Research Article

www.molinf.com

molecular informatics

doi.org/10.1002/minf.202100118

Quasi-Supervised Strategies for Compound-Protein Interaction Prediction

Onur Çakı[a] and Bilge Karaçalı*[a]

Abstract: In-silico compound-protein interaction prediction addresses prioritization of drug candidates for experimental biochemical validation because the wet-lab experiments are time-consuming, laborious and costly. Most machine learning methods proposed to that end approach this problem with supervised learning strategies in which known interactions are labeled as positive and the rest are labeled as negative. However, treating all unknown interactions as negative instances may lead to inaccuracies in real practice since some of the unknown interactions are bound to be positive interactions waiting to be identified as such. In this study, we propose to address this problem using the Quasi-Supervised Learning (QSL) algorithm. In this framework, potential interactions are predicted by estimating the overlap between a true positive dataset of compound-protein pairs with known interactions and an unknown dataset of all the remaining compound-protein pairs. The potential interactions are then identified as those in the unknown dataset that overlap with the interacting pairs in the true positive dataset in terms of the associated similarity structure. We also address the class-imbalance problem by modifying the conventional cost function of the QSL algorithm. Experimental results on GPCR and Nuclear Receptor datasets show that the proposed method can identify actual interactions from all possible combinations.

Keywords: Compound-Protein Interactions · Compound Similarity · Chemoinformatics · Drug Discovery · Machine Learning

1 Introduction

Identification of compound-protein interactions (CPI) plays an essential role in a wide range of pharmacological applications. The initial step of drug discovery is to detect effective interactions between drug candidate compounds and a target protein.[1] A number of studies have shown that complex diseases such as cancer and Alzheimer's Disease are associated with multiple targets necessitating the elucidation of the interaction profiles of candidate drugs with more target proteins.[2] Identification of such interactions also allows predicting undesired side-effects of drugs by detecting off-target interactions.[3] Furthermore, it is a key part of drug repositioning, i.e. discovering new clinical usage of existing drugs.[4]

The experimental validation of compound-protein interactions in laboratory environments remains time-consuming, laborious and extremely costly even when using high-throughput screening technologies. As a result, only a small number of experimentally validated interacting compound-protein pairs exist compared to the large numbers of compounds and proteins: There are ~1.5 million human protein sequences out of which ~20000 are reviewed and ~96 million compounds in the databases of NCBI Entrez system against only ~1.2 million recorded interactions.[5] In recent years, there has been growing interest in using computational tools for CPI prediction. In-silico prediction of CPI aims to narrow the search space for future wet-lab experiments by suggesting the most likely interactions, thereby accelerating pharmacological research processes, decreasing costs, and increasing research productivity.[6]

There are three main computational approaches in virtual screening for potential compound-protein interactions. Structure-Based Virtual Screening (SBVS) aims to utilize the 3D structure of a target protein to determine whether or not a compound would interact with the target protein.[7] The disadvantage of this approach is that obtaining the 3D structure of a target protein may not always be possible, especially for membrane proteins such as Ion Channels and GPCRs.[8] In Ligand-Based Virtual Screening (LBVS) that relies on Chemical Similarity Principle stating that shared structural elements may indicate similar bioactivity, potential interactions are identified by comparing the structure of compounds that are known to interact with a target protein against candidate compounds.[9] However, this approach becomes unfeasible if the target protein of interest has few or no known interactions. Although there are some studies on integrating LBVS and SBVS in which LBVS methods are mostly used as prefiltering

[a] *O. Çakı, B. Karaçalı*
Electrical and Electronics Engineering Department,
Izmir Institute of Technology,
Urla, Izmir 35430, Turkey
phone/fax: +90(232)750 6534
+90(232)750 6599
E-mail: bilgekaracali@iyte.edu.tr

Supporting information for this article is available on the WWW under https://doi.org/10.1002/minf.202100118

for more time-consuming SBVS methods, they do not compensate for the mentioned drawbacks of each approach.[10] Finally, Chemogenomics-Based Virtual Screening (CGBVS) aims to address the issues associated with the earlier two approaches.[8] The idea behind this approach is again that the compounds that have similar structure would tend to interact with same or similar proteins, but unlike the LBVS, information that comes from both compounds and proteins are considered simultaneously. In this way, CGBVS aims to compensate for the lack of known interactions of target proteins by considering the known interactions of similar proteins, and to develop a unifying prediction model for the whole compound-protein data at hand.

To date, various machine learning-based methods have been proposed for CPI prediction based on CGBVS.[11–13] These methods can be categorized further into feature-vector based approaches and similarity-based approaches. In feature-vector based approaches, compound-protein pairs are represented by fixed-length feature vectors that are used as input to a machine learning algorithm. For example, Radkar et al. (2020) construct the feature vector for a compound-protein pair by combining the two feature vectors, one from the protein, and the other from the compound. They selected the features that may be important for CPI using Wrapper Feature Selection to overcome high dimensionality of the pair space.[14] However, feature extraction is a challenging process especially when it comes to CPI prediction due to complex relationships between the chemical and the genomic spaces. Since many factors may affect the establishment of interaction in a given compound-protein pair, fixed-length vectors may not adequately reflect the critical pharmacological properties. In similarity-based approaches, machine learning algorithms can be constructed to evaluate compound-compound similarities and protein-protein similarities to predict interactions of compound-protein pairs. In the literature, machine learning algorithms based on similarity offer promising results when the problem has unstructured data.[15] It is also important to note that this strategy is inherently suitable for a CGBVS method as structural similarities that are key for molecular interaction may not necessarily be represented adequately through numeric features.[11]

In a similarity-based scheme, Yamanashi et al. (2008) approached the CPI prediction problem as link prediction in a bipartite graph. They used compound-compound similarities and protein-protein similarities to embed them into a pharmacological vector space in which the Euclidean distances between linked vectors are minimized.[16] Jacob and Vert (2008) developed a pairwise kernel method to obtain a similarity matrix for compound-protein pairs from similarities between compounds and similarities between proteins. They then trained a Support Vector Machine (SVM) classifier using this similarity matrix as a kernel matrix.[17] Laarhoven et al. (2011) treated interaction profiles of each protein and each compound as binary feature vectors. They constructed similarity matrices from these vectors using a Gaussian kernel and integrated them with a compound similarity matrix and a protein similarity matrix. A predicted interaction score matrix was calculated from these combined similarities using Regularized Least Squares (RLS).[18] Laarhoven and Marchiori (2013) later expanded Gaussian Interaction Profile kernels with a Weighted Nearest Neighbor approach to predict interactions for new proteins and compounds for which no known interactions exist.[19] Gönen (2012) combined non-linear dimension reduction and matrix factorization to project compounds and proteins into a unified low-dimensional space through their similarity matrices and estimate an interaction matrix.[20] Zheng et al. (2013) used Collaborative Matrix Factorization to estimate a binary interaction matrix between compounds and proteins in such a way that the latent features of the matrix approximate protein and compound similarity matrices.[21] Several recent studies proposed new machine learning techniques for CPI prediction problem. Reker et al. (2017) construct a model from machine-picked informative samples using active learning method instead of fitting the model to whole data.[22] Tsubaki et al. (2019) used a unified deep learning model in which a graph neural network screens the chemical space and a convolutional neural network screens the genomic space in order to predict the interaction profile of a given pair.[23] In all these studies, the CPI prediction problem is addressed within a Supervised Learning framework in which known interactions are labelled as positive, and everything else is labelled as negative. However, treating the compound-protein pairs that have no known interactions as negative leads to unrealistic recognition models as these pairs undoubtedly include some positive interactions that are as of yet unknown. To address this issue, Liu et al. (2015) construct a dataset with highly credible negative samples that are selected from unlabeled compound-protein pairs based on the assumption that "the proteins dissimilar to every known/predicted target of a compound are not much likely to be targeted by the compound and vice versa".[24] However, there is a notable lack of studies to tackle this problem in a realistic manner.[12]

The inherent issue with current machine learning approaches is that the true negative samples of non-interacting compound-protein pairs are rarely found in experimental literature. Even the most notable datasets, such as DUD and DUD-E[25] that are designed as benchmark datasets for molecular docking programs suffer from this problem. On the other hand, Tox21, which is constructed to provide a comparison of toxicity prediction models, is a dataset where all compounds were tested versus all proteins.[26,27] However, such datasets are rare and usually missing for general CPI instances. Since conventional classification-based strategies require a true negative dataset to contrast with the true positive dataset of known interactions, the only option is to manufacture true

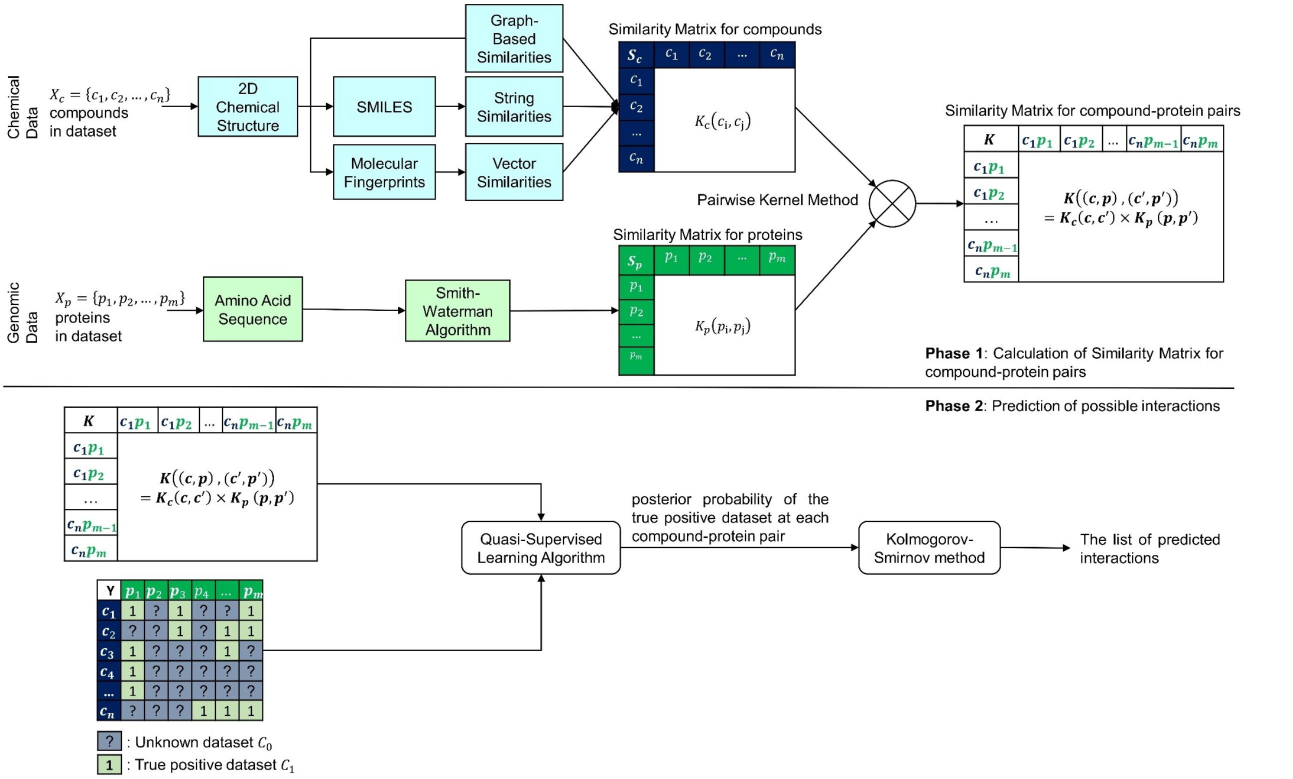

Chemical Data
$X_c = \{c_1, c_2, \dots, c_n\}$ compounds in dataset
2D Chemical Structure
SMILES
Molecular Fingerprints
Graph-Based Similarities
String Similarities
Vector Similarities
Similarity Matrix for compounds
S_c c_1 c_2 … c_n
$K_c(c_i, c_j)$
Pairwise Kernel Method
Similarity Matrix for compound-protein pairs
K c_1p_1 c_1p_2 … c_np_{m-1} c_np_m
$K((c,p),(c',p')) = K_c(c,c') \times K_p(p,p')$
Genomic Data
$X_p = \{p_1, p_2, \dots, p_m\}$ proteins in dataset
Amino Acid Sequence
Smith-Waterman Algorithm
Similarity Matrix for proteins
S_p p_1 p_2 … p_m
$K_p(p_i, p_j)$
Phase 1: Calculation of Similarity Matrix for compound-protein pairs
Phase 2: Prediction of possible interactions
Quasi-Supervised Learning Algorithm
posterior probability of the true positive dataset at each compound-protein pair
Kolmogorov-Smirnov method
The list of predicted interactions
Y p_1 p_2 p_3 p_4 … p_m
c_1 1 ? 1 ? ? 1
c_2 ? ? 1 ? 1 1
c_3 1 ? ? ? 1 ?
c_4 1 ? ? ? ? ?
… 1 ? ? ? ? ?
c_n ? ? ? 1 1 1
? : Unknown dataset C_0
1 : True positive dataset C_1

Figure 1. Schematic diagram of the proposed method.

negative datasets from pairings of existing compounds and proteins. This, however, entails several additional issues: Firstly, since many different true negative datasets can be manufactured based on different principles of non-interaction, classifier outputs incur a conditional bias on the selected true negative dataset and differ depending on the choice of the true negative dataset. Secondly, the presence of the unknown positive interactions in manufactured true-negative datasets of effectively untested interactions contaminates the inferred interaction recognition mechanism. In the absence of a validated true negative dataset of non-interacting compound-protein pairs, the only viable option for an unbiased and uncontaminated machine learning strategy is to contrast the set of previously untested interactions containing all possible pairings between the compounds and proteins at hand with the true positive dataset of interacting pairs and seek those pairs in the unknown and effectively untested dataset that differentiate from the rest towards the positive interactions.

In this paper, we use the quasi-supervised learning algorithm[28] to contrast the true positive dataset with known compound-protein interactions against the untested dataset of all possible pairings. For machine learning purposes, we define a similarity between compound-protein pairs from protein-protein similarity and compound-compound similarity measures and apply the quasi-supervised learning algorithm on the combined similarity measure to calculate estimates for the posterior probability of a given compound-protein pair to belong to the true positive dataset, for all pairs in both datasets. Finally, we determine the optimal threshold for predicted positive interactions in the untested dataset using Kolmogorov-Smirnov statistics applied on the posterior probability estimates.

This paper is organized as follows. We describe the details of the proposed method for protein-compound interaction prediction using the quasi-supervised learning algorithm in the next section along with the various techniques with which we characterize the similarity between compound and protein pairs, and the protein-compound interaction datasets used in this study. We provide descriptions for each operational block shown in the schematic diagram of the proposed method in Figure 1. We present the results of a comparative analysis of the proposed method across different configurations involving alternative similarity measures and alternative approaches from the literature in the Results Section. We conclude the paper with a discussion on a general evaluation of the proposed method and several potential extensions for future work.

2 Material and Methods

2.1 Dataset

In this study, we used the publicly available dataset published by Yamanishi et al.[16] In this widely referenced paper, the authors point out that screening all compound-protein pairs is computationally infeasible, and construct a modular dataset to build machine learning models separately for four major protein classes (i.e. enzymes, ion channels, GPCRs, and nuclear receptors) which are commonly considered as drug targets. This dataset has since become a benchmark in CPI prediction studies.[12] The interaction information between compound-protein pairs were retrieved from DrugBank,[29] KEGG,[30] BRENDA,[31] and SuperTarget[32] databases by Yamanishi. Table 1 shows the number of proteins and compounds and known interactions between all possible compound-protein pairs for each protein class dataset in the collection.

Table 1. Datasets of Yamanishi.[14]

Protein Class Dataset	Enzyme	Ion Channel	GPCR	Nuclear Receptor
Compound	445	210	223	54
Protein	664	204	95	26
Interaction	2926	1476	635	90
Fraction of Annotated Interaction	%0.99	%3.45	%2.10	%6.41

2.2 Chemical Data

We retrieved the chemical structures of compounds in mol format from KEGG DRUG database.[30] Similarity matrices for chemical space that evaluate the similarity between different compounds, denoted by a matrix $\boldsymbol{S_c}$ of compound-compound similarity, are constructed using a variety of methods. The methods used in this study can be classified into three main categories: Graph-based methods, SMILES-based methods and Molecular Fingerprints-based methods.

2.2.1 Graph-Based Methods

SIMCOMP[33] algorithm is used to calculate the chemical structural similarity between compounds. This algorithm treats the 2D structure of compounds as graphs in which atoms are mapped to vertices and bonds are mapped to edges. The vertices are labelled with 68 KEGG atom types instead of the usual atomic species. The KEGG atom types consist of three letters: The first letter corresponds to the element symbol of the atom, while the second and third letters indicate its hierarchical classification depending on its hybrid orbital and atomic environments. These micro-environments are introduced in order to distinguish molecules in a biochemical manner in addition to their structures. The algorithm finds the maximum common subgraph between two compound graphs and then calculates a similarity score using the Jaccard coefficient, defined by

$$\boldsymbol{S_c}(G,G') = \frac{|G\cap G'|}{|G\cup G'|} = \frac{MCS(G,G')}{|G|+|G'|-MCS(G,G')} \qquad (1)$$

where the intersection and the union operations between graphs G and G' are defined as the maximum common subgraph and the nonredundant subgraph, respectively. In addition, the $|.|$ operator calculates the cardinality of its argument graph.

In our implementation of the algorithm, the maximum common subgraph was found using RDKit chemoinformatics library[34] where vertices are labelled by atomic species instead of 68 KEGG atoms. In addition to their types, vertices are also distinguished by their valance and bonds are distinguished by their aromaticity and ring information.

2.2.2 SMILES-Based Methods

Simplified Molecular Input Line Entry System (SMILES) is a 1D string representation that encodes the structural information of compounds.[35] A Study by Öztürk et al. (2016) suggested that text similarity between two SMILES strings can be considered as a measure of structural similarity between two compounds for compound-protein interaction prediction tasks.[36] They showed that similarity measures using various SMILES kernels performed as well as graph-based methods with an additional computational time advantage. We used molconverter console program of JCHEM (developed by ChemAxon, https://www.chemaxon.com/) to convert MOL files into canonical SMILES as in the original paper.[36] The program defines SMILES by following Daylight's SMILES specification rules.[37]

We used Normalized Longest Common Subsequence (NLCS), Combination of Longest Common Subsequence Models (CLCS), LINGO-*q* Similarity, LINGO Based Term Frequency (TF) Cosine Similarity, and LINGO Based Term Frequency-Inverse Document Frequency (TF-IDF) Cosine Similarity which are proposed by Öztürk et al. to calculate similarity between two compounds.[36] In NLCS, they find the longest common subsequence between two SMILES strings and calculate a similarity score by cosine normalization. Note that the longest common subsequence is not required to be consecutive. In order to achieve a more meaningful semantic similarity between two strings, they also find Maximal Consecutive Longest Common Subsequence (MCLCS) starting from the first and the *n*'th characters. CLCS is then defined as the equal weighted average of their cosine normalizations and NLCS.

LINGO-q stands for consecutive q-character substrings that can be created from a SMILES string.[38] For instance, LINGOs($q = 4$) that can be extracted from the SMILES string of Gabapentin "NCC1(CC(O)=O)CCCCC1" are {'NCC0', 'CC0(', 'C0(C', …, 'CCC0'}. Note that all ring numbers must be replaced with 0 s before the LINGO extraction process. A similarity function, LINGOsim,[36] then calculates a similarity score between the two SMILES strings using the unique LINGOs that are extracted from them. We used $q = 3, 4, 5$ as in the original study.[36] In order to calculate a similarity score between two SMILES strings, the strings are also mapped into vectors whose length equals the total number of unique LINGOs in the two strings. In Lingo-based Term Frequency (TF) cosine similarity, the TF of each unique LINGO reflects the occurrence frequency of the LINGO in SMILES and are collected into feature vectors. In LINGO based TF-IDF similarity, the TF values of LINGOs are multiplied with their Inverse Document Frequency that reflects the occurrence frequency of the LINGOs in the whole SMILES dataset and then collected into reference vectors. TF assigns higher values to more frequently occurring LINGOs in a SMILES, while IDF assigns lower values to more frequently occurring LINGOs in the dataset. Cosine similarity between two vectors then provides a similarity score between the two SMILES strings. These processes are applied to all possible SMILES pairs in the dataset to construct a similarity matrix, $\boldsymbol{S}_c$, between all compounds.

2.2.3 Molecular Fingerprints-Based Methods

Fingerprints encode the structure of compounds into fixed-length bit vectors depending on whether a substructure occurs in a compound or not. A study by Sawada et al. (2014) investigated the performance of different types of fingerprints and similarity functions in compound-protein interaction prediction problems.[39] In our study, we used Extended-Connectivity Fingerprints (ECFP),[40] Functional-Class Fingerprints (FCFP),[40] Molecular ACCess System (MACCS) fingerprints,[41] KEGG Chemical Function and Substructures (KCF–S) descriptors[42] that have previously been identified as useful in CPI prediction.[39]

ECFP describes the structure of a molecule by encoding substructures formed by a circular neighborhood of each atom within an atom radius into 1024-length binary vectors, an approach that is also known as a circular fingerprint, or Morgan Fingerprints.[34] We set the radius as 2 providing a maximum range between fingerprint atoms of 4 (ECFP4). FCFP is an extension of ECFP in which pharmacophore roles of atoms are also added to fingerprints to encode for functional substructural features instead of just atom environments. The MACCS fingerprints describe the structure of a molecule with a 166-length bit vector whose elements correspond to a substructure key. These publicly

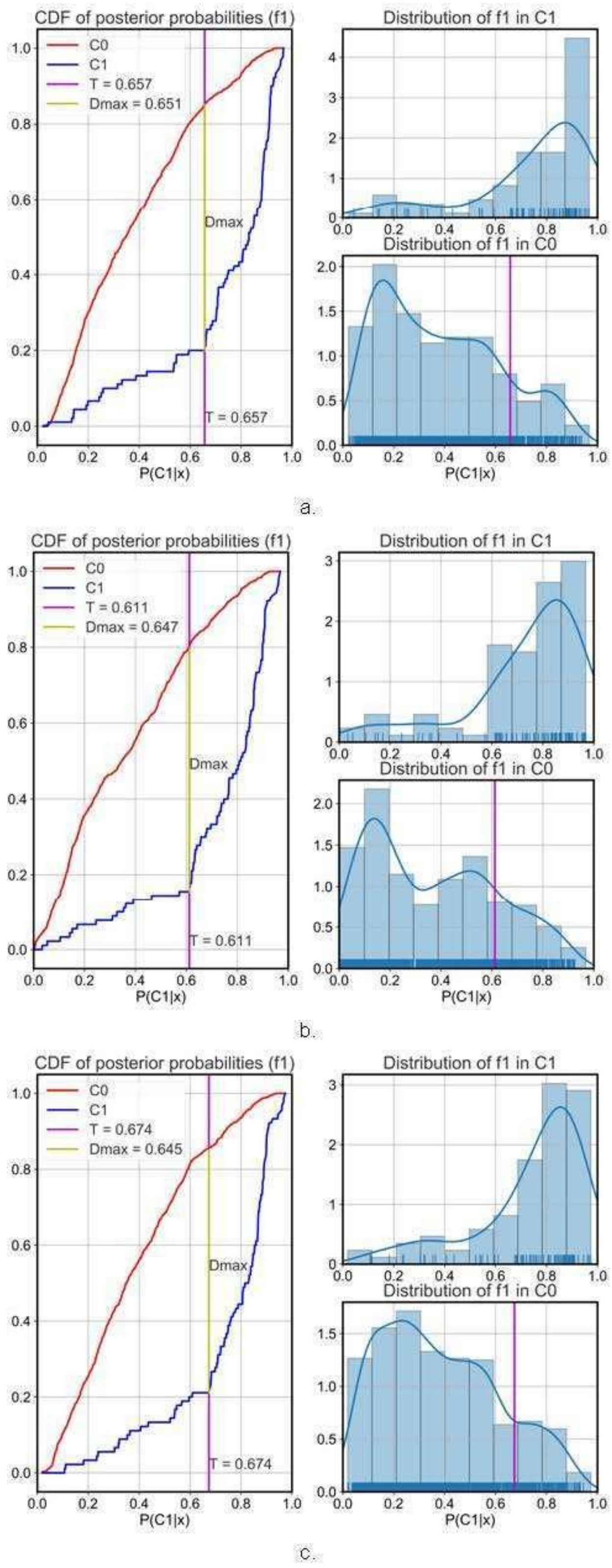


Figure 2. Posterior probability distributions of compound-protein pairs in the Nuclear Receptor dataset to belong to the true positive dataset and their Kolmogorov-Smirnov Analysis for (a) KCF–S Fingerprints, (b) Maximum Common Substructure (RDkit), and (c) LINGO based TF-IDF Cosine Similarity.

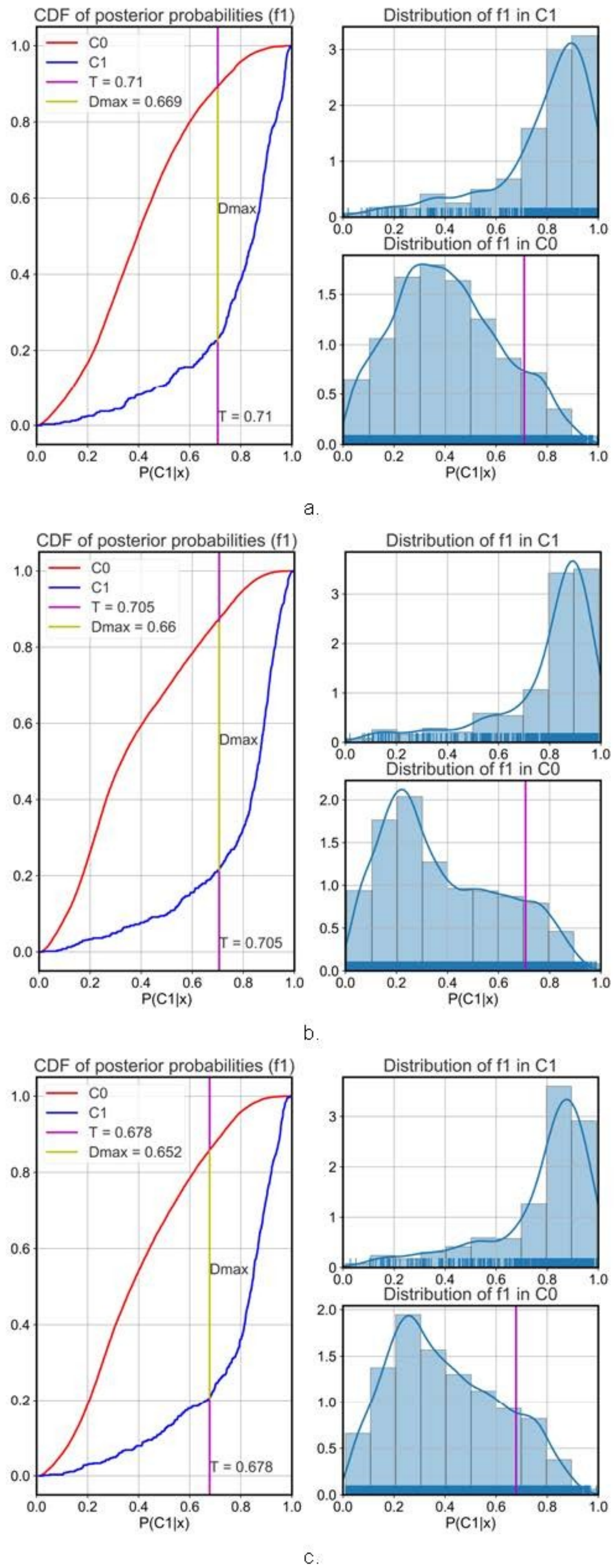


Figure 3. Posterior probability distributions of compound-protein pairs in the GPCR dataset to belong to the true positive dataset and their Kolmogorov-Smirnov Analysis for (a) LINGO based TF-IDF Cosine Similarity, (b) LINGOsim ($q = 3$), and (c) LINGOsim ($q = 4$).

available sub-structure keys are developed by a private company (previously MDL Information Systems, now BIOVIA, at the URL address https://www.3ds.com/products-services/biovia/) in order to calculate a molecular similarity. We used the RDKit python library to construct these fingerprints and calculated similarity scores between each fingerprint pair using cosine similarity.

One known drawback of fingerprints is that they encode for only the presence or absence of substructures and disregard the copy number for multiply present substructures. KCF–S addresses this problem using integer-valued vectors of counts instead of binary vectors: It treats the 2D chemical structure of a compound as a graph and characterizes the structure by an integer-valued vector in which each element of the vector corresponds to the number of distinct copies of a substructure that the compound possesses. Moreover, instead of atomic species such as C, H, O, N, P, and so on, it uses the 68 KEGG atoms. Substructures are constructed from the graph of a compound using seven chemical structural attributes: atom, bond, triplet, vicinity, ring, skeleton and inorganic. The dimension of these vectors equals the number of unique substructures listed in a database of substructures that can be extracted from the compounds. We used KCF-Convoy python package[43] to construct fingerprint vectors and calculated the similarity between these vectors using a weighted Tanimoto similarity.[43]

2.3 Genomic Data

We retrieved the amino acid sequence information of all proteins in the dataset in FASTA format from KEGG GENES database.[30] Similarity Matrix for genomic space, denoted by a matrix $\boldsymbol{S_p}$ of protein-protein similarity, was constructed by calculating the similarity between each protein pair in the dataset using Normalized Smith-Waterman Algorithm. Smith-Waterman is a sequence alignment algorithm which returns an alignment score for the conserved regions between the two sequences.[44] Since these conserved regions can be responsible for common bioactivity and functional similarity, the Normalized Smith-Waterman algorithm offers a more biologically meaningful assessment compared to other sequence alignment algorithms. The alignment score is normalized as

$$\boldsymbol{S_p}(p_i, p_j) = \frac{SW(p_i, p_j)}{\sqrt{SW(p_i, p_i) \times SW(p_j, p_j)}} \tag{2}$$

to obtain a similarity score between 0 and 1 where $SW(p_i, p_j)$ denotes the alignment score of the Smith-Waterman algorithm between the amino acid sequences of proteins p_i and p_j. In this study, we used the default values for the parameters of the algorithm as in Pairwise Sequence Alignment Tool of EMBOSS (https://www.ebi.ac.uk/Tools/psa/emboss_water/).

2.4 Pairwise Kernel Method

A similarity matrix must satisfy Mercer's Theorem to be used in a machine learning algorithm as a kernel matrix, which means that it has to be symmetric and positive semi-definite, i.e. all eigen values must be non-negative.[15] In order to ensure that compound-compound and protein-protein similarity matrices satisfy these criteria, we firstly calculated symmetric and regularized compound and protein similarity matrices $\boldsymbol{K_c}$ and $\boldsymbol{K_p}$ by

$$\boldsymbol{K_p} = \frac{S_p + S_p{}^T}{2} + \left|\lambda_{min}\left(\boldsymbol{S_p}\right)\right| I \tag{3}$$

and

$$\boldsymbol{K_c} = \frac{S_c + S_c{}^T}{2} + \left|\lambda_{min}(\boldsymbol{S_c})\right| I \tag{4}$$

where the diagonal entries of the symmetric similarity matrices are augmented by the minimum eigenvalue of the corresponding compound or protein similarity matrices. We then used the pairwise kernel method[17] to calculate joint similarity between compound-protein pairs (c, p) and (c', p') as

$$K\left((c,p)\ ,(c^{'},p^{'})\right) = \boldsymbol{K_c}(c,c^{'}) \times \boldsymbol{K_p}(p,p^{'}) \tag{5}$$

which is tantamount to constructing a similarity matrix $\boldsymbol{K}$ between compound-protein pairs by the Kronecker product of $\boldsymbol{K_c}$ and $\boldsymbol{K_p}$ as (Jacob et al., 2008)

$$K = \boldsymbol{K_c} \otimes \boldsymbol{K_p}\ . \tag{6}$$

Note that several other similarity pair kernels could be used in order to determine the similarity structure between compound-protein pairs.[45] However, the current manuscript aims to introduce a new machine learning paradigm that can make interaction predictions in the absence of validated true negative interaction instances as well as based on various compound and protein similarity measures. Thus, while alternative similarity kernels can be incorporated into the proposed framework in future studies, an exhaustive evaluation of all possible kernels falls outside the scope of the manuscript.

2.5 The Quasi-Supervised Learning Algorithm

The Quasi-Supervised Learning Algorithm (QSL) was developed by Karacali (2010) to address one of the major problems of biomedical data analysis, the possible lack of ground-truth labeled data for a class of interest.[28] In this learning strategy, given a two-class recognition scenario with labeled samples of only one of the classes, the data at hand are divided into two datasets (C_0 and C_1): One dataset, say C_1, consists of the labelled samples of the known class, while the other one, C_0, consists of all samples without any label. A numerical machine learning algorithm then allows nonparametric estimation of posterior probability of each sample belonging to C_0 and C_1 using the asymptotic properties of nearest neighbor classification rule. Using the estimated posterior probabilities, we can estimate the overlap between C_0 and C_1 for automatic labelling of the samples in the unlabeled dataset C_0 that appear among C_1 samples.

Table 2. Performance Comparison of Compound Similarity Measure Methods.

Compound Similarity Measure Methods	D_{max} Nuclear Receptors	D_{max} GPCR	
ECFP4	0.587	0.604	Fingerprint
FCFP4	0.512	0.601	Fingerprint
KCF-S	**0.651**	0.642	Fingerprint
MACCS	0.475	0.518	Fingerprint
SIMCOMP	0.622	0.584	Graph
MCS – RDkit	**0.647**	0.635	Graph
NLCS	0.543	0.582	SMILES
CLCS	0.592	0.584	SMILES
LINGOsim ($q = 3$)	0.624	**0.660**	SMILES
LINGOsim ($q = 4$)	0.630	**0.652**	SMILES
LINGOsim ($q = 5$)	0.607	0.629	SMILES
LINGO based TF	0.604	0.646	SMILES
LINGO based TF-IDF	**0.645**	**0.669**	SMILES

The QSL algorithm can be described briefly as follows: Given M reference sets $\{R_1, R_2, \ldots, R_M\}$ for nearest neighbor classification constructed with n numbers of samples from each of C_0 and C_1, the average rate of assigning a sample x to C_0 and C_1using nearest neighborhood classification with reference to $R_1, R_2, \ldots, R_M$ will approximate the posterior probabilities. Mathematically, this can be expressed as

$$\begin{aligned} P(C_1|x) &\simeq f_1(x) \\ &= \frac{1}{M}\sum_{m=1}^{M} 1(x \text{ is assigned to } C_1 \text{ with reference to } R_m) \end{aligned} \tag{7}$$

and

$$\begin{aligned} P(C_0|x) &\simeq f_0(x) \\ &= \frac{1}{M}\sum_{m=1}^{M} 1(x \text{ is assigned to } C_0 \text{ with reference to } R_m) \end{aligned} \tag{8}$$

where the assignment of a sample x to C_0 and C_1 is made using a nearest neighbor classifier using the indicated reference set. Since carrying out great numbers of nearest neighbor classification is not feasible due to the associated computational expense, the Quasi-Supervised Learning Algorithm provides a fast and efficient numerical calculation of the rates $f_1(x)$ and $f_0(x)$ for each sample x_i in the

Table 3. The list of top 30 predicted positive interactions in the unknown dataset C_0 for the Nuclear Receptor Dataset. The identified positive interactions are indicated by the letter Y and potential interactions are indicated by the letter P, respectively.

Compound	Protein	Posterior Probability	SuperTarget	DrugBank	KEGG	ChEMBL
Nandrolone phenpropionate	estrogen receptor 1	0,92284				
Fluoxymesterone	progesterone receptor	0,92249				
Testosterone	**progesterone receptor**	**0,92149**	**Y**	**Y**		**Y**
Hydrocortisone	**progesterone receptor**	**0,91974**				**P**
Norethindrone	estrogen receptor 1	0,91877				
Spironolactone	**progesterone receptor**	**0,91811**		**Y**		**Y**
Nandrolone phenpropionate	**progesterone receptor**	**0,91511**				**P**
Eplerenone	**progesterone receptor**	**0,91061**	**Y**	**Y**		
Testosterone	**estrogen receptor 1**	**0,90441**	**Y**	**Y**		**Y**
Oxandrolone	**progesterone receptor**	**0,90432**				**P**
Budesonide	**progesterone receptor**	**0,90063**				**P**
Mifepristone	**estrogen receptor 1**	**0,9002**	**Y**			
Loteprednol etabonate	progesterone receptor	0,89722				
Amcinonide	progesterone receptor	0,88766				
Isotretinoin	**retinoic acid receptor beta**	**0,88641**			**Y**	**Y**
Pregnenolone	**progesterone receptor**	**0,88571**				**P**
Isotretinoin	**retinoic acid receptor gamma**	**0,87976**	**Y**	**Y**	**Y**	**Y**
Oxandrolone	estrogen receptor 1	0,87834				
Hydrocortisone	estrogen receptor 1	0,87657				
Dydrogesterone	estrogen receptor 1	0,86944				
Spironolactone	**estrogen receptor 1**	**0,8678**				**Y**
Ethinyl estradiol	**progesterone receptor**	**0,86463**				**P**
Isotretinoin	**retinoid X receptor gamma**	**0,86261**	**Y**			
Chenodiol	progesterone receptor	0,86051				
Tazarotene	estrogen receptor 1	0,85617				
Cholesterol	**progesterone receptor**	**0,85454**				**P**
Eplerenone	**estrogen receptor 1**	**0,85175**	**Y**	**Y**		
Isotretinoin	retinoid X receptor alpha	0,846				
Etretinate	estrogen receptor 1	0,84411				
Chenodiol	estrogen receptor 1	0,8422				

collection. Finally, the optimal value for the parameter n is found by minimizing the cost function

$$E(n) = 4\sum_i f_1(x_i) f_0(x_i) + 2n \tag{9}$$

where the first term penalizes large overlaps between C_0 and C_1, and the second term penalizes large n for better generalization. Mathematical foundations and a more detailed explanation of numerical algorithm and cost function can be found in the paper of Karacali (2010).[28]

We applied the QSL strategy to predict potential interactions between the compound-protein pairs that do not have any known interaction. To this end, we constructed two datasets: The unknown dataset C_0 which includes the untested compound-protein pairs that do not have a documented interaction, and the true positive dataset C_1 which includes the compound-protein pairs whose interactions are experimentally validated. The samples in C_0 are assigned with a label of 0 $(y = 0)$ and the samples in C_1 are assigned with a label of 1 $(y = 1)$. Then, the QSL algorithm calculates the posterior probability of the true positive dataset $P(C_1|x)$ at each compound-protein pair $x = (c, p)$ using the similarity matrix, $\boldsymbol{K}$ of all compound-protein pairs.

We adapted the numerical algorithm developed by Karacali (2010) to CPI prediction task in such a way that similarities between pairs are used instead of distances between feature vectors, where the most similar pair to a query pair corresponds to its nearest neighbor. This allows formulating the nearest neighbor classification and by extension the quasi-supervised learning algorithm in terms of a similarity measure between pairs, which eliminates the need to construct feature vectors for the unstructured chemical and genomics data on which a distance metric is otherwise to be defined and calculated for compound-protein pairs. The labels of compound-protein pairs in the unknown dataset C_0 are then predicted based on a set of known interactions between a relatively small number of established compound-protein pairs in C_1 in terms of pairwise similarities between compound-protein pairs. By virtue of the QSL paradigm, only positive interaction information and well-defined similarity measures between chemical and protein data are enough to carry out our proposed method. Nevertheless, all compound-protein pairs can still be embedded into a Euclidean space via

Table 4. The list of top 30 predicted positive interactions in the unknown dataset C_0 for the GPCR Dataset The identified positive interactions are indicated by the letter Y and potential interactions are indicated by the letter P, respectively.

Compound	Protein	Posterior Probability	SuperTarget	DrugBank	KEGG	ChEMBL
Isoetharine	**adrenoceptor beta 2**	**0,96806**		**Y**	**Y**	
Albuterol	**adrenoceptor beta 1**	**0,96794**	**Y**	**Y**		
Clozapine	**dopamine receptor D3**	**0,96584**	**Y**	**Y**		**Y**
Metoprolol	**adrenoceptor beta 2**	**0,9627**	**Y**	**Y**	**Y**	
Denopamine	**adrenoceptor beta 2**	**0,95763**				**P**
Levodopa	adrenoceptor beta 2	0,952				
Ritodrine	**adrenoceptor beta 1**	**0,95129**		**Y**		**P**
Dipivefrin	**adrenoceptor beta 2**	**0,94935**		**Y**	**Y**	
Epinephrine	**adrenoceptor beta 3**	**0,94265**			**Y**	**Y**
Methoxamine hydrochloride	adrenoceptor beta 2	0,94259				
Albuterol sulfate	**adrenoceptor beta 1**	**0,9424**	**Y**	**Y**		**P**
Levodopa	adrenoceptor beta 1	0,94178				
Methoxamine hydrochloride	adrenoceptor beta 1	0,94074				
Dipivefrin	**adrenoceptor beta 1**	**0,93835**			**Y**	
Bisoprolol	**adrenoceptor beta 3**	**0,93605**	**Y**			
Atenolol	**adrenoceptor beta 3**	**0,93445**	**Y**			
Cicloprolol hydrochloride	**adrenoceptor beta 3**	**0,93416**			**Y**	
Betaxolol hydrochloride	adrenoceptor beta 3	0,93132				
Clozapine	**adrenoceptor alpha 2C**	**0,93129**		**Y**		**Y**
Chlorpromazine	**histamine receptor H1**	**0,93035**		**Y**	**Y**	**Y**
Fenoldopam mesylate	adrenoceptor beta 2	0,92823				
Terbutaline sulfate	**adrenoceptor beta 1**	**0,9279**	**Y**	**Y**		**P**
Methixene hydrochloride	histamine receptor H1	0,92739				
Clozapine	**cholinergic receptor muscarinic 3**	**0,9265**		**Y**		
Perphenazine	**histamine receptor H1**	**0,92473**				**P**
Chlorpromazine phenolphthalinate	**histamine receptor H1**	**0,92187**		**Y**	**Y**	
Chlorpromazine hibenzate	**dopamine receptor D2**	**0,92114**		**Y**	**Y**	
Oxymetazoline hydrochloride	adrenoceptor beta 2	0,9203				
Albuterol	**adrenoceptor beta 3**	**0,91992**	**Y**			
Mesoridazine	histamine receptor H1	0,91985				

dedicated feature vectors, and distances between vectors can be used instead: Sorting distances between a query pair and all pairs in an ascending order will be equivalent to sorting similarities in a descending order.

Another problem with compound-protein interaction data is class imbalance: Since only a small portion of samples are marked as true positive, the number of samples in C_0 is much greater than C_1. Therefore, we modified the cost function in the QSL algorithm to find the optimum n parameter as

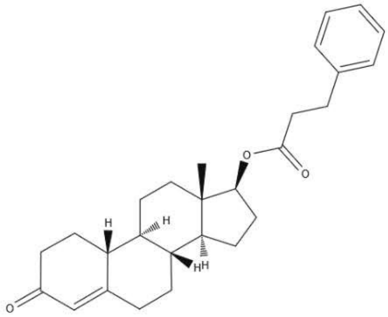
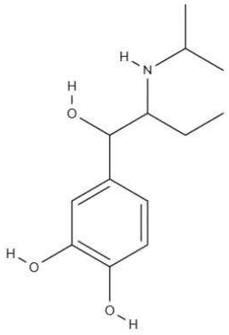
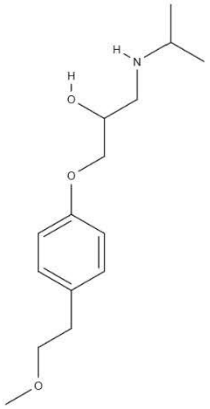
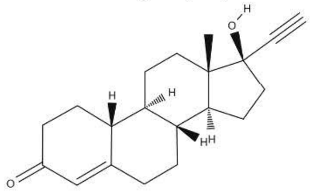
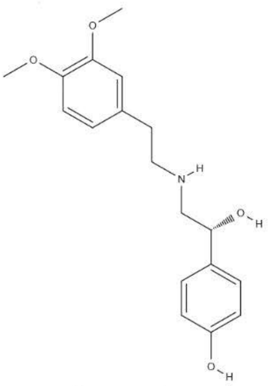
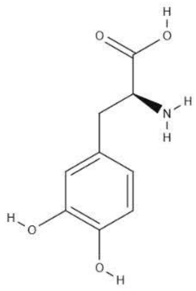
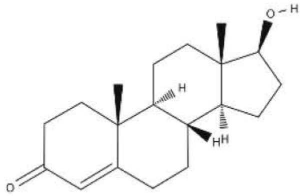
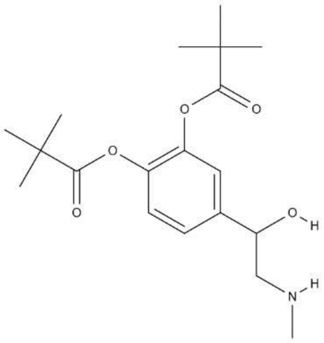
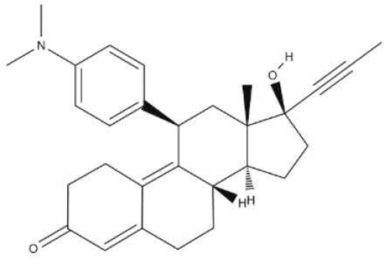
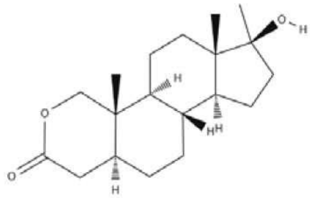
$$E(n) = 4\frac{|C_1|}{|C_0|}\sum_{x_i \in C_0} f_1(x_i)f_0(x_i) + 4\sum_{x_i \in C_1} f_1(x_i)f_0(x_i) + 2n \quad (10)$$

where $|C_0|$ and $|C_1|$ denotes the number of compound-protein pairs in C_0 and C_1, respectively. Note that this correction simply balances the contributions of the two datasets to the overall cost function as opposed to multiply-sampling the smaller dataset as proposed by the SMOTE approach.[14]

2.6 Kolmogorov-Smirnov Method

Once estimates of the posterior probability of the true positive dataset $P(C_1|x_i)$ for samples x_i in C_0 and C_1 are obtained, the samples in C_0 that would have been labeled as positives had they been tested are expected to exhibit greater posterior probability of belonging to C_1 compared to the actual negatives in C_0. Such hidden positive samples in C_0 can then be identified as $x_i \in C_0$ for which $P(C_1|x_i) \geq T$ using a suitable threshold T. Note that in this formulation, the threshold T draws the boundary of the overlap between the true positive dataset C_1 and the unknown dataset C_0. We used the Kolmogorov-Smirnov method to determine the optimal posterior probability threshold, T. To this end, observed posterior probability values $\{P(C_1|x_1), P(C_1|x_2), P(C_1|x_3), \ldots, P(C_1|x_{|C_1|+|C_0|})\}$ of samples from C_1 and C_0 are combined in a list and sorted in an ascending order. Empirical distribution functions of the true positive dataset $F_{C_1}(t)$ and the unknown dataset $F_{C_0}(t)$ are calculated separately as

Table 5. Structure graphs of top five compounds that have been predicted to interact with estrogen receptor 1 and adrenoceptor beta 2. The drawings have been generated using MolView (https://molview.org/).

estrogen receptor 1	adrenoceptor beta 2	
 Nandrolone phenpropionate	 Isoetharine	 Metoprolol
 Norethindrone	 Denopamine	 Levodopa
 Testosterone	 Dipivefrin	
 Mifepristone		
 Oxandrolone		

$$F_{C_1}(t) = \frac{1}{|C_1|}\sum_i 1(P(C_1|x_i) < t \text{ for } x_i \in C_1) \tag{11}$$

and

$$F_{C_0}(t) = \frac{1}{|C_0|}\sum_i 1(P(C_1|x_i) < t \text{ for } x_i \in C_0) \tag{12}$$

for all $t \in [0,1]$. Finally, the maximum difference between the empirical cumulative distribution functions of the two sample sets was identified as

$$D_{max} = \max_t |F_{C_1}(t) - F_{C_0}(t)| \tag{13}$$

by a line search.

Conventionally, the D_{max} statistic is used to decide whether samples in two different sets come from the same distribution or not with respect to a statistical significance level.[46] In this study, the posterior probability value at which D_{max} is observed is used as the optimal threshold that separates the hidden positive samples in C_0 from the rest. We also used the value of D_{max} as a measure pertaining to the ability of the proposed approach and the associated similarity metrics to separate the hidden positive and the actual negative samples in C_0 for performance comparison purposes between different similarity measures.

3 Results

In this section, we first provide an analysis of the proposed methodology regarding its ability to separate the hidden positives from the actual negatives in the unknown dataset using different combinations of compound and protein similarity measures. Then, we present the most likely interactions that are predicted by the proposed method and the current records about these interactions in up-to-date compound-protein interaction databases.

We calculated a total of thirteen compound similarity matrix alternatives and one protein similarity matrix for the compounds and proteins in the Nuclear Receptor dataset and the GPCR dataset separately using the methods described earlier. By applying the quasi-supervised learning algorithm on the resulting thirteen combined compound-protein similarity matrices, we calculated the posterior probability of the true positive dataset for all compound-protein pairs. The quality of the separation between the hidden positives and the actual negatives in the unknown dataset for the thirteen different similarity matrix choices was calculated in terms of the D_{max} values obtained by Kolmogorov-Smirnov analysis. These values indicate the ability of the proposed framework to contrast the true positive dataset against the unknown dataset, and by extension, how good the interacting and non-interacting compound-protein pair classes are distinguished from each other.

Table 2 presents the D_{max} values of all techniques with which we calculated the similarity between compounds, as there is only one similarity measure for proteins, for the Nuclear Receptor dataset and the GPCR dataset. The results obtained from the Nuclear Receptor dataset indicate that KCF–S Fingerprints achieves the greatest separation between the compound-protein pairs in the true positive dataset C_1 and the unknown dataset C_0, followed by Maximum Common Substructure (RDkit) and TF-IDF cosine similarity. The top three techniques that achieved the greatest separation in the GPCR dataset are TF-IDF cosine similarity, LINGO similarity with $q = 3$, and LINGO similarity with $q = 4$, respectively. The resulting separation between the hidden positive compound-protein pairs and the actual negatives in the unknown dataset is also apparent in the histograms of the posterior probability of the true positive dataset obtained using these compound similarity methods as shown in Figure 2 and Figure 3. The compound-protein pairs for which the posterior probability of the true positive dataset was greater than the indicated threshold were identified as hidden positives representing predicted interactions.

Finally, we constructed two lists of predicted interactions, one for the Nuclear Receptor dataset by taking the intersection of the compound-protein pairs predicted separately by KCF–S, SIMCOMP, MCS, LINGOsim ($q = 3$), LINGOsim ($q = 4$), LINGOsim ($q = 5$), LINGO based TF and LINGO based TF-IDF, and another for the GPCR dataset by taking the intersection of the compound-protein pairs predicted separately by ECFP4, FCFP4, KCF–S, MCS, LINGOsim ($q = 3$), LINGOsim ($q = 4$), LINGOsim ($q = 5$), TF LINGO based TF and LINGO based TF-IDF for which Kolmogorov-Smirnov Analysis resulted in D_{max} values greater than 0.6. For each predicted interaction, we calculated the geometric mean of the posterior probabilities by each method to obtain a unique posterior probability.

The top thirty predicted interactions in both lists are provided in Table 3 and Table 4, respectively, in the descending order of posterior probability of belonging to the set of true interactions along with the current record on them in DrugBank,[29] KEGG,[30] SuperTarget,[32] ChEMBL.[47] Note that since the publication of the Yamanashi dataset in 2008, interactions of some unlabeled pairs in C_0 have been experimentally validated and incorporated in the interaction databases listed above, providing a means for independent evaluation for the predicted interactions. In Table 3 and Table 4, the pairs that have interaction record in least one dataset and the potential interactions suggested by ChEMBL[47] were indicated by bold characters. In the tables, identified positive interactions are indicated by the letter Y and potential interactions are indicated by the letter P, respectively. Note also that a considerable number of predicted interactions are now categorized as positive interactions indicating the success of the proposed approach in identifying unknown true interactions among all possible compound-protein combinations.

Lastly, we collected the structures of top five compounds that have been predicted to interact with estrogen receptor 1 and adrenoceptor beta 2 in Tables 3 and 4, respectively. The 2D structure graphs shown in Table 5 reveal striking similarities: The common feature of the top five compounds that have been predicted to interact with estrogen receptor 1 is the similarity of the chemical structures with each other especially with testosterone. Among the top five compounds that have been predicted to interact with adrenoceptor beta 2, Isoetharine, Denopamine, Levodopa and Dipivefrin all contain a catechol group while Metoprolol and Denopamine have a phenol group as common substructures. Based on these observations, it is not surprising that these compounds have been predicted to interact with their respective target proteins.

4 Discussion

In this paper, we have proposed a quasi-supervised learning approach for compound-protein interaction prediction that addresses the issues associated with the lack of ground-truth negative instances in compound-protein interaction datasets. As mentioned in the literature review, there are very few studies in the literature that address the absence of reliable negatives as well as data imbalance between true positives and unlabeled compound-protein pairs. The present study offers an alternative strategy for an adequate evaluation of unlabeled compound-protein pairs. The results show that the quasi-supervised learning algorithm can make accurate predictions on interaction status of unlabeled compound-protein pairs without requiring an experimentally validated set of true negatives; or compound-protein pairs that have been established not to interact.

The quasi-supervised learning algorithm is well-suited to the compound-protein interaction prediction problem due to two reasons. Firstly, it uses only ground-truth positive compound-protein pairs without making any unrealistic and potentially erroneous presumptions on the interaction status of the unlabeled pairs. Instead, it successfully contrasts the set of all unlabeled compound-protein pairs with no known interaction with the true-positive dataset and identifies the pairs most likely to interact with each other automatically. Secondly, it can operate on the similarity structure between protein and compound pairs directly without requiring a feature vector representation for either of them, a common requirement for most other machine learning strategies. In this manner, it avoids the issues and shortcomings associated with feature-extraction processes that constitute major challenges especially for unstructured compound and protein data. This also allows incorporating alternative notions of similarity between protein and compound pairs from a larger, non-numeric class of similarity measures and enhances the breadth of the analysis.

On a final note, the proposed methodology can be extended in several ways. First, we applied quasi-supervised learning algorithm on only Nuclear Receptor and GPCR datasets due to computational limitations. These limitations are associated with the quadratic complexity of the kernel approach that entails calculating a similarity matrix for all compound-protein pairs. This, in turn, imposes restrictions on the dataset size to be evaluated unless the computational cost is alleviated using resource-friendly techniques. The proposed methodology can also be applied on datasets of other common target protein families such as Enzyme and Ion Channels of the Yamanashi dataset using more powerful computing resources. The Quasi-Supervised Learning algorithm appears particularly suitable for parallelization allowing for wider-scale applications on parallel computation architectures. Apart from this, further research can explore additional similarity measures that reflect the correlation between chemical and genomic spaces for potentially more efficient prediction. For instance, LINGO-like similarity measures for proteins can be explored in terms of protein motifs and domains that may incorporate the established functional characteristics of the proteins into the similarity structure more adequately.

Data Availability Statement

These data were derived from the following resources available in the public domain:

References

[1] T. 1. Engel, J. Gasteiger, *Appl. Chemoinf.: Achievements and Future Opportunities*, Germany: Wiley-VCH, **2018**.

[2] J. Medina-Franco, M. Giulianotti, G. Welmaker, R. Houghten, *Drug Discovery Today* **2013**, *18*, 495–501.

[3] E. Lounkine, M. J. Keiser, S. Whitebread, D. Mikhailov, J. Hamon, J. L. Jenkins, P. Lavan, E. Weber, A. K. Doak, S. Côté, B. K. Shoichet, L. Urban, *Nature* **2012**, *486*, 486, 361–367.

[4] N. Novac, *Trends Pharmacol. Sci.* **2013**, *34*, 267–272.

[5] E. W. Sayers, R. Agarwala, E. E. Bolton, J. R. Brister, K. Canese, K. Clark, R. Connor, N. Fiorini, K. Funk, T. Hefferon, J. B. Holmes, S. Kim, A. Kimchi, P. A. Kitts, S. Lathrop, Z. Lu, T. L. Madden, A. Marchler-Bauer, L. Phan, V. A. Schneider, C. L. Schoch, K. D. Pruitt, J. Ostell, *Nucleic Acids Res.* **2019**, *47*, D23-D28.

[6] A. L. Hopkins, *Nature* **2009**, *462*, 167–68.

[7] A. C. Cheng, R. G. Coleman, K. T. Smyth, Q. Cao, P. Soulard, D. R. Caffrey, A. C. Salzberg, E. S. Huang, *Nat. Biotechnol.* **2007**, *25*, 71–75.

[8] J. B. Brown, *Computational Chemogenomics.* Humana Press., **2018**.

[9] M. J. Keiser, B. L. Roth, B. N. Armbruster, P. Ernsberger, J. J. Irwin, B. K. Shoichet, *Nat. Biotechnol.* **2007**, 25, 197–206.

[10] K. Heikamp, J. Bajorath, *Chem. Biol. Drug Des.* **2012**, *81*(1), 33–40.

[11] H. Ding, I. Takigawa, H. Mamitsuka, S. Zhu, *Briefings Bioinf.* **2013**, *15*, 734–747.

[12] T. Cheng, M. Hao, T. Takeda, S. H. Bryant, Y. Wang, *AAPS J.* **2017**, *19*, 1264–1275.

[13] A. Ezzat, M. Wu, X. L. Li, C. K. Kwoh, *Briefings Bioinf.* **2018**, *20*, 1337–1357.

[14] S. Redkar, S. Mondal, A. Joseph, K. Hareesha, *Mol. Inf.* **2020**, *39*(5), 1900062.

[15] Y. Chen, E. K. Garcia, M. R. Gupta, A. Rahimi, L. Cazzanti, *J. Mach. Learn. Res.* **2009**, *10*, 747–776.

[16] Y. Yamanishi, M. Araki, A. Gutteridge, W. Honda, M. Kanehisa, *Bioinformatics* **2008**, *24*, i232-i240.

[17] L. Jacob, J. P. Vert, *Bioinformatics* **2008**, *24*, 2149–2156.

[18] T. van Laarhoven, S. B. Nabuurs, E. Marchiori, *Bioinformatics* **2011**, *27*, 3036–3043.

[19] T. van Laarhoven, E. Marchiori, *PLoS One* **2013**, *8*, e66952.

[20] M. Gönen, *Bioinformatics* **2012**, *28*, 2304–2310.

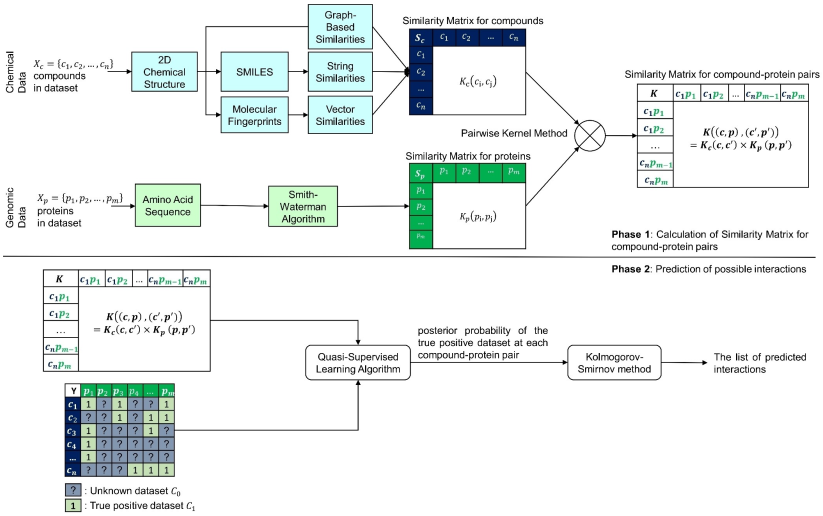
[21] X. Zheng, H. Ding, H. Mamitsuka, S. Zhu, in *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, New York, NY, USA, **2013**.

[22] D. Reker, P. Schneider, G. Schneider, J. Brown, *Future Med. Chem.* **2017**, *9*(4), 381–402.

[23] M. Tsubaki, K. Tomii, J. Sese, *Bioinformatics* **2018**, *35*(2), 309–318.
[24] H. Liu, J. Sun, J. Guan, J. Zheng, S. Zhou, *Bioinformatics* **2015**, *31* (12), i221-i229.
[25] M. Mysinger, M. Carchia, J. Irwin, B. Shoichet, *J. Med. Chem.* **2012**, *55*(14), 6582–6594.
[26] A. Mayr, G. Klambauer, T. Unterthiner, S. Hochreiter, *Front. Environ. Sci., 3*:80.
[27] R. Huang, M. Xia, D. T. Nguyen, T. Zhao, S. Sakamuru, J. Zhao, S. Shahane, A. Rossoshek, A. Simeonov, *Front. Environ. Sci.* **2016**, *3*:85.
[28] B. Karaçali, *Pattern Recognit. Lett.* **2010**, *43*, 3674–3682.
[29] D. S. Wishart, C. Knox, A. C. Guo, D. Cheng, S. Shrivastava, D. Tzur, B. Gautam, M. Hassanali, *Nucleic Acids Res.* **2008**, *36*, D901-D906.
[30] M. Kanehisa, *Nucleic Acids Res.* **2006**, 34, D354-D357.
[31] I. Schomburg, *Nucleic Acids Res.* **2004**, 32, 431D-433.
[32] S. Gunther, M. Kuhn, M. Dunkel, M. Campillos, C. Senger, E. Petsalaki, J. Ahmed, E. G. Urdiales, A. Gewiess, L. J. Jensen, R. Schneider, R. Skoblo, R. B. Russell, P. E. Bourne, P. Bork, R. Preissner, *Nucleic Acids Res.* **2007**, *36*, D919-D922.
[33] M. Hattori, Y. Okuno, S. Goto, M. Kanehisa, *J. Am. Chem. Soc.* **2003**, *125*, 11853–11865.
[34] "RDKit: Open-source cheminformatics", can be found under http://www.rdkit.org, **2021**.
[35] D. Weininger, A. Weininger, J. L. Weininger, *J. Chem. Inf. Model.* **1989**, *29*, DOI 97–101.
[36] H. Öztürk, E. Ozkirimli, A. Özgür, *BMC Bioinf.* **2016**, *17*, DOI 10.1186/s12859-016-0977-x.
[37] "Daylight Theory: SMILES", can be found under https://www.daylight.com/dayhtml/doc/theory/theory.smiles.html, **2021**.
[38] D. Vidal, M. Thormann, M. Pons, *J. Chem. Inf. Model.* **2005**, *45*, 386–393.
[39] R. Sawada, M. Kotera, Y. Yamanishi, *Mol. Inf.* **2014**, *33*, 719–731.
[40] D. Rogers, M. Hahn, *J. Chem. Inf. Model.* **2010**, *50*, 742–754.
[41] J. L. Durant, B. A. Leland, D. R. Henry, J. G. Nourse, *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1273–1280.
[42] M. Kotera, Y. Tabei, Y. Yamanishi, Y. Moriya, T. Tokimatsu, M. Kanehisa, S. Goto, *BMC Syst. Biol.* **2013**, *7*, S2.
[43] M. Sato, H. Suetake, M. Kotera, *bioRxiv* **2018**, DOI 10.1101/452383.
[44] T. F. Smith, M. S. Waterman, *J. Mol. Biol.* **1981**, *147*, 195–197.
[45] J. Vert, J. Qiu, W. Noble, *BMC Bioinf.* **2007**, *8*(S10).
[46] B. Bagwell, *Clinical Immunology Newsletter* **1996**, *16*, 33–37.
[47] A. Gaulton, L. J. Bellis, A. P. Bento, J. Chambers, M. Davies, A. Hersey, Y. Light, S. McGlinchey, D. Michalovich, B. Al-Lazikani, J. P. Overington, *Nucleic Acids Res.* **2012**, *40*, D1100-D1107.

Received: October 17, 2021
Accepted: [redacted], [redacted]
Published online on [redacted], [redacted]

RESEARCH ARTICLE



Chemical Data
$X_c = \{c_1, c_2, \dots, c_n\}$ compounds in dataset
2D Chemical Structure
SMILES
Molecular Fingerprints
Graph-Based Similarities
String Similarities
Vector Similarities
Similarity Matrix for compounds
S_c c_1 c_2 … c_n
$K_c(c_i, c_j)$
Pairwise Kernel Method
Genomic Data
$X_p = \{p_1, p_2, \dots, p_m\}$ proteins in dataset
Amino Acid Sequence
Smith-Waterman Algorithm
Similarity Matrix for proteins
S_p p_1 p_2 … p_m
$K_p(p_i, p_j)$
Similarity Matrix for compound-protein pairs
K c_1p_1 c_1p_2 … c_np_{m-1} c_np_m
$K((c,p),(c',p')) = K_c(c,c') \times K_p(p,p')$
Phase 1: Calculation of Similarity Matrix for compound-protein pairs
Phase 2: Prediction of possible interactions
Quasi-Supervised Learning Algorithm
posterior probability of the true positive dataset at each compound-protein pair
Kolmogorov-Smirnov method
The list of predicted interactions
Y p_1 p_2 p_3 p_4 … p_m
? : Unknown dataset C_0
1 : True positive dataset C_1

*O. Çakı, B. Karaçalı**

1 – 14

Quasi-Supervised Strategies for Compound-Protein Interaction Prediction