of COX2 and TRPV4, K–L divergence could not support the finding. However, the result can be explained by the experimental data: (1) Ki against TRPV4 was more than 10 μM and (2) indirect regulation of COX2 was recorded through the Prostaglandin H2 receptor in ChEMBL DB. When compared with a 2D fingerprint based Top5 prediction of the additional target classes [36], our method can provide how much each query is quantitatively different with each target class from the raw data without any refinements such as assay, activity index, and duplicated ligands. This point is very important for investigating unprecedented drug scaffolds having weak activity out of the Top5 of a target class.

Table 3. K–L divergence of ligands shared with eight target classes *.

Query	Targets	ESR	VDR	COX2	CTSD	HIV1	HSP90	TRPV4	TOP1
CHEMBL 163 (RITONAVIR)	VDR/COX2/HIV1	1.2649	2.2088	1.6702	0.6982	**0.3587**	1.6040	1.9256	1.2754
CHEMBL 164 (MYRICETIN)	VDR/COX2/HSP90/TOP1	0.0718	**0.0526**	0.1148	0.0475	0.0393	0.1655	0.5684	0.0915
CHEMBL 772 (RESERPINE)	ESR/VDR/COX2/TOP1	**0.3075**	0.4963	0.6972	0.2792	0.1685	0.8460	0.7630	0.5009
CHEMBL 1813048	COX2/TPRV4	0.2385	0.3053	**0.4731**	0.2322	0.1704	0.6374	0.6669	0.5810

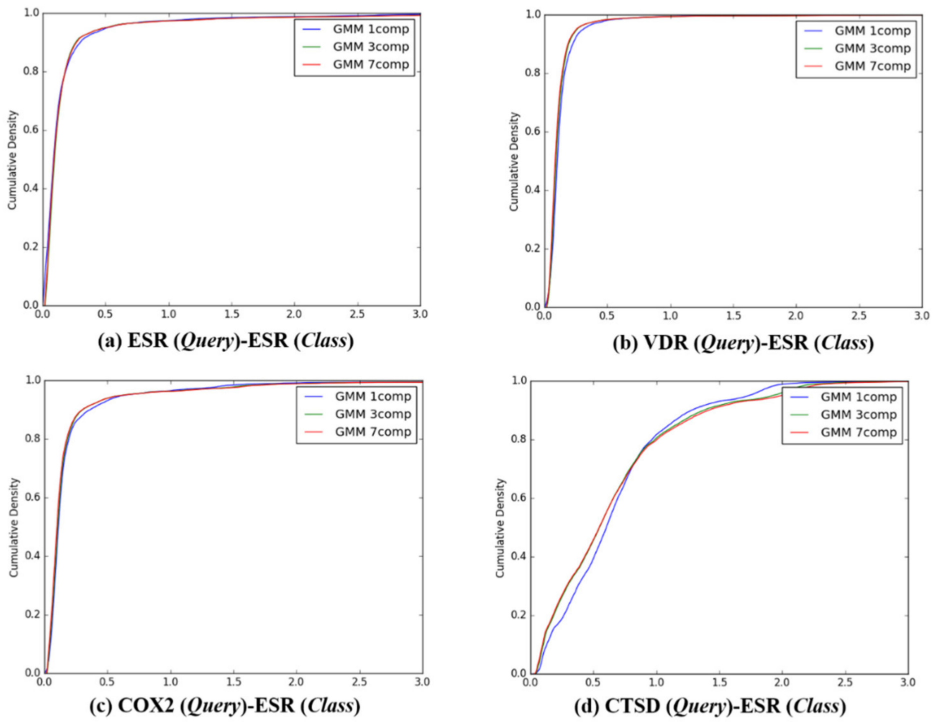
* The smallest K–L divergence value among the experimentally tested targets of each query is presented in bold.

After the individual K–L divergence comparisons of each query, comparisons between the target classes were quantified. In sequence, the K–L divergence between the Gaussian distributions of 13,957 queries and the Q-distributions (K = 1, 3, and 7) for the four target classes were presented as a cumulative distribution, as seen in Figures 4–7. To investigate the feasibility of the information, the following distribution was defined:

$$\mathbb{P}(\nu(l_m) \;=\; i) \text{ for } i \;=\; 1,2,3,4, \tag{23}$$

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

$$\nu(l_m) := \arg\min_n\{D\{\Xi_{ML}(\phi_{mn}^{l_m}(x)|g,\omega,\mu,\sigma,1) \parallel \Xi_{EM}(\phi_n(x)|g,\omega,\mu,\sigma,1)\}|1 \le n \le 4, 1 \le \mathrm{l}_m \le 13,957\} \tag{24}$$



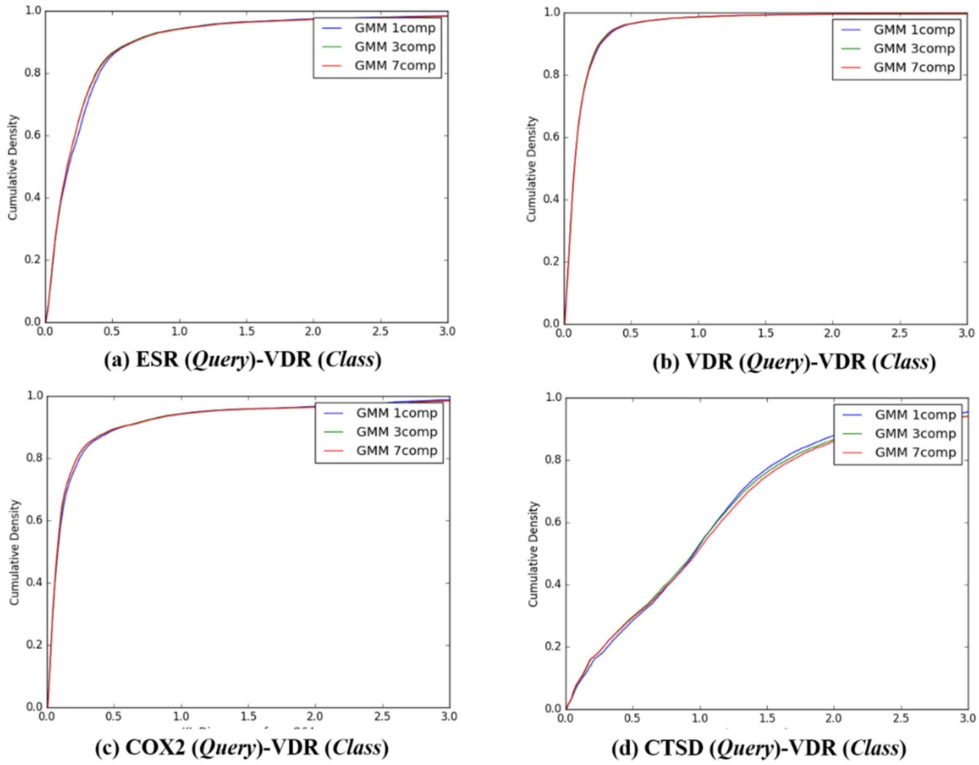
(a) ESR (*Query*)-ESR (*Class*)

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

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

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

Figure 4. The cumulative densities of K–L distance between *Q*-distribution (Target class: ESR) and queries. *X*-axis: K–L divergence, *Y*-axis: cumulative density; *Q*-distribution of ESR through GMM and the distribution of queries were calculated. (**a**) ESR(Query)-ESR(Class), (**b**) VDR(Query)-ESR(Class), (**c**) COX2(Query)-ESR(Class), and (**d**) ESR(Query)-ESR(Class).



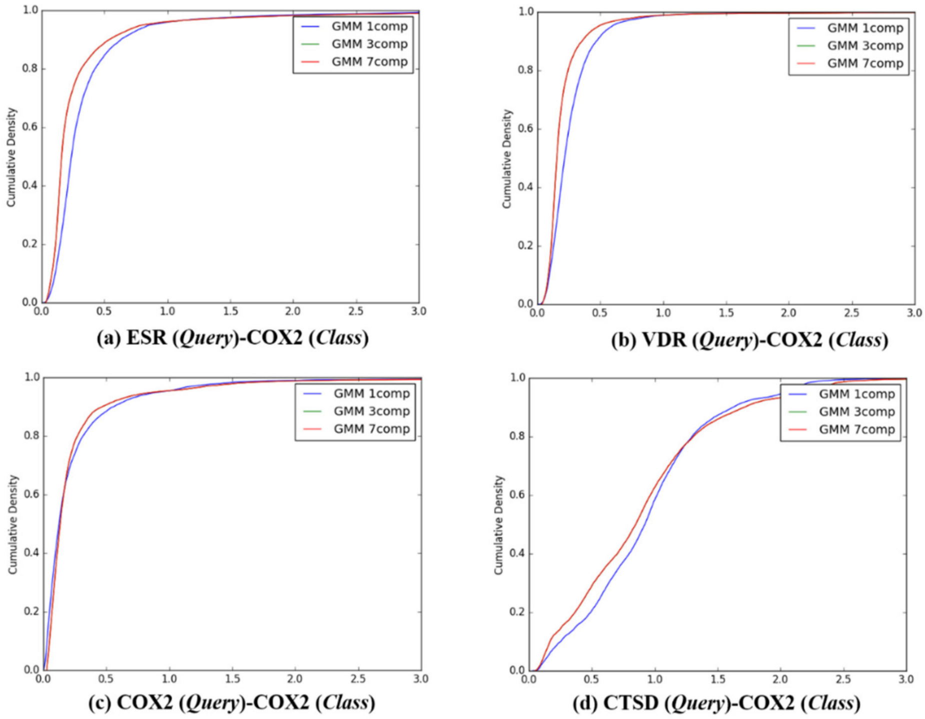
(a) ESR (*Query*)-VDR (*Class*)

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

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

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

Figure 5. The cumulative densities of K–L distance between *Q*-distribution (Target class: VDR) and queries. *X*-axis: K–L divergence, *Y*-axis: cumulative density; *Q*-distribution of VDR through GMM and the distribution of queries were calculated. (**a**) ESR(Query)-VDR(Class), (**b**) VDR(Query)-VDR(Class), (**c**) COX2(Query)-VDR(Class), and (**d**) ESR(Query)-VDR(Class).



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

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

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

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

Figure 6. The cumulative densities of K–L distance between *Q*-distribution (Target class: COX2) and queries. *X*-axis: K–L divergence, *Y*-axis: cumulative density; *Q*-distribution of COX2 through GMM and the distribution of queries were calculated. (**a**) ESR(Query)-COX2(Class), (**b**) VDR(Query)-COX2(Class), (**c**) COX2(Query)-COX2(Class), and (**d**) ESR(Query)-COX2(Class).

If the K–L divergence (Equation (20)) is an ideal measurement for discrimination between target classes, $(\nu(l_m) = i)$ would satisfy the following conditions:

- Necessary condition:

$$\mathbb{P}(\nu(l_m) = m) \geq \max_{i \neq m} \mathbb{P}(\nu(l_m) = i) \tag{25}$$

- Sufficient condition: The feasibility index, F_m, is defined as

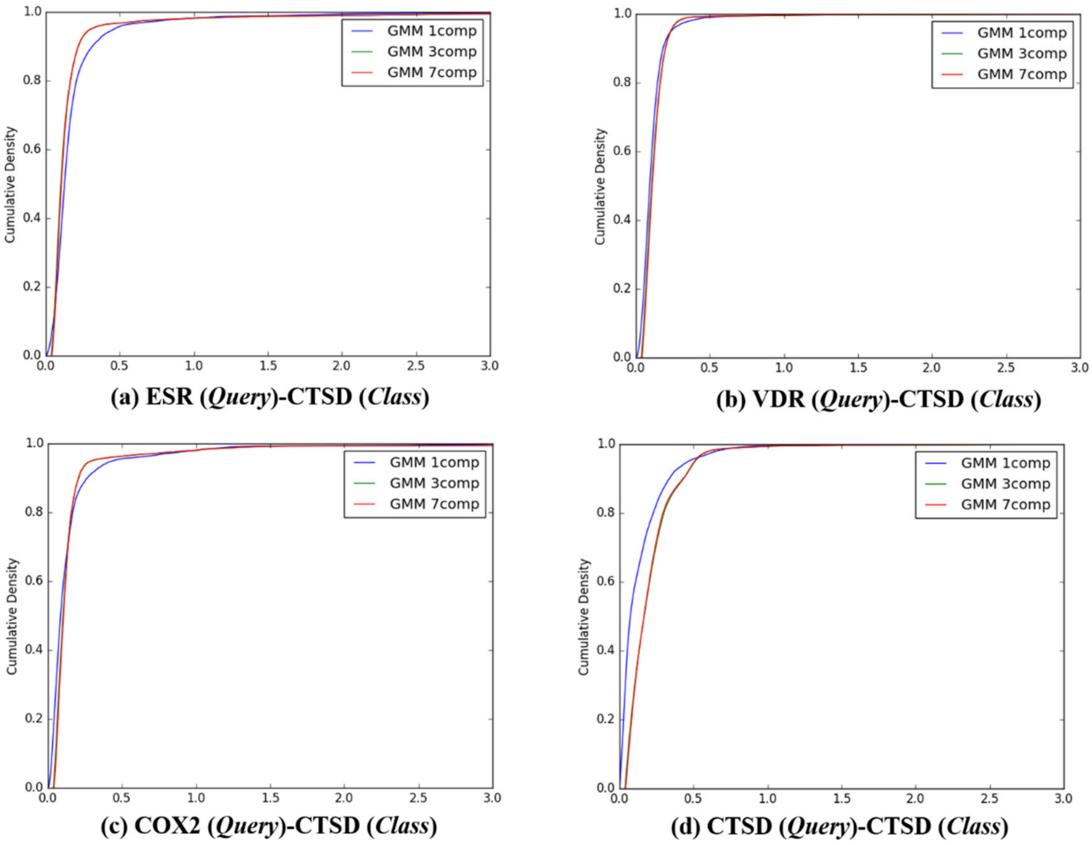
$$\mathrm{F}_m := \sqrt{\frac{\mathbb{P}(\nu(l_m) = m)}{1 - \mathbb{P}(\nu(l_m) = m)}} \geq 1 \tag{26}$$

The above conditions implied a quantitative measurement for the discrimination. In particular, F_m in the sufficient condition represents the ratio between two probabilities (i.e., that a query compound belonged to a class of itself as well as belonging to other classes). A larger value of F_m indicated better feasibility or resolution of discrimination. Table 4 depicts the probability of the K–L divergence $\mathbb{P}(\nu(l_m) = i)$ for $1 \leq i, m \leq 4$, indicating that, except for example $m = 3$ where the class was COX2, the tested classes met the necessary conditions $\mathbb{P}(\nu(l_m) = m) \geq \max_{i \neq m} \mathbb{P}(\nu(l_m) = i)$ in Equation (25) with respect to the feasibility index in Equation (26), it was easier to distinguish a query compound in the CTSD class where $m = 4$ from every class except itself (Figure 8). When the feasibility index resulting from the GMM ($K = 1$) was compared with the index calculated from the GMM ($K = 3$) and ($K = 7$) for the *Q*-distributions, GMM ($K = 1$) showed superior feasibility for class discrimination using GMM ($K = 3$) or ($K = 7$), as shown in Table 4.

Table 4. The description on $\mathbb{P}(\nu(l_m) = i)$ and F_m according to the number of components of Gaussian Mixture Model K, and the class $\nu(l_m)$ of queries l_m [a].

K = 1		$\mathbb{P}(\boldsymbol{\nu}(\boldsymbol{l_m})=\boldsymbol{i})$ Class of representative distributions (i)				$\mathbf{F}_m$ [b]
		ESR	VDR	COX2	CTSD	
Class $\nu(l_m)$ of queries l_m	ESR	**0.4623**	0.2172	0.0082	0.3123	0.9272
	VDR	0.1116	**0.5101**	0.0054	0.3729	1.0205
	COX2	0.0882	0.3216	0.2046	0.3856	0.5071
	CTSD	0.0051	0.0489	0.0057	0.9404	3.9718
K = 3		$\mathbb{P}(\boldsymbol{\nu}(\boldsymbol{l_m})=\boldsymbol{i})$ Class of representative distributions (i)				$\mathbf{F}_m$ [b]
		ESR	VDR	COX2	CTSD	
Class $\nu(l_m)$ of queries l_m	ESR	0.3289	0.2616	0.0725	0.3370	0.7001
	VDR	0.1653	0.5199	0.0517	0.2631	1.0406
	COX2	0.1024	0.4922	0.1534	0.2520	0.4257
	CTSD	0.1348	0.0741	0.0128	0.7783	1.8738
K = 7		$\mathbb{P}(\boldsymbol{\nu}(\boldsymbol{l_m})=\boldsymbol{i})$ Class of representative distributions (i)				$\mathbf{F}_m$ [b]
		ESR	VDR	COX2	CTSD	
Class $\nu(l_m)$ of queries l_m	ESR	0.3669	0.2553	0.0713	0.3065	0.7613
	VDR	0.2164	0.5005	0.0476	0.2356	1.0009
	COX2	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 $\nu(\mathrm{l_m})$ of queries $\mathrm{l_m}$. [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 = COX2, Cathepsin D = CTSD.



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

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

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

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

Figure 7. The cumulative densities of K–L distance between *Q*-distribution (Target class: CTSD) and queries. *X*-axis: K–L divergence, *Y*-axis: cumulative density; *Q*-distribution of CTSD through GMM and the distribution of queries were calculated. (**a**) ESR(Query)-CTSD(Class), (**b**) VDR(Query)-CTSD(Class), (**c**) COX2(Query)-CTSD(Class), and (**d**) ESR(Query)-CTSD(Class).

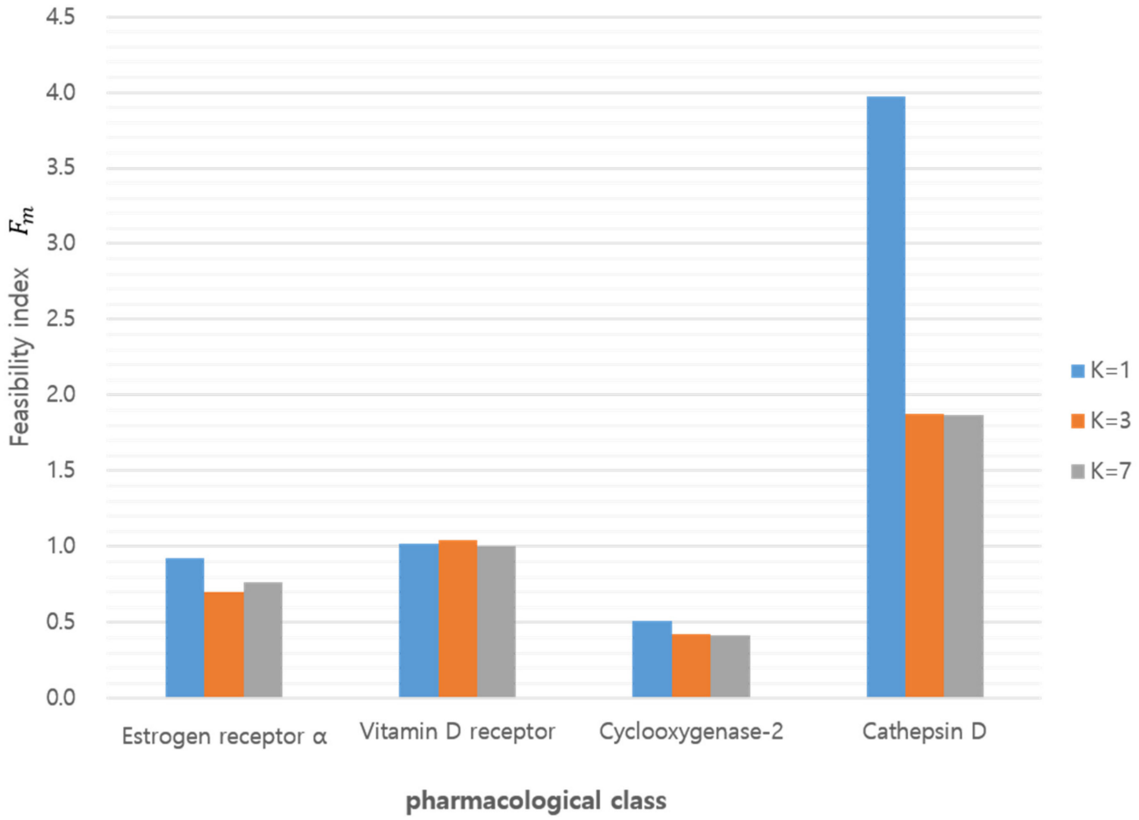


Figure 8. Feasibility index according to target class and GMM component (K).

Representative ligands for better discriminative predictions: According to the results described in Figures 4–7 and Table 4, 3D similarity-based K–L divergence together with $\mathbb{P}(\nu(l_m) = m)$ and F_m showed discriminative power with regard to some query–class associations. The question therefore remains regarding the efficient use of the 3D-chemocentric approach under the current discriminative power, where it can be applied to investigate the novel pharmacology of an unprecedented compound. For this purpose, K–L divergence of an unprecedented compound should be calculated to compare known ligands and target classes. In detail, representative ligands within each target class were chosen for the comparison. For example, we selected four representative queries based on their Tanimoto–Jaccard coefficients (x), and K-L divergence value, namely, (1) x is the nearest to the mean of the Q distribution (GMM, K = 1), (2) x is the nearest to an outlier of the Q distribution (mean ± 2SD), (3) the range of K–L divergence between two target classes, and (4) the highest similarity with an unprecedented compound (Table 4). As an example, BNDS-A, a recently reported in-house compound [7], was used as the unprecedented compound due to the absence of ChEMBL DB. The first query compound close to the mean of the Q distribution showed a smaller K–L divergence than the other compounds (Table 5). The initial assumption and initial selection of the target class of BNDS-A (in other words, the selection of the Q distribution), resulted in a critical effect on the K–L divergence of BNDS-A as a query compound to predict the 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 for the ESR Q distribution) and CHEMBL604989 (which exhibited the biggest K–L divergence gap), and was less ESR-like than CHEMBL499809 (at the mean for the ESR Q distribution) and CHEMBL2 (at the mean + 2SD). Under the Q of ESR assumption, BNDS-A showed the lowest K–L divergence with the VDR ligands (0.0588 of VDR < 0.2116 of ESR), suggesting that BNDS-A was more VDR ligand-like than ESR ligand-like. When the initial target was transferred to VDR or COX2, BNDS-A showed the lowest K–L divergence required to satisfy the assumption (chosen Q). In all BNDS-A rows of Table 4, while the order of K–L divergence of BNDS-A (VDR < ESR < CTSD) was retained under the assumed every target class of BNDS-A, COX2 showed the lowest K–L divergence under only COX2 *Q* distribution and did not show consistent prediction. Therefore, BNDS-A was more VDR ligand-like than COX2 ligand-like. Experimentally, BNDS-A regulated the expression level of targets in a concentration-dependent manner (VDR > CTSD >> ESR) [7]. Notably, K–L divergence of 3D similarity distributions can be an additional comparison method of known methods to predict the target of a novel compound such as (1) the rank of 3D similarity score [7,15,16] or (2) p-value of one 3D similarity distribution [20]. Whenever achieving the relevance between a novel query and a target class is the aim, K–L divergence can be used for 3D-chemocentric informatics, as seen in the example of BNDS-A.

Table 5. 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 BNDS-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 BNDS-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 BNDS-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 BNDS-A	0.0835	0.1467	0.0952	0.0129
	BNDS-A	Unknown	**0.0556**	0.26421	0.2092	0.087

4. Materials and Methods

Data collection: All data, except for the in-house compound (BNDS-A), were extracted from the ChEMBL database (1. ESR, VDR, COX2, and CTSD: version 23 through KNIME community node, 2. HIV1, HSP90, TRPV4, and TOP1: version 25 through MySQL) [37]. Version 23 was available in both the ChEMBL community node of KNIME and in-house MySQL built from the dump file from ChEMBL ftp (ftp://ftp.ebi.ac.uk/pub/databases/chembl/ChEMBLdb/releases/). HIV1 protease, HSP90, TRPV4, and TOP1 data were chosen based on the literature [36] and downloaded from the ChEMBL 25 version.

Conformational sampling: Extracted compounds were converted from 2D structures into 3D conformation using Omega of the Openeye software [38] under the following conditions: (1) the MMFF94 force field excluding Coulomb interactions and the attractive part of Van der Waals interactions (option: *mmff94s_Trunc*) to retain the forces: bonding stretching, angle bending, stretch-bend interaction, out-of-plane bending at tricooridnate centers, torsion interaction, and the repulsive part of Van der Waals interactions; (2) 15 kcal/mol as the energy window; (3) hydrogen deletion from the input file fragment prior to the substructure search (option: *deleteFixHydrogens*); (4) permission to generate stereoisomers; and (5) maximum acceptable number of rotatable bonds of 25 [39]. Due to computational burden and space limitation to write similarity into a matrix during calculation at posterior work, 3D structures of every compound were merged into the structure files (file extension: sdf) according to target class, and 13,957 3D structures (with duplication due to different conformation) from the files were chosen via stratified sampling in KNIME to produce the dataset for similarity matrices as shown in Supplementary Table S1.

Alignment method: In order to align the 3D-structures of compound pairs, center of the mass was used [40]. In detail, it is reported that SIMPLEX algorithm for the alignment is already implemented in ROCS [15]. Shape Toolkit in the Openeye software [40,41] provides '*OEBOOrientation*' used in *OEBestOverlay*. To optimize the alignment of each paired 3D structures, the starting point should be chosen before finding centers-of-mass of two conformers and *OEBestOverlay* uses an inertial frame alignment method to decide on starting positions by default. Under the default condition ('*OEBOOrientation_Inertial*'), the first 3D structure (refmol in the python code in the Supplementary Materials) was aligned by its principal moments of inertia, then the second structure (fitmol in the python code in the Supplementary Materials) object was aligned in four positions with the primary and secondary moments of inertia in both possible directions. Therefore, the alignment of a compound pair (A, B) is approximately the same and absolutely not identical with the alignment (B, A).

3D Descriptors: In order to describe a molecular shape, atom-centered Gaussian sphere model was implemented in OE-MPI/ROCS and the Shape Toolkit [40,41]. OE-MPI, a kind of MPI (message passing interface), was also provided by Openeye for thread parallel calculation with a high number of CPUs. The Gaussian sphere model describing the 3D shape of compounds used the sum of Gaussian functions of individual heavy atoms except for hydrogen. f and g are characteristic functions to present the 3D atomic structure of each compound, I: self-volume overlaps of each entity, independent; O: the overlap between the two functions, dependent on orientation of two molecules.

$$\text{Shape}(f,g) = \sqrt{\int [f(x,y,z) - g(x,y,z)]^2 dV} \tag{27}$$

$$\text{Shape}(f,g)^2 = \int [f(x,y,z)]^2 dV + \int [g(x,y,z)]^2 dV - 2\int f(x,y,z)g(x,y,z)dV$$

$$\text{Shape}(f,g) = I_f + I_g - 2O_{f,g}$$

$$\text{Jaccard–Tanimoto coefficient of Shape}(f,g) = \frac{O_{f,g}}{I_f + I_g - O_{f,g}}$$

Color features of every query were generated under the default algorithm of the Shape Toolkit. Color features were defined by pharmacophore types (H-bond donor, H-bond acceptor, negative charge, positive charge, hydrophobic, and ring) in a color force field (*Implicit Mills Dean*) and color atoms were described by Gaussian functions as being relatively hard with a steep gradient.

3D Similarity matrix: The Jaccard–Tanimoto coefficient of two features, shape and color were calculated, combined, and written into 3D similarity matrices using the functions in the supplementary python script [42].

- OEOverlay(): optimization of the alignment(overlap) between query and database
- OEBestOverlayScoreIter(): sorting all scores to highest Tanimoto coefficient before writing similarity score into an empty matrix.

In this study, while the dimension of 3D similarity matrices for *Q* distributions (GMM) was 13,957 by 13,957, the dimension of 3D similarity matrices for query distributions (ML estimation) was 1 by 13,957. Every sampled compound of four target classes (13,957 conformers x four target classes) was used as the query to show the performance of K–L divergence. The BNDS-A compound is only one query not existing in any target class.

Script for K–L divergence. In order to realize (1) the GMM model, (2) the ML estimation, and (3) K–L divergence, python scripts were written using python libraries such as pandas [43], numpy [44], and scipy [45] under anaconda installation [46], so that every code was uploaded to GitHub [47].

5. Conclusions

We developed a quantitative method comparing query compounds to target classes. The discriminative comparison was achieved by K–L divergence of 3D similarity distributions. The distributions were generated from 3D structures (sampled multi-conformers) with target annotation and optimized with parameters to best fit to frequent histograms. The feasibility index, F_m, and the probability, $P(\nu(l_m) = i)$, derived from the K–L divergence demonstrates the discrimination of queries against target classes. The feasibility index resulting from the GMM ($K = 1$) showed better feasibility for class discrimination than the GMM ($K = 3$) and ($K = 7$). Among the target classes, CTSD showed the most desirable feasibility and COX2 was the least desirable target for chemocentric informatics. K–L divergence comparison of an unprecedented compound, BNDS-A showed the consistent order of K–L divergence of BNDS-A (VDR < ESR < CTSD) under different target assumptions of BNDS-A so that our method is applicable for discriminative predictions of unknown query compounds in chemocentric informatics. This study will contribute to 3D chemocentric target deconvolution for unprecedented drug scaffolds. In the recent future, this quantitative method should be further studied with regard to the field of chemical optimization between the chemical space and pharmacological space.

Supplementary Materials: Supplementary Materials can be found at http://www.mdpi.com/1422-0067/21/12/4208/s1.

Author Contributions: M.-h.K. conceived and designed for the study. S.-H.L. and S.A. carried out all computational experiments. M.-h.K., S.-H.L., and S.A. analyzed all the data. M.-h.K. and S.-H.L. wrote the manuscript and M.-h.K. revised it. M.-h.K. provided the research work facility and funding. All authors have read and agreed to the published version of the manuscript.

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).

Acknowledgments: The authors would like to thank OpenEye Scientific Software providing the academic free license.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

ESR	Estrogen receptor alpha
VDR	Vitamin D receptor
COX2	Cyclooxygenase-2
CTSD	Cathepsin D

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. [CrossRef] [PubMed]
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. [CrossRef]
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. [CrossRef] [PubMed]
4. Eckert, H.; Bajorath, J. Molecular similarity analysis in virtual screening: Foundations, limitations and novel approaches. *Drug Discov. Today* **2007**, *12*, 225–233. [CrossRef] [PubMed]
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. [CrossRef]
6. Taylor, R.D.; MacCoss, M.; Lawson, A.D. Rings in drugs: Miniperspective. *J. Med. Chem.* **2014**, *57*, 5845–5859. [CrossRef]
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.; et al. 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. [CrossRef]
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. [CrossRef]
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. [CrossRef]
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. [CrossRef]
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. [CrossRef] [PubMed]
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.* **2019**, *30*, 911–923. [CrossRef]
13. 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. [CrossRef]
14. 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. [CrossRef] [PubMed]
15. Kumar, A.; Zhang, K.Y. Advances in the development of shape similarity methods and their application in drug discovery. *Front. Chem.* **2018**, *6*, 315. [CrossRef] [PubMed]
16. Shin, W.-H.; Zhu, X.; Bures, M.G.; Kihara, D. Three-dimensional compound comparison methods and their application in drug discovery. *Molecules* **2015**, *20*, 12841–12862. [CrossRef]
17. Lo, Y.-C.; Senese, S.; Damoiseaux, R.; Torres, J.Z. 3D chemical similarity networks for structure-based target prediction and scaffold hopping. *ACS Chem. Biol.* **2016**, *11*, 2244–2253. [CrossRef]
18. Seo, S.; Lee, T.; Kim, M.H.; Yoon, Y. Prediction of side effects using comprehensive similarity measures. *BioMed Res. Int.* **2020**, *2020*, 1–10. [CrossRef]
19. Méndez-Lucio, O.; Kooistra, A.J.; Graaf, C.D.; Bender, A.; Medina-Franco, J.L. Analyzing multitarget activity landscapes using protein–Ligand interaction fingerprints: Interaction cliffs. *J. Chem. Inf. Model.* **2015**, *55*, 251–262. [CrossRef]
20. Pérez-Nueno, V.I.; Venkatraman, V.; Mavridis, L.; Ritchie, D. Detecting drug promiscuity using gaussian ensemble screening. *J. Chem. Inf. Model.* **2012**, *52*, 1948–1961. [CrossRef]
21. Maaten, L.V.D.; Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579–2605.
22. Hershey, J.R.; Olsen, P.A. Approximating the Kullback Leibler Divergence Between Gaussian Mixture Models. In Proceedings of the 2007 IEEE International Conference on Acoustics, Speech and Signal Processing—ICASSP '07, Honolulu, HI, USA, 15–20 April 2007.
23. Kullback, S.; Leibler, R.A. On information and sufficiency. *Ann. Math. Stat.* **1951**, *22*, 79–86. [CrossRef]

24. Burnham, K.P.; Anderson, D.R. Kullback-Leibler information as a basis for strong inference in ecological studies. *Wildl. Res.* **2001**, *28*, 111–119. [CrossRef]
25. Nalewajski, R.F.; Parr, R.G. Information theory, atoms in molecules, and molecular similarity. *Proc. Natl. Acad. Sci. USA* **2000**, *97*, 8879–8882. [CrossRef] [PubMed]
26. Koller, D.; Sahami, M. Toward Optimal Feature Selection. In Proceedings of the Thirteenth International Conference on Machine Learning, Bari, Italy, 3–6 July 1996; pp. 284–292.
27. Kümmerer, M.; Wallis, T.S.; Bethge, M. Information-theoretic model comparison unifies saliency metrics. *Proc. Natl. Acad. Sci. USA* **2015**, *112*, 16054–16059. [CrossRef]
28. Duchi, J. Derivations for linear algebra and optimization. *Berkeley Calif.* **2007**, *3*, 2325–5870.
29. McLachlan, G.J.; McGiffin, D.C. On the role of finite mixture models in survival analysis. *Stat. Methods. Med. Res.* **1994**, *3*, 211–226. [CrossRef]
30. Singh, R.; Pal, B.C.; Jabr, R.A. Statistical representation of distribution system loads using gaussian mixture model. *IEEE Trans. Power Syst.* **2009**, *25*, 29–37. [CrossRef]
31. Duda, R.O.; Hart, P.E.; Stork, D.G. *Pattern Classification and Scene Analysis*, 2nd ed.; John Wiley & Sons: Hoboken, NJ, USA, 1995.
32. 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**, *39*, 1–22. [CrossRef]
33. Hartley, H.O. Maximum likelihood estimation from incomplete data. *Biometrics* **1958**, *14*, 174–194. [CrossRef]
34. McLachlan, G.; Krishnan, T. *The EM Algorithm and Extensions*; John Wiley & Sons: Hoboken, NJ, USA, 2007; Volume 382.
35. Bajusz, D.; Rácz, A.; Héberger, K. Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? *J. Cheminf.* **2015**, *20*, 1–13. [CrossRef] [PubMed]
36. Montaruli, M.; Alberga, D.; Ciriaco, F.; Trisciuzzi, D.; Tondo, A.R.; Mangiatordi, G.F.; Nicolotti, O. Accelerating Drug Discovery by Early Protein Drug Target Prediction Based on a Multi-Fingerprint Similarity Search. *Molecules* **2019**, *24*, 2233. [CrossRef]
37. Berthold, M.R.; Cebron, N.; Dill, F.; Gabriel, T.R.; Kötter, T.; Meinl, T.; Ohl, P.; Thiel, K.; Wiswedel, B. KNIME—the Konstanz information miner. *ACM SIGKDD Explor. Newsl.* **2009**, *11*, 26–31. [CrossRef]
38. OpenEye Scientific. *OMEGA Software (ver. 2.4.6)*; OpenEye Scientific: Santa Fe, NM, USA, 2015. Available online: https://www.eyesopen.com/omega (accessed on 18 May 2020).
39. 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**, *9*, 21. [CrossRef] [PubMed]
40. 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. [CrossRef]
41. OpenEye Scientific. *Shape TK Software (ver. 1.9.3)*; OpenEye Scientific: Santa Fe, NM, USA, 2015. Available online: https://www.eyesopen.com/shape-tk (accessed on 18 May 2020).
42. Shape Toolkit 2.0.4. Available online: https://docs.eyesopen.com/toolkits/python/shapetk (accessed on 18 May 2020).
43. Pandas documentation, Version: 1.0.4. Available online: https://pandas.pydata.org/docs/ (accessed on 18 May 2020).
44. NumPy v1.18 Manual. Available online: https://numpy.org/ (accessed on 18 May 2020).
45. SciPy. Available online: https://www.scipy.org/ (accessed on 18 May 2020).
46. Anaconda.Documentation. Available online: https://docs.anaconda.com/anaconda/install/ (accessed on 18 May 2020).
47. GitHub. Available online: https://github.com/college-of-pharmacy-gachon-university/KLD-Pharmacological_Class_Similarity (accessed on 18 May 2020).