Supplementary Information

1 Figures

Figure 1: An example of a supervised classification method for predicting protein interactions

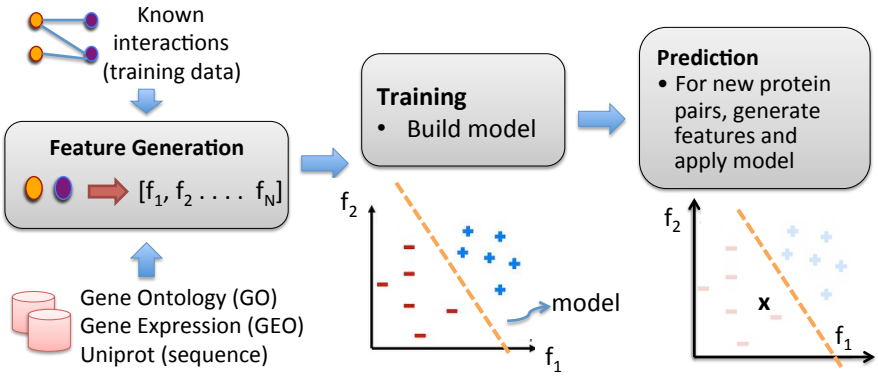
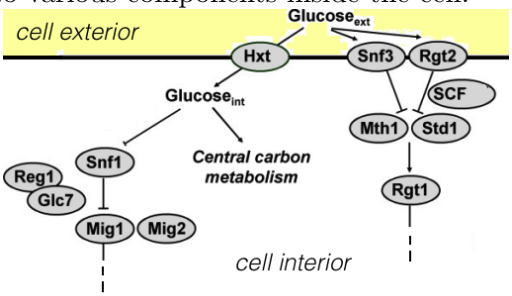


Figure 2: Part of the "glucose transport pathway" in human. Grey nodes represent the human proteins (genes) involved. Edges represent causality in the process. This pathway involves the transport of glucose from outside the cell to various components inside the cell.



2 Feature description and details

The following features were defined for the 3 high throughput human-bacterial datasets for bacterial species: *B. anthracis*, *F. tularensis*, *Y. pestis*. The features for *S. typhi.* were downloaded from the authors' website.

1. **GO similarity features**: These features model the similarity between the functional properties of two proteins. Gene Ontology (GO) provides GO-term annotations for three important protein properties: molecular function (F), cellular component (C) and biological process (P). We derive 3 types of features using these three properties. The similarity between two individual GO terms was computed using the G-Sesame algorithm. This feature can be considered as a matrix M of all the GO term combinations found in a given protein pair: $< p_b, p_h >$. The rows of the matrix represent GO terms from protein p_b and the columns represent GO terms from p_h. In this matrix M, we set the features corresponding to each of the GO-term combinations to be 1 and the remaining features to be 0. This feature thus records the co-occurrance of GO terms.

2. **Protein sequence n-mer or n-gram features**: Since the sequence of a protein determines its structure and consequently its function, it may be possible to predict PPIs using the amino

acid sequence of a protein pair. Shen et al. (2007) introduced the "conjoint triad model" for predicting PPIs using only amino acid sequences. Shen et al. (2007) partitioned the twenty amino acids into seven classes based on their electrostatic and hydrophobic properties. For each protein, they counted the number of times each distinct three-mer (set of three consecutive amino acids) occurred in the sequence. To account for protein size, they normalized these counts by linearly transforming them to lie between 0 and 1 (see (Shen et al., 2007) for details). They represented the protein with a 343-element feature vector, where the value of each feature is the normalized count for each of the 343 (73) possible amino acid three-mers. We use two-, three-, four-, and five-mers. For each hostpathogen protein pair, we concatenated the feature vectors of the individual proteins. Therefore, each hostpathogen protein pair had a feature vector of length at most 98, 646, 4802, and 33614, in the cases of two-, three-, four-, and five-mers, respectively.

3. **Graph based features using the human interactome**: These features are derived using only the human protein 'p_h' from the pair. Pathogens generally target host proteins that are important in several host processes; these host proteins interact with many other host proteins to carry out their tasks. This insight is captured in the form of three graph properties: *degree*, *between-ness centrality* and *clustering coefficient* of the human protein "node" in the human interactome graph. The human interactome was downloaded from HPRD. The degree of a node is the number of its neighbouring nodes in the graph. The clustering coefficient of a node 'n' is defined as: the ratio of the number of edges present amongst n's neighbors to the number of all possible edges that could be present between the neighbours. Betweenness centrality for a node 'n', is defined as the sum over all pairs of nodes (u, v), the fraction of shortest paths from u to v, that pass through n. Mathematically, it is: $\sum_{u,v \in V \setminus n} \frac{\text{shortest\_paths}_n(u,v)}{\text{shortest\_paths}(u,v)}$. Intuitively, nodes that occur on many shortest paths between other vertices have higher betweenness than those that do not.

4. **Gene expression features**: The intuition behind this feature is that genes that are significantly differentially regulated upon being subject to bacterial infection, are more likely to be involved in the infection process. These features are derived using the gene of the human protein 'p_h' from the pair. We selected transcriptomic datasets GSE12131, GSE14390, GSE5966 for *B. anthracis*, GSE12108, GSE22203 for *F. tularensis* and GSE22299, GSE18293 for *Y. pestis* from the GEO database. These give the differential gene expression of human genes infected by the bacteria, under different control conditions.

3 Precision Recall curves from 10 fold CV results

We plot the recall vs the precision obtained by our method, MTPL on the 4 tasks in Figure 3. We used the results from the 10 fold CV experiments. The classifier score for each test instance was aggregated from the various pairwise models a manner similar to what is explained in Section 4.3 of the main paper. Let the classifier scores (i.e $\mathbf{w}^\intercal \mathbf{x}$) from each model for a given test instance $\mathbf{x}$ be $\{s_1, s_2 \dots s_{m-1}\}$. The aggregated multi-task classifer score of $\mathbf{x}$ is given by:

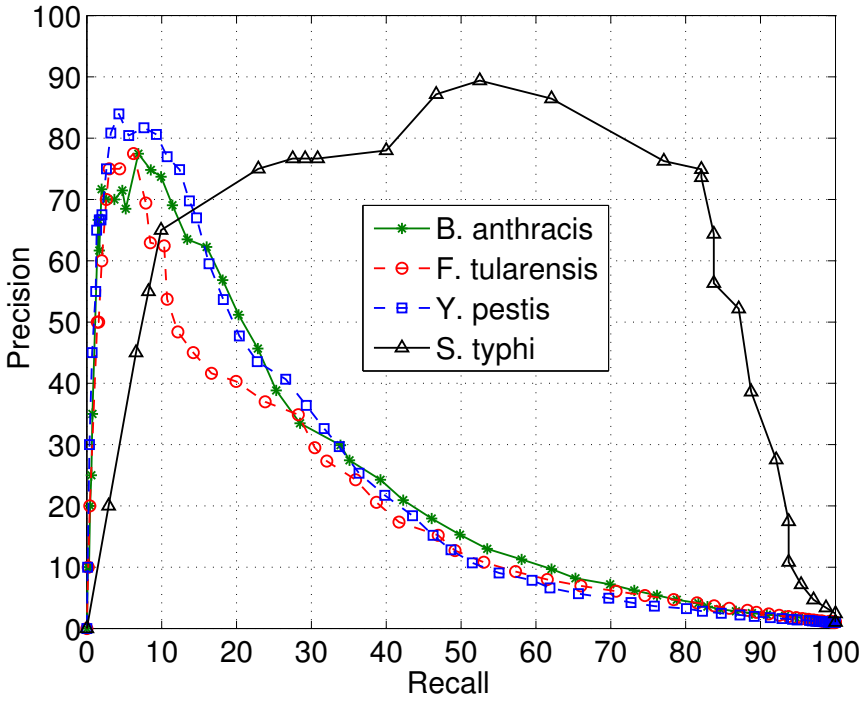
$$s(\mathbf{x}) = \begin{cases} \max\limits_i \ s_i, & \text{if } \left(\sum_i I(s_i \geq 0)\right) \geq 1 \\ \min\limits_i \ s_i, & \text{otherwise} \end{cases} \tag{1}$$

The classifier threshold was then varied and the precision (P), recall (R) were computed for each threshold. The final curve was obtained by aggregating the P-R curves from each of the ten folds.

4 Results from 50 bootstrap sampling experiments

We validate the improvement in performance by checking for statistical significance using Paired t-tests on 50 bootstrap sampling experiments. We only compare MTPL (our method) with Indep.

Figure 3: Precision-Recall curves for MTPL



(independent models). Each bootstrap sampling experiment consists of the following procedure: we first make two random splits of 80% and 20% of the data, such that the class ratio of 1:100 is maintained in both. The training set is then constructed using a bootstrap sample from the 80% split and the test data from the 20% split. A total of 50 models are thus trained and evaluated. We do not tune parameters again for each model and instead use the optimal setting of parameter-values from our 10 fold CV experiments. The F1 is computed for each experiment thereby giving us 50 values, which will be our samples for the hypothesis test. The F1 values averaged from 50 experiments are in the table below.

Note that the overall results of both methods are worse than the 10 fold CV. This happens due to the bootstrap sample, which allows an instance to be sampled multiple times. A sample of size n will have fewer than n unique instances. Hence, the effective size of the training data seen by the classifier is less than 80%. Compare this to the 10 fold CV experiments, where the classifier has access to 80% of the data during training.

The performance of MTPL is better than Indep for the three high throughput datasets. We see an improvement of 3.8 for *B. anthracis*, 2.7 for *F. tularensis*, 3.5 for *Y. pestis*. The Indep results are better for *S. typhi* by 3.2 F points.

Table 1: Averaged F1 from 50 bootstrap sampling experiments.

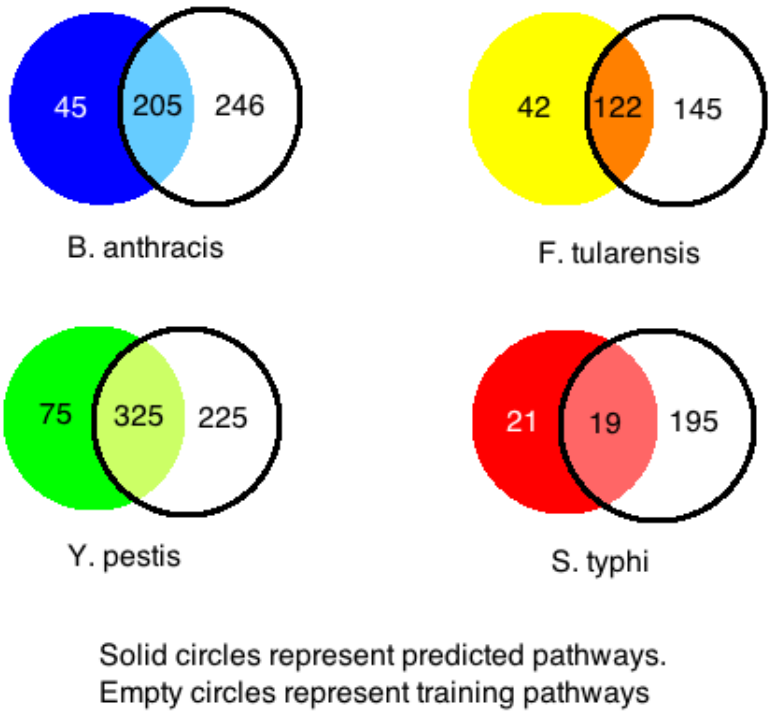
Method	*B. anthracis*	*F. tularensis*	*Y. pestis*	*S. typhi*
Indep.	24± 5.6	23± 5.7	25.3± 5.2	**65.2** ± 9.6
MTPL	**27.8** ±5.5	**25.7**± 5.7	**28.8** ± 4.4	62 ± 13.6

5 Intersection of pathways enriched in the PPIs from training and predicted.

Here we show the intersection between the pathways enriched in the predicted interactions and the pathways enriched in the gold-standard positives. For both enrichment computation, the human genes from the interactions are considered. We used Fisher's exact test and a p-value cut-off of 1e-7. The filled circles on the left of each intersection represent the enriched pathways in the predictions. The

empty circles on the right show the enriched pathways in the training data. We can see that there are several new pathways enriched in the predictions as compared to those enriched in the gold-standard data.

Figure 4: Enrichment intersection between training PPIs and predicted PPIs. Cut-off used for enrichment: **1e-7**.



6 List of 17 pathways commonly enriched from predictions across all bacterial datasets

Table 2: Five of the 17 commonly enriched pathways in the predicted interactions from MTPL.

Adaptive Immune System
Developmental Biology
E-cadherin signaling events
E-cadherin signaling in the nascent adherens junction
Glypican pathway
Immune System
Integrin alphaIIb beta3 signaling
Integrin cell surface interactions
L1CAM interactions
N-cadherin signaling events
Platelet activation, signaling and aggregation
Platelet Aggregation (Plug Formation)
Posttranslational regulation of adherens junction stability and dissassembly
Signalling by NGF
Signal Transduction
Stabilization and expansion of the E-cadherin adherens junction
TNF alpha/NF-kB