Bioinformatics, 37(16), 2021, 2275–2281
doi: 10.1093/bioinformatics/btab092
Advance Access Publication Date: 9 February 2021
Original Paper

OXFORD

Sequence analysis

A novel method for data fusion over entity-relation graphs and its application to protein–protein interaction prediction

Daniele Raimondi *, Jaak Simm, Adam Arany and Yves Moreau

ESAT-STADIUS, KU Leuven, 3001 Leuven, Belgium

*To whom correspondence should be addressed.

Associate Editor: Jinbo Xu

Received on October 26, 2020; revised on January 14, 2021; editorial decision on February 2, 2021; accepted on February 4, 2021

Abstract

Motivation: Modern bioinformatics is facing increasingly complex problems to solve, and we are indeed rapidly approaching an era in which the ability to seamlessly integrate heterogeneous sources of information will be crucial for the scientific progress. Here, we present a novel non-linear data fusion framework that generalizes the conventional matrix factorization paradigm allowing inference over arbitrary entity-relation graphs, and we applied it to the prediction of protein–protein interactions (PPIs). Improving our knowledge of PPI networks at the proteome scale is indeed crucial to understand protein function, physiological and disease states and cell life in general.

Results: We devised three data fusion-based models for the proteome-level prediction of PPIs, and we show that our method outperforms state of the art approaches on common benchmarks. Moreover, we investigate its predictions on newly published PPIs, showing that this new data has a clear shift in its underlying distributions and we thus train and test our models on this extended dataset.

Contact: daniele.raimondi@kuleuven.be

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Although significant progress has been made in the development of data fusion methods using statistics, kernel methods and matrix factorization (MF), a systematic framework able to perform inference over arbitrarily complex data structures is still missing. MF methods allow the factorization of sparsely observed matrices (Mnih and Salakhutdinov, 2008; Srebro and Jaakkola, 2003) by approximating them as a product of two or more low-rank matrices and thus provide an attractive methodology for data fusion (Žitnik and Zupan, 2014a), and they have indeed been consistently used in several domains. In bioinformatics, MF has been recently used for gene prioritization (Žitnik *et al.*, 2015; Žitnik and Zupan, 2014a), drug–protein interaction (Arany *et al.*, 2015) and *in silico* diagnosis of Crohn's disease (Jeong and Kim, 2016). Recent efforts toward the generalization of MF methodology involved (i) adding side information, which corresponds to features in the classic ML sense, (ii) performing full scalable Bayesian inference (Simm *et al.*, 2015) and (iii) extending the method to the factorization of tensors (Arany *et al.*, 2015).

Here, we aim at pushing the generalization of MF-based data fusion frameworks even further, by rephrasing the problem in terms of inference over entity-relation (ER) graphs, where entities are classes of objects belonging to a particular domain and relations describe the interactions between entities. Such an arbitrary data fusion model is completely general and allows inference on any domain whose data structure can be represented as ER graph, which could be conceptually likened to a relational database. One of the benefits introduced by this novel formalism is the ease in which entities can be connected through relations, thus allowing the seamless inclusion in the model of datasets that are apparently only loosely related with the *main task* under investigation, thus increasing the data available to the model (Žitnik and Zupan, 2014a) by integrating data that is usually discarded but that could provide further contextual information. The first approach in this sense is DFMF (Data Fusion by Matrix tri-Factorization) (Žitnik and Zupan, 2014a), which provides multi-entity and multi-relational data fusion through matrix trifactorization. The model infers bilinear latent representations of both the entities and the relations by minimizing the error in reconstructing the input matrices describing relations, subject to some constraints. DFMF has been successfully used for gene prioritization (Žitnik *et al.*, 2015) and gene function prediction (Žitnik and Zupan, 2014b). Nonetheless, DFMF is still not fully general since (i) it infers only bi-linear relations from the data, (ii) it does not support the use of side information, which is usually helpful for the contextualization of entities with few observed values and (iii) it does not model self-relations (relations between an entity and itself) since it uses such data to define similarity constraints between the

© The Author(s) 2021. Published by Oxford University Press. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com

Downloaded from https://academic.oup.com/bioinformatics/article/37/16/2275/6131674 by KU Leuven Libraries user on 29 October 2024

latent representation of instances belonging to the same class. In biological applications, it is relatively easy to imagine self-relations that do not describe a similarity measure between the objects involved, such as the case of protein–protein interaction (PPI), in which very different proteins may interact. Some of these drawbacks, such as the inherent (bi)linearity of the MF algorithms have been overcome by other approaches (Fan and Cheng, 2018; Trigeorgis *et al.*, 2016), but only in the context of fixed single-matrix ER models with two entities and one relation and never in the context of arbitrary ER graphs. A fully general nonlinear approach to data fusion is thus, to the best of our knowledge, currently missing, even though it could be crucial to learn complex patterns from large scale intertwined data.

In this article, we present NXTfusion, a novel data fusion framework for non-linear inference over arbitrary ER graphs, and we apply it to the prediction of PPI networks in humans. Proteins are *social entities* in the sense that, in order to perform their molecular function they interact with other proteins (PPI) and other molecules (Goh *et al.*, 2007; Meyer *et al.*, 2013; Rolland *et al.*, 2014), giving rise to the notion of *interactome*. Investigating how proteins dynamically interact is indeed crucial to understand the biological mechanisms driving cell life (Chen *et al.*, 2019; Luck *et al.*, 2020), such as signaling pathways, protein function, diseases and drug mechanisms (Kotlyar *et al.*, 2015; Snider *et al.*, 2015).

Although it is difficult to estimate the human interactome size (Kotlyar *et al.*, 2015; Luck *et al.*, 2020), huge experimental efforts have been made (Luck *et al.*, 2020; Rolland *et al.*, 2014; Schwartz *et al.*, 2009), leading to a steady increase of the known PPIs (Kotlyar *et al.*, 2015). Nevertheless, current experimental approaches are expensive, time consuming and have limited sensitivity (Chen *et al.*, 2019; Luck *et al.*, 2020). To partially overcome these problems and help speeding up the discovery of PPIs, computational methods have been devised to prioritize the candidates for experimental screening (Chen *et al.*, 2019; Kotlyar *et al.*, 2015).

Some computational approaches use as input just the protein sequences (Chen *et al.*, 2019; Sprinzak and Margalit, 2001), other rely on the detection of the co-evolutionary signal (Bitbol *et al.*, 2016; Cong *et al.*, 2019; Zeng *et al.*, 2018) or on structural information (Zhang *et al.*, 2012). These methods nevertheless generally are devised to work on subset of the human proteome/interactome, due to (i) the limited availability of protein structures, (ii) the limited availability of homologous sequences (Orlando *et al.*, 2016) or (iii) the excessive presence of missing information, leading to completely un-annotated proteins. For these reasons, here we focus on a specific class of large scale data mining-based prediction approaches, such as FPClass (Kotlyar *et al.*, 2015), which integrate heterogeneous sources of information and aim at modeling proteome-scale interactome networks involving millions of interactions (or lack thereof) among 17651 human proteins.

In this study, we used NXTfusion to devise three increasingly complex ER graph-based data fusion models corresponding to (i) the classical MF paradigm, (ii) tensor factorization and (iii) arbitrary multi-relation ER graph. We show that our approach outperforms state of the art data mining methods such as FPClass (Kotlyar *et al.*, 2015), providing a 12% improvement in terms of area under the precision-recall curve (AUPRC). Moreover, we investigate its predictions on recently discovered PPIs, showing that this new data has a clear shift in the distributions underlying these interactions and that it could thus be beneficial to include it in the training set in order to improve the ability of PPI predictors to generalize over the entire interactome. We thus trained and tested our models also on this extended dataset.

2 Materials and methods

2.1 Benchmark dataset

We benchmarked the performances of our data fusion approaches against state of the art similar methods by adopting the main PPI dataset used in FPClass (Kotlyar *et al.*, 2015). This dataset contains 20454 interactions observed in at least two experiments and 2 045 400 randomly selected negatives over a total of 17 651 human proteins. The random selection of the negative samples is certainly not ideal but is necessary to overcome the fact that there is no objective way to identify large numbers of highly likely non-interacting pairs. This sampling approach exploits the expected sparsity of the actual proteome-scale PPI network, ensuring that the vast majority of the randomly sampled protein pairs will indeed correspond to non-interacting pairs. We also used the same train-test splits to reproduce the 5-folds cross-validation (CV) used to evaluate FPClass performance, in order to ensure a fair comparison.

Moreover, from Braun *et al.* (2009), we extracted the 92 high-confidence interacting pairs of proteins and the 92 random pairs (negatives) used to validate FPClass, providing an extra validation.

Finally, from the novel PPIs listed in the recently published human reference interactome (HI-III-19) dataset (Luck *et al.*, 2020), we built a blind test set containing 43 663 interactions between 8125 proteins that were not known when the FPClass training was built. We refer to this dataset as HURI.

2.2 Additional data sources

Our data fusion models aim at integrating heterogeneous sources of information represented as (possibly) sparsely observed matrices which conceptually represent relations between pairs of entities. In this study we use the following data sources.

We retrieved Pfam domains annotations for the 17 651 proteins in the benchmark dataset with PfamScan (Finn *et al.*, 2014). We used such annotations to represent the (Protein, Domain) relation as a sparsely observed binary matrix of size $17\,651 \times 5972$ where a 1 in position (i, j) indicates that the protein i contains the domain j.

We retrieved associations between proteins and diseases from Disgenet (Piñero *et al.*, 2017) using only the 'curated' annotations. This data consists in a sparsely observed real-valued matrix of size $17\,651 \times 6087$ and it has been used to represent the (Protein, Disease) relation.

We retrieved expression data for each protein in 53 tissues from Liu *et al.* (2016). This data is stored into a dense $17\,651 \times 53$ real-valued matrix and it has been used to represent the (Protein, Tissue) relation.

The network topology information matrices are dense $17\,651 \times 17\,651$ real-valued matrices that have been computed as described in Kotlyar *et al.* (2015) and Scott and Barton (2007). The goal of the two topology scores that we compute is to quantify, for each pair of proteins (i, j) in the training set, the propensity of i and j to share interaction partners (Kotlyar *et al.*, 2015). Further details about the construction of these matrices are available in Supplementary Material.

2.3 Nxtfusion: non-linear data fusion over entity-relation graphs

In the classical MF settings, a partially observed matrix $Y = UV$ is factorized into two low-dimensionality rectangular matrices U and V. In order to minimize the regularized reconstruction error

$$\underset{U,V}{\operatorname{argmin}} ||Y - UV||_F + \lambda(||U||_F + ||V||_F)$$

U and V end up containing the latent representations of the objects represented as rows and columns in Y. $||\cdot||_F$ is the Frobenius norm and λ the regularization factor. This supervised MF approach paradigm is commonly used to build recommender systems and perform data fusion (Arany *et al.*, 2015; Simm *et al.*, 2015; Žitnik *et al.*, 2015; Žitnik and Zupan, 2014a).

Here, we propose a full generalization of this approach, abstracting from the factorization details and focusing on the conceptual connections between the data on which inference must be performed. We thus organize the data as an arbitrarily connected ER graph $G = (\mathcal{E}, \mathcal{R})$ where the entities $E_i \in \mathcal{E}$ (the nodes of the graph) represent classes of objects and $R_{i,j} \in \mathcal{R}$ the relations (edges) between them (see Fig. 1 for an illustrative example). Each $R_{i,j}$ connects two entities $E_i, E_j \in \mathcal{E}$ and it is represented by a (possibly sparsely observed) matrix.

Downloaded from https://academic.oup.com/bioinformatics/article/37/16/2275/6131674 by KU Leuven Libraries user on 29 October 2024

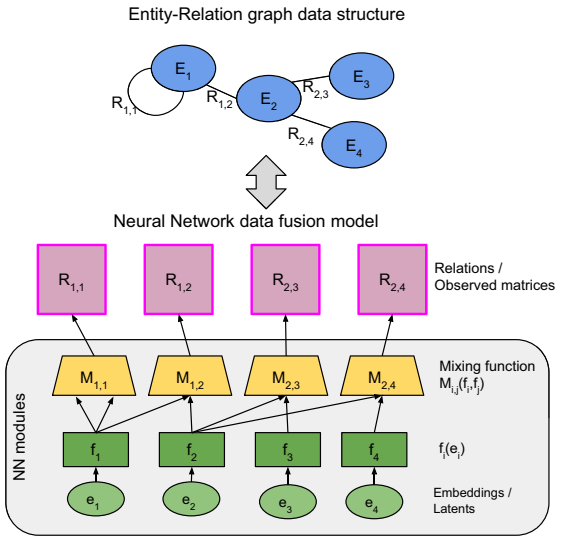


Fig. 1. A hypothetical entity-relation graph (top half) and the corresponding modular neural network architecture that we designed to perform inference over the graph. Entities E_i are classes of objects and relations $R_{i,j}$ correspond to the observed matrices, which contain the labels for the supervised training of the NN

Given an ER graph G (see Fig. 1), which describes how the available data is *conceptually* structured, in order to perform inference over G we construct a modular neural network (NN) model able to factorize G (bottom half of Fig. 1). We represent each entity E_i as a set of trainable latent variables e_i (called *embeddings* in the NN jargon) which participate in the reconstruction of each relation (matrix) $R_{i,j}$ in which E_i is involved. Without loss of generality, for each $R_{i,j} \in \mathcal{R}$ the loss function to be minimized is $L_{i,j}(R_{i,j}, M_{i,j}(f_i(e_i), f_j(e_j)))$, where e_i are the trainable embeddings representing the latent variables, each f_i is an entity-specific function (e.g. a NN module containing one or multiple feed-forward layers), and $M(\cdot,\cdot)$ is a relation-specific mixing function that takes as input a pair of entity-specific functions f_i and outputs the $\hat{R}_{i,j}$ prediction of the model (e.g. a bilinear layer). Each $L_{i,j}$ is a relation-specific loss function, that matches the type of data contained in the matrix representing the relation $R_{i,j}$, to ensure full flexibility.

As shown in Figure 1, this optimization is performed concurrently over all the relations (matrices) involved in the ER graph G (all the edges), effectively performing the factorization over the entire graph. The resulting objective function to optimize is thus:

$$\underset{W,e}{\operatorname{argmin}} \sum_{R_{i,j} \in \mathcal{R}} \omega_{i,j} L_{i,j}(R_{i,j}, M_{i,j}(f_i(e_i), f_j(e_j)))$$

where W are the trainable parameters of the functions $M_{i,j}$ and f_i, e_i are the embeddings, $L_{i,j}$ are the relation-specific loss functions, $\omega_{i,j}$ are loss-specific scale factors and $R_{i,j}$ the observed matrices describing the relations between pairs of entities.

This approach basically transforms the problem of factorizing ER graphs into an extreme multi-task learning problem, where the goal is to populate the ER graphs with additional tasks (relations between entities) that could have a positive synergistic effect over the learning of the main task that we want to solve. Moreover, existing approaches such as DFMF (Žitnik and Zupan, 2014a) or the classic MF are subsumed by our framework as special cases obtained by specific choices of the $M_{i,j}$ and f_i functions.

2.4 Three ER graphs data fusion models for protein–protein interaction prediction

To show the flexibility of our data fusion framework, we built three increasingly complex ER graphs-based models able to tackle the problem of PPI prediction, and we benchmarked them against state-of-the-art approaches. The first model (G1) is equivalent to a non-linear single MF, the second (G2) involves non-linear tensor factorization and the third (G3) performs non-linear inference over an ER graph with multiple entities and relations.

2.4.1 Model G1: non-linear MF

The simplest model, we propose (see Model G1 in Fig. 2) is the ER graph non-linear data fusion equivalent of the classic MF approach, where a single matrix Y is factorized into two matrices U and V. Model G1 in Figure 2 shows that the corresponding ER graph consists in a self-relation involving the single entity representing the 'protein' class. Y is the $17\,651 \times 17\,651$ sparsely observed binary target matrix that we want to reconstruct. It contains 20 454 interacting protein pairs, labeled as 1 s, and 2 045 400 randomly assigned non-interacting pairs, labeled as 0 s. Due to the heavy unbalancement in the data, we used the FocalLoss (Lin *et al.*, 2017) with $\gamma = 4$ as loss function, which has been specifically designed for unbalanced data.

2.4.2 Model G2: non-linear tensor factorization

The second model, we propose extends the classic MF paradigm by including multiple relations between the same entities, which are factorized concurrently (see Model G2 in Fig. 2). G2 extends G1 by introducing additional self-relations with the goal of providing the model with additional information with synergistic effects for the reconstruction of the main task Y. We followed the approach of FPClass (Kotlyar *et al.*, 2015) and (Scott and Barton, 2007) and we computed two matrices containing network topology data derived from Y. The goal of these additional relations is to quantify, for each pair of proteins (i, j), the propensity of i and j to share interaction partners. Details on how these scores are computed are available in Supplementary Material. These matrices have the same size ($17\,651 \times 17\,651$) of the main task Y but are dense and real-valued. Our framework allows relation-specific loss functions and so we used the FocalLoss for the Y matrix and the mean square error (MSE) for the topology matrices.

2.4.3 Model G3: non-linear inference over arbitrary ER graph

The last model presented here involves multiple relations between multiple entities, thus providing a further generalization of the tensor factorization (G2). Figure 2 shows the ER graph used for the G3 model, which adds to the G2 graph the Protein-Domain and Protein-Disease and Protein-Tissue relations, thus including annotations from Pfam (Finn *et al.*, 2014), Disgenet (Piñero *et al.*, 2017) and expression data (Liu *et al.*, 2016). We used MSE loss functions for the Protein-Domain and Protein-Disease relations, and SmoothL1Loss for the Protein-Tissue relation.

2.5 Implementation details

The general NN architecture used to factorize the ER graphs is shown in Figure 1 and it is implemented with Pytorch (Paszke *et al.*, 2017). During training, we used the Adam optimizer with a learning rate of 0.001 and an L2 regularization with $\lambda = 0.001$. The trainable embeddings representing the Protein, Domain, Disease and Tissue entities have respectively 70, 50, 50 and 10 dimensions. We chose these sizes to be roughly proportional to the cardinality of the entities they represent, without searching for the optimal sizes. The $f_i(\cdot)$ functions used here are feed forward (FF) layers with 50 output neurons and the mixing functions $M_{i,j}(\cdot,\cdot)$ are bilinear layers with (50, 50) input neurons and a 10 output neurons, followed by a FF layer with 10 input neurons that yields the final scalar prediction for each task. All the trainable parameters in the network (i.e. the embeddings, the f_i and $M_{i,j}$ modules) are randomly initialized using the xavier_uniform pytorch function before starting the training.

Since each relation may consist in a classification or regression problem, we used Tanh activations, always preceded by a normalization layer (LayerNorm). A dropout with $p = 0.1$ is applied to the final layer predicting the values associated to each relations. When dealing with symmetric matrices, we computed the reconstruction error only for the upper half. The losses associated to each relation

Downloaded from https://academic.oup.com/bioinformatics/article/37/16/2275/6131674 by KU Leuven Libraries user on 29 October 2024

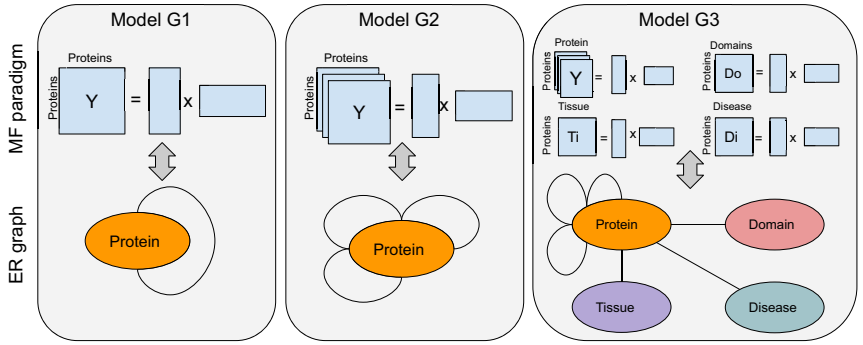


Fig. 2. The three increasingly complex models G1, G2 and G3 that we devised for protein–protein interaction prediction. For each model, we show the dualism between the conventional matrix factorization (MF) paradigm and the corresponding entity-relation (ER) graph that represents the same data fusion problem

have been weighted with suitable ω_{ij} in order to ensure that they all have comparable values. More details concerning the NN architecture are available in Supplementary Material. The code is available at https://bitbucket.org/eddiewrc/nxtppi/src. The NXTfusion data fusion library is available at https://bitbucket.org/eddiewrc/nxtfusion/ and https://pypi.org/project/NXTfusion/.

2.6 Evaluation procedure

In order to compare the prediction of our model with state of the art approaches, we reproduced the 5-folds CV used in FpClass (Kotlyar *et al.*, 2015), measuring the area under the ROC curve (AUC), but focusing on the AUPRC more, due to the heavy unbalancement between positive and negative cases.

3 Results

3.1 Protein–protein interaction prediction

We benchmarked our three ER graph-based data fusion models by evaluating their ability to predict interactions in a human interactome-scale PPI network. This network involves 20 454 experimentally determined interactions and 2 million randomly assigned negative samples (non-interacting pairs of proteins) among 17 651 human proteins. We compared the prediction performance of our G1, G2 and G3 models against the results obtained by FPClass (Kotlyar *et al.*, 2015), a data mining-based method which relies on the integration of heterogeneous sources of information via feature selection and a Noisy-OR model (Kotlyar *et al.*, 2015). To ensure a fair comparison, we reproduced the same CV procedure used by FPClass, with the same 5-folds split of the data.

In Table 1, we show the comparison between the results obtained by our G1, G2 and G3 models (see Fig. 2) and FPClass. We can see that AUC is very high for all the methods, due to the extreme unbalancement of the dataset, which contains two orders of magnitude more negative cases than positives. For this reason, the AUC does not show a sufficient dynamics to evaluate these methods, and we rely instead on the AUPRC, which provides an indication of how well the methods can *rank* the few positive cases higher than the negatives.

3.1.1 The single relation approach shows that integrating more data is beneficial

The model G1 corresponds to the non-linear version of the classical MF setting (see Fig. 2), where a single matrix is factorized into two latent matrices. In the ER formalism, this is equivalent to modeling a single relation between two entities, which in this case are coincident, making it become a self-relation, since the matrix is symmetric. In the model G1, no additional sources of information are added in the model, which focuses only on the PPI prediction. Although the performances are far better than random, the AUPRC is suboptimal, being nearly 50% lower than PFClass. This provides indications that more information need to be fused in order to obtain satisfactory PPIs predictions.

Table 1. The performance of the G1, G2 and G3 ER graph data fusion models compared with FPClass (Kotlyar *et al.*, 2015)

Methods	AUC	AUPRC
FPClass	96.0	70.2
G1	93.3	37.1
G2	98.9	78.5
G3	98.8	77.8

Note: The results have been obtained while reproducing the same 5-folds cross-validation procedure used for FPClass.

3.1.2 Integrating network topology information through tensor factorization shows a synergistic effect

In order to provide more information to the model, we devised the graph G2 (see Fig. 2). The G2 model is equivalent to a non-linear tensor factorization problem, or, in the ER formalism, a factorization over a graph containing multiple self-relations involving the Protein entity. G2 contains two additional relations (matrices) containing data computed from the topology of the PPI network used as training, following the approach proposed by Scott and Barton (2007) and also used by FPClass. Both papers show that this feature provides the highest AUPRC contribution among the features tested. We thus computed these topology scores in the same way (see Supplementary Material), and as expected adding them to our model provides a huge improvement, allowing G2 to reach an AUPRC 12% higher than FPClass. Since the model G2 integrates information by training its embeddings to solve multiple tasks at the same time, we can conclude that learning to predict the topology relations has a synergistic effect on the *main* PPI prediction task.

3.1.3 Inference over a multi-entity and multi-relation graph

The G3 model provides an example of the full extent to which our framework for data fusion over ER graphs can be used to integrate even loosely connected data sources into a single model. It contains six relations among four entities, introducing the Protein-Domain, Protein-Tissue and Protein-Disease relations, in order to provide a better contextualization of biological aspects that may provide valuable information when it comes to predict PPIs. We decided to introduce protein Domain and Expression data in G3 because previous approaches (Kotlyar *et al.*, 2015) showed a quantifiable positive contribution. The intuition behind adding data related to the involvement of proteins (genes) in known disease was that it could be an alternative way to contextualize how proteins interact at a high level of abstraction, such as the systemic disease level. Our G3 model still performs 11% better than FPClass in terms of AUPRC, but slightly lower than our G2 model, indicating that on this dataset the saturation has probably been reached with the topology

Downloaded from https://academic.oup.com/bioinformatics/article/37/16/2275/6131674 by KU Leuven Libraries user on 29 October 2024

information, which has indeed been shown to be among the most relevant features in Scott and Barton (2007) and FPClass (Kotlyar *et al.*, 2015).

3.2 Additional validations show consistent performances

To provide a further validation for the performances of our models, we ran two additional tests. First, we tested the G2 and G3 models on the dataset (Braun *et al.*, 2009), which contains 92 high-confidence interactions (positive cases) and 92 random protein pairs that are used as negatives (Kotlyar *et al.*, 2015). Supplementary Table S1 shows that the AUC and AUPRC obtained by the G2 and G3 methods are very similar to the ones obtained by FPClass method on the same data, showing a consistency in the prediction performance besides the CV. Interestingly, in this case the G3 model slightly outperforms the G2.

Second, we ran an additional CV procedure, ensuring that the proteins in different folds shared less than 20% sequence identity. This is more stringent with respect to the results presented in Table 1, where we reproduced the same CV performed by FPClass in order to ensure a fair comparison. The folds used in Table 1 have been built by FPClass authors by randomly partitioning the interacting and non-interacting protein pairs in five sets, but in this case, the same proteins might be present, albeit involved in different interactions, both in the training and test sets of each CV iteration. This behavior could produce an information leakage between the train and test sets, possibly leading to over-optimistic predictions.

Supplementary Table S2 shows the cross-validated performance of our models once this leakage is removed by ensuring that all the proteins in the folds used as test (and their homologs) were previously unseen by the model during the training. The G2 and G3 models show a negligible drop in terms of AUC and a 9% drop in terms of AUPRC, indicating that indeed the random selection of the interactions in the CV folds was leading to certain degree of overestimation of the performance. On the other hand the G1 model, which corresponds to the non-linear version of the classical MF approach, shows a drastic drop of the performance, with the AUPRC that goes to 0. This is indeed expected, because since the proteins in the test set are unseen by the model and since the G1 model consists of only a single relation (the target matrix), the latent representations (the embeddings) corresponding to the proteins in the test set did not undergo any optimization during training, and thus the predictions are deemed to be random (the Matthews Correlation Coefficient is equal to 0.09). Incidentally, this stringent CV shows how the addition of *auxiliary* relations in models G2 and G3 allows our data fusion framework to learn a meaningful latent representation for all the entities in the ER graph even when they have no actually observed values in the *main task*.

3.3 The human reference interactome shows that novel interactions are hard to predict

We then retrieved the recently published Human Reference Interactome HI-III-19 data (Luck *et al.*, 2020), in which 17 500 genes have been screened with multiple Y2H assays, identifying 53 000 PPIs. Since the training set used by FPClass and by us has been published in 2014, the novel PPIs published during this 6 years hiatus can indeed provide a true blind test validation for PPI predictors.

We thus mapped the HI-III-19 genes to SwissProt (UniProt Consortium, 2015) protein IDs and we removed all the PPIs that were already present in the training set used in this study, obtaining 43 662 previously unseen PPIs (we refer to this dataset as HURI from now on). We used the G2 model to assign a likelihood score to these PPIs, and we compared the distribution of the predictions on HURI with the distribution of the cross-validated predictions obtained for the positive and negative classes on the training set, as shown in Supplementary Figure S1. From this plot, it appears that identifying the interacting pairs in the HURI dataset is indeed a challenging task, since the distribution of the HURI predictions is very

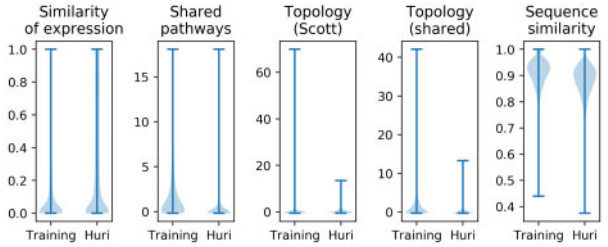


Fig. 3. Comparison of the characteristics of PPIs in the HURI and in the training dataset by analyzing (from left to right) the similarity of the protein expression profiles, the number of shared pathways, the two topology-derived scores used in the model and the sequence similarity between proteins. In each panel, the *y* axis label is indicated above the plot. All ranksums *P*-values are $< 10^{-300}$

similar to the negative samples in the training set. Similar results are obtained for FPClass predictions, as shown in the same figure.

By applying to the HURI predictions the best discrimination threshold (0.096) computed on the CV predictions, G2 correctly classifies 2232 PPIs as positive. A similar result is produced by FPClass, which ranks as positive 5220 PPIs with an optimal threshold of 0.0226. To further investigate these predictions, we built a dataset of 44 000 non-likely interacting pairs from the 7288 proteins involved in HURI PPIs by randomly pairing them, and we used them as negatives against the available PPIs in HURI. We tested both G2 and FPClass on this dataset with positive and likely negative samples, and their AUC was respectively 0.51 and 0.52, indicating that the HURI dataset presents indeed a hard ML challenge given current models and training datasets.

To determine the possible reasons behind the poor performances obtained on HURI, we analyzed several factors that might highlight a *shift* in the distributions underlying training set and HURI. In Figure 3, we compared the positive samples in the training data and in the HURI data while characterizing different biological aspects that are relate to PPI. From left to right, we compared the distributions of the (i) Pearson correlation coefficient between the interacting proteins' expression profiles, (ii) the number of shared pathways, as evinced from dbNFSP (Liu *et al.*, 2016), (iii) the distributions of the topology scores used as features and (iv) the sequence similarity computed as the χ^2 pairwise distance between the 20-dimensional vectors describing the aminoacid frequency composition of each protein. From Figure 3, we can indeed see that the proteins in HURI tend to cluster less together than the pairs used in the training set (topology scores) and tend to share fewer pathways. This causes the topology scores, which have been reported as critical feature for PPI prediction (Kotlyar *et al.*, 2015; Scott and Barton, 2007) to have noticeably lower discrimination power, causing the models that rely on them to mistakenly identify many HURI pairs as not likely to interact. The sequence similarity between proteins has been used as feature to predict interactions in previous works (Zaki *et al.*, 2009), but again we see that HURI PPIs tend to have lower overall similarity with respects to the training set. Finally, we used the Uniprot (UniProt Consortium, 2015) annotation score, which is a measure of the amount of annotations available for a UniProtKB entry (e.g. a protein) to test whether HURI contained relatively less studied proteins. Although the difference is not striking, the average annotation score of the pairs of interacting proteins in HURI is significantly lower than the ones in the training set (p-value$= 6 \times 10^{-35}$, Cohen's $d = -0.23$).

3.4 Extending the current training sets with HURI data

From the results on the HURI dataset, it appears that PPIs in HURI have significantly shifted underlying distributions, and this data could thus be extremely valuable to improve the generalization ability of predictors and reduce their biases toward well-studied regions of the Human interactome, since it largely consists of PPIs that are not trivial to identify given the currently used features and data.

We thus merged the training set from FPClass with the HURI dataset, and we evaluated our models in CV. Table 2 shows the

Downloaded from https://academic.oup.com/bioinformatics/article/37/16/2275/6131674 by KU Leuven Libraries user on 29 October 2024

Table 2. The performance of our G2 and G3 models in cross-validation on the data obtained by merging the training set and the HURI dataset

Methods	AUC	AUPRC
G1	91.0	50.7
G2 (old topology)	94.9	65.2
G2 (new topology)	95.9	72.4
G3 (old topology)	94.8	64.5
G3 (new topology)	95.9	71.4

results obtained. The AUC values are slightly lower than the ones obtained in Table 1, but the problem is still heavily unbalanced, even though there are now around 60 000 positives instead of 20 000 against 2 million negative samples. The AUPRC of the G1 model improved significantly in this benchmark, reaching 0.51. The topology features used by model G2 and G3 still help the prediction even when they are computed from the original training data (old topology), but clearly provide a larger contribution if they are computed from the training+HURI data (new topology). Also in this scenario, G2 model slightly outperforms G3.

4 Discussion

4.1 Data fusion as an extreme multi-task learning problem

In this article, we present an NN-based data fusion method which generalizes the classic MF paradigm by allowing non-linear inference over arbitrarily complex data represented as ERs graphs. This framework performs inference by learning to reconstruct the observed values in the data matrix (representing each relation between two entities) by optimizing the latent variables associated to each Entity. The inference over the ER graph is thus approached by translating it into a multi-task learning problem where the reconstruction error over each matrix is simultaneously minimized, allowing the model to learn entity-specific latent representations e_i that are thus influenced by every relation in which e_i is involved. This dualism between the conceptual organization of the data (ER graph) and the actual model is shown in Figure 1.

Our approach is completely general and allows inference on an extremely broad class of problems, since the NN formalism on which it is based guarantees (i) its scalability to big data problems, thanks to mini-batching and the use of GPUs, (ii) its modularity and customizability in the form of full flexibility for what concerns model complexity, number and types of layers, loss functions and optimizers. Finally, (iii) the ease with which the ER graph can be extended by connecting new entities to it by adding novel relations can allow the inclusion of data that is only loosely related with the problem under investigation and that may be discarded by a more conventional method.

Since the ER graph is processed in a multi-task learning fashion, the leading intuition for its design is that the prediction of the *main task Y* (e.g. the PPI in model G3) will benefit from the introduction of other relations if training on these tasks has a *synergistic* effect with respect to predicting *Y*. In this case, the auxiliary tasks would act as a sort of *informed regularization* and thus help learning *Y* and the generalization ability of the model. On the contrary, adding tasks that might have an *antagonistic* effect with respect to the main task *Y*, or that are simply totally unrelated could impair the learning of meaningful latents, since the model is *distracted* by an additional irrelevant task. Unfortunately, the role of antagonistic and synergistic tasks in multi-task learning has not been explicitly defined and investigated yet, since this phenomenon is heavily intertwined with the problem of assigning suitable weights to each task, but it is not exactly coincident (Kendall *et al.*, 2018).

Given an ML problem with a main task *Y* and k auxiliary tasks T_i, we define the optimal weights $W = \{w_0, \ldots, w_k\}$ associated to each task-specific loss function $L_i(\cdot,\cdot)$ as the values of W that maximize the generalization and prediction accuracy of the main task *Y*. These W are indeed influenced by at least three aspects, which are non-trivial to disentangle. These factors are: (i) the different number of labeled samples in the datasets associated to each task T_i, (ii) the type of prediction and loss function associated to each T_i (e.g. binary classification or regression with labels with unbounded domains) and (iii) the actual synergistic or antagonistic effect that the simultaneous learning of each task has on the main task *Y*. To the best of our knowledge, current studies proposing weighting methods for multi-task learning problems (Bengio *et al.*, 2009; Chen *et al.*, 2018; Guo *et al.*, 2018; Kendall *et al.*, 2018; Li *et al.*, 2016) do not explicitly distinguish between these effects and do not directly address the optimization of the synergy, possibly because it is not trivial to rigorously define and measure it in the ML context. This is thus an interesting open problem for multi-task learning that could greatly benefit the data fusion approaches based on the multi-task learning paradigm.

5 Conclusions

In this article, we presented a novel data fusion method that provides a full generalization of the existing MF-based approaches by allowing non-linear inference over arbitrary ER graphs, and we applied it to the PPI prediction task. We trained and tested three different models based on different underlying ER graphs and we show that both the tensor factorization (model G2) and the full ER graph approach (G3) outperform a state of the art data mining model over a interactome-scale dataset. We further validated it on a small dataset of high-quality PPIs and finally, we applied it on the recently published Human Reference Interactome HI-III-19 data which contains 50 000 novel PPIs. We show that this dataset (called HURI in this article) contains PPIs that are particularly difficult to predict for our G2 model and for the state of the art method. We thus investigated the possible reasons for this issue and we found that the PPIs in the HURI dataset show a significant shift in the distributions describing several biological properties underlying its interacting proteins, effectively making them unrecognizable by the models trained on precedent data. We thus conclude that the HURI data could be beneficial to improve the generalization ability of PPI predictors over less studied regions of the proteome and we show the results of our G1-3 models when they are benchmarked on this extended data.

Acknowledgements

D.R. is grateful to A.L. Mascagni, M. Kotlyar, G. Orlando, N. Verplaetse, and A. Passemiers for the constructive discussion and to M. Kotlyar for kindly providing FPClass data.

Funding

D.R. is funded by a Research Foundation – Flanders (FWO) post-doctoral fellowship.

Conflict of Interest: none declared.

Data availability

The PPI data and the NXTppi predictor are available at https://bitbucket.org/eddiewrc/nxtppi. The NXTfusion data fusion library is available at https://bitbucket.org/eddiewrc/nxtfusion/ and as a python package at https://pypi.org/project/NXTfusion/.

References

Arany,A. *et al.* (2015). Highly scalable tensor factorization for prediction of drug-protein interaction type. *arXiv preprint arXiv:1512.00315*.

Bengio,Y. *et al.* (2009). Curriculum learning. In: *Proceedings of the 26th Annual International Conference on Machine Learning*. pp. 41–48.

Downloaded from https://academic.oup.com/bioinformatics/article/37/16/2275/6131674 by KU Leuven Libraries user on 29 October 2024

Bitbol,A.-F. *et al.* (2016) Inferring interaction partners from protein sequences. *Proc. Natl. Acad. Sci. USA*, **113**, 12180–12185.

Braun,P. *et al.* (2009) An experimentally derived confidence score for binary protein-protein interactions. *Nat. Methods*, **6**, 91–97.

Chen,M. *et al.* (2019) Multifaceted protein–protein interaction prediction based on siamese residual rcnn. *Bioinformatics*, **35**, i305–i314.

Chen,Z. *et al.* (2018). Gradnorm: gradient normalization for adaptive loss balancing in deep multitask networks. In: *International Conference on Machine Learning*. PMLR, pp. 794–803.

Cong,Q. *et al.* (2019) Protein interaction networks revealed by proteome coevolution. *Science*, **365**, 185–189.

Fan,J. and Cheng,J. (2018) Matrix completion by deep matrix factorization. *Neural Netw.*, **98**, 34–41.

Finn,R.D. *et al.* (2014) Pfam: the protein families database. *Nucleic Acids Res.*, **42**, D222–D230.

Goh,K.-I. *et al.* (2007) The human disease network. *Proc. Natl. Acad. Sci. USA*, **104**, 8685–8690.

Guo,M. *et al.* (2018). Dynamic task prioritization for multitask learning. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. pp. 270–287.

Jeong,C.-S. and Kim,D. (2016) Inferring Crohn's disease association from exome sequences by integrating biological knowledge. *BMC Med. Genomics*, **9**, 1–9.

Kendall,A. *et al.* (2018). Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 7482–7491.

Kotlyar,M. *et al.* (2015) *In silico* prediction of physical protein interactions and characterization of interactome orphans. *Nat. Methods*, **12**, 79–84.

Li,C. *et al.* (2016). Self-paced multi-task learning. *arXiv preprint arXiv: 1604.01474*.

Lin,T.-Y. *et al.* (2017). Focal loss for dense object detection. In: *Proceedings of the IEEE International Conference on Computer Vision*. pp. 2980–2988.

Liu,X. *et al.* (2016) dbnsfp v3. 0: a one-stop database of functional predictions and annotations for human nonsynonymous and splice-site snvs. *Hum. Mutat.*, **37**, 235–241.

Luck,K. *et al.* (2020) A reference map of the human binary protein interactome. *Nature*, **580**, 402–408.

Meyer,M.J. *et al.* (2013) Instruct: a database of high-quality 3d structurally resolved protein interactome networks. *Bioinformatics*, **29**, 1577–1579.

Mnih,A. and Salakhutdinov,R.R. (2008). Probabilistic matrix factorization. In: *Advances in Neural Information Processing Systems*. pp. 1257–1264.

Orlando,G. *et al.* (2016) Observation selection bias in contact prediction and its implications for structural bioinformatics. *Sci. Rep.*, **6**, 36679.

Paszke,A. *et al.* (2019). PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems* 32, pp. 8024–8035.

Piñero,J. *et al.* (2017). Disgenet: a comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic Acids Res.*, **45**, D833–D839.

Rolland,T. *et al.* (2014) A proteome-scale map of the human interactome network. *Cell*, **159**, 1212–1226.

Schwartz,A.S. *et al.* (2009) Cost-effective strategies for completing the interactome. *Nat. Methods*, **6**, 55–61.

Scott,M.S. and Barton,G.J. (2007) Probabilistic prediction and ranking of human protein-protein interactions. *BMC Bioinformatics*, **8**, 239.

Simm,J. *et al.* (2015). Macau: scalable Bayesian multi-relational factorization with side information using mcmc. *arXiv preprint arXiv:1509.04610*.

Snider,J. *et al.* (2015) Fundamentals of protein interaction network mapping. *Mol. Syst. Biol.*, **11**, 848.

Sprinzak,E., and Margalit,H. (2001) Correlated sequence-signatures as markers of protein-protein interaction. *J. Mol. Biol.*, **311**, 681–692.

Srebro,N. and Jaakkola,T. (2003). Weighted low-rank approximations. In: *Proceedings of the 20th International Conference on Machine Learning (ICML-03)*. pp. 720–727.

Trigeorgis,G. *et al.* (2016) A deep matrix factorization method for learning attribute representations. *IEEE Trans. Pattern Anal. Mach. Intell.*, **39**, 417–429.

UniProt Consortium. (2015) Uniprot: a hub for protein information. *Nucleic Acids Res.*, **43**, D204–D212.

Zaki,N. *et al.* (2009) Protein-protein interaction based on pairwise similarity. *BMC Bioinformatics*, **10**, 150.

Zeng,H. *et al.* (2018) Complexcontact: a web server for inter-protein contact prediction using deep learning. *Nucleic Acids Res.*, **46**, W432–W437.

Zhang,Q.C. *et al.* (2012) Structure-based prediction of protein–protein interactions on a genome-wide scale. *Nature*, **490**, 556–560.

Žitnik,M. and Zupan,B. (2014a) Data fusion by matrix factorization. *IEEE Trans. Pattern Anal. Mach. Intell.*, **37**, 41–53.

Žitnik,M. and Zupan,B. (2014b). Matrix factorization-based data fusion for gene function prediction in baker's yeast and slime mold. *Pac Symp Biocomput.*, 400–411.

Žitnik,M. *et al.* (2015) Gene prioritization by compressive data fusion and chaining. *PLoS Comput. Biol.*, **11**, e1004552.

Downloaded from https://academic.oup.com/bioinformatics/article/37/16/2275/6131674 by KU Leuven Libraries user on 29 October 2024