Large-Scale Protein-Protein Post-Translational Modification Extraction With Distant Supervision And Confidence Calibrated BioBERT

Aparna Elangovan
The University of Melbourne
Yuan Li
The University of Melbourne
Douglas E.V. Pires
The University of Melbourne
Melissa J. Davis
Walter and Eliza Hall Institute of Medical Research
Karin Verspoor (karin.verspoor@rmit.edu.au)
RMIT University

Research Article

Keywords: Protein-protein interaction, Post-translational modifications, BioBERT, Natural language processing, Deep learning, Distant supervision

Posted Date: September 22nd, 2021

DOI: https://doi.org/10.21203/rs.3.rs-898489/v1

Version of Record: A version of this preprint was published at BMC Bioinformatics on January 4th, 2022.
See the published version at https://doi.org/10.1186/s12859-021-04504-x.

RESEARCH

Large-scale protein-protein post-translational modification extraction with distant supervision and confidence calibrated BioBERT

Aparna Elangovan[1], Yuan Li[1], Douglas E. V. Pires[1], Melissa J. Davis[2,3] and Karin Verspoor[1,4*]

*Correspondence:
karin.verspoor@rmit.edu.au
[4]School of Computing Technologies RMIT University, Melbourne, Australia
Full list of author information is available at the end of the article

Abstract

Motivation: Protein-protein interactions (PPIs) are critical to normal cellular function and are related to many disease pathways. A range of protein functions are mediated and regulated by protein interactions through post-translational modifications (PTM). However, only 4% of PPIs are annotated with PTMs in biological knowledge databases such as IntAct, mainly performed through manual curation, which is neither time- nor cost-effective. Here we aim to facilitate annotation by extracting PPIs along with their pairwise PTM from the literature by using distantly supervised training data using deep learning to aid human curation. We further assessed model generalisation in a real-world scenario, evaluating its performance on a randomly sampled subset of predictions from 18 million PubMed abstracts.

Method: We use the IntAct PPI database to create a distant supervised dataset annotated with interacting protein pairs, their corresponding PTM type, and associated abstracts from the PubMed database. We train an ensemble of BioBERT models – dubbed PPI-BioBERT-x10 – to improve confidence calibration. We extend the use of ensemble average confidence approach with confidence variation to counteract the effects of class imbalance to extract high confidence predictions.

Results and conclusion: The PPI-BioBERT-x10 model evaluated on the test set resulted in a modest F1-micro 41.3 (P=58.1, R=32.1). However, by combining high confidence and low variation to identify high quality predictions, tuning the predictions for precision, we retained 19% of the test predictions with 100% precision. We evaluated PPI-BioBERT-x10 on 18 million PubMed abstracts and extracted 1.6 million (546507 unique PTM-PPI triplets) PTM-PPI predictions, and filter ≈5,700 (4584 unique) high confidence predictions. Of the 5700, human evaluation on a small randomly sampled subset shows that the precision drops to 33.7% despite confidence calibration and highlights the challenges of generalisability beyond the test set even with confidence calibration. We circumvent the problem by only including predictions associated with multiple papers, improving the precision to 58.8%. In this work, we highlight the benefits and challenges of deep learning-based text mining in practice, and the need for increased emphasis on confidence calibration to facilitate human curation efforts.

Keywords: Protein-protein interaction; Post-translational modifications; BioBERT; Natural language processing; Deep learning; Distant supervision

Background

Critical biological processes, such as signaling cascades and metabolism, are regulated by protein-protein interactions (PPIs) that modify other proteins in order to modulate their stability or activity via post-translational modifications (PTMs). PPIs are curated in large online repositories such as IntAct [1] and HPRD [2]. However, most PPIs are not annotated with a function, for example, we found the IntAct database has over 100,000 human PPIs, but less than 4000 of these are annotated with PTMs such as phosphorylation, acetylation or methylation. Understanding the nature of PTM between an interacting protein pair is critical for researchers to determine the impact of network perturbations and downstream biological consequences. PPIs and PTMs in biological databases are usually manually curated, which is time consuming and requires highly trained curators. [3]. Orchard *et al.* [4] have also highlighted additional challenges in maintaining manually curated databases, ensuring they are up to date, as well as the economic aspects of manual curation. Hence, the adoption of automated curation methods is essential for sustainability of this work.

Here we extract PTMs by text mining PubMed abstracts, extracting protein pairs along with their corresponding PTM. Given an input journal abstract, the output is a triplet of the form [Protein1, PTM function, Protein2] where Protein1 and Protein2 are Uniprot identifiers [5] of the proteins. We also aim to aid human curation of PTM-PPIs, hence we assess how well machine learning models generalise by applying them to 18 million PubMed abstracts to extract PTM-PPI triplets. In this paper, we use confidence calibration [6, 7] as a mechanism to understand generalisability to know when a prediction works to extract high quality predictions. We believe our paper is the first to study the practical applicability and challenges of large scale PTM-PPI extraction using NLP with deep learning and distant supervision.

We focus on extracting PTMs including phosphorylation, dephosphorylation, methylation, ubiquitination, deubiquitination, and acetylation (these PTMs were selected based on the availability of training data). We create a training dataset using distant supervised approach [8, 9] using IntAct [1] as the source knowledge base to extract PTM-PPI triplets from PubMed abstracts. We train an ensemble of BioBERT [10] models to improve neural confidence calibration [6, 7]. We then applied the trained model on 18 million PubMed abstracts to extract PPI pairs along with their corresponding PTM function and attempt to ensure high quality predictions using neural confidence calibration techniques [6, 7] to augment and facilitate human curation efforts.

Related works in protein interaction extraction through deep learning

The datasets in PPI extraction such as AIMed [11] and BioInfer [12] used to evaluate text mining approaches have remained the same for over a decade (since 2007) and focus on extracting protein interactions but not the nature of the PTM interaction between them. These datasets have also been used to evaluate the latest machine learning approaches including deep learning [13, 14] in protein pair extraction. However, the latest deep learning trends do not seem to be widely popular in PPI curation outside the limited context of benchmarking methods using the AIMed [11] and BioInfer [12] datasets. Automated PPI curation attempts using text mining and rule-based approaches seem more prevalent [15, 16, 17].

STRING v11 [18], one of the most popular PPI databases, uses text mining as a curation method. Their text mining pipeline has largely remained the same since STRING v9.1 [15]. STRING v9.1 uses a weighted PPI co-occurrence rule-based approach, where the weights depend on whether a protein pair occurs together within the same document, the same paragraph and/or the same sentence. Rule based approaches can be quite effective [19, 15] even with limited training data, depending on the task. For instance, Szklarczyk *et al.* define an interaction unit

in STRING v11 database [18] as a *"functional association, i.e. a link between two proteins that both contribute jointly to a specific biological function"*. This definition allows for co-occurrence rule-based approach to be quite effective, *i.e.* if a protein pair co-occurs in text frequently then the pair is highly likely to be related.

iPTMnet [20] consolidates information about PPIs and PTMs from various manually curated databases such as HPRD [2] and PhosphoSitePlus [21] as well as text mining sources. For text mining, iPTMnet uses RLIMS-P [17] and eFIP [16] to automatically curate enzyme-substrate-site relationships. These tools use rule-based approaches using text patterns to extract proteins involved in PTM. The iPTMnet statistics dated Nov 2019 (`https://research.bioinformatics.udel.edu/iptmnet/stat`) indicate that the total number of enzyme-substrate pairs curated using RLIMS-P [17] is fewer than 1,000 pairs. This modest number highlights the main challenge using text patterns: while they can extract relationships with fairly high precision, they are not robust to variations in how PPI relations can be described in text. We therefore explore machine learning-based methods.

Automatic extraction of PPIs using deep learning can be beneficial as it has the potential to extract PPIs from a variety of text where the PPI relationships are described in ways that cannot be easily captured by a manually crafted rule-based system. However, deep learning requires a much larger volume of training data. Generalisability of the model to ensure prediction quality is key to its widespread adoption for automatic extraction of PPI relationships from text at scale. Enhancing quality of prediction at large scale needs to focus on reducing false positives to minimise corrupting existing knowledge base entries (*i.e.* predicting that a relationship exists when it doesn't) and, hence, confidence calibration approaches to reduce poor quality predictions become a crucial step in large scale text mining. Confidence calibration is the problem of predicting probability estimates representative of the true correctness [7] and in this paper we use confidence calibration to know when

a prediction is likely correct and use it as a mechanism to improve generalisation. The aspects of generalisability have been largely limited to the evaluation of a test set and the limitations of using the test set performance as a proxy for real world performance have been challenged in previous studies [22, 23].

Creating gold standard training data with fine-grained annotations is a manual, labor-intensive task and is a limiting factor in applying machine learning to new domains or tasks. Being able to leverage one or more existing data sources is key to using machine learning in new domains or for new tasks. Distant supervision [8, 9] exploits existing knowledge bases, such as IntAct [1], instead of annotating a new dataset. However there are two main limitations to the use of distant supervision datasets: **(a)** noisy labels require noise reduction techniques to improve label quality [24] **(b)** they require negative samples to be generated as the databases usually only contain positive examples of a relationship.

Deep learning architectures such as BiLSTM and BioBERT have been previously used to benchmark methods for protein relation extraction using Natural Language Processing (NLP) and the AIMed dataset [13, 25, 22]. However, these works [13, 25, 22] do not measure the ability of these models to calibrate confidence scores. We chose a state-of-the-art deep learning approach, BioBERT [10], train an ensemble to enhance confidence calibration [6] and use confidence variation to counteract the effects of class imbalance during confidence calibration.

Methods

Dataset

We obtain the dataset of human PTM-PPI interactions from the IntAct database [1], a database that is part of the International Molecular Exchange (IMEX) [26] Consortium. The database contains Uniprot identifiers, protein aliases, interaction type (where available), and the PubMed identifier of the paper describing the interaction. Of the over 100,000 human PPIs in IntAct [1], only 3,381 PPIs describe PTM

interactions. Hence, we start with an initial dataset containing these 3,381 PPIs. We then remove duplicate entries, *i.e.* the interacting proteins, the PubMedId and the interaction type are the same resulting in 2797 samples. We then pre-process this data to remove self-relations (where participant protein1 is the same as protein2).

Table 1 Illustration of data preparation. The IntAct database has the PubMedId, the Uniprot identifiers of the participating proteins and the interaction type. More than 1 pair of interacting proteins can be annotated against a given PubMed Id and not all of these interactions are described in the abstract. This forms the noisily labelled training data.

IntAct	**Pubmed Id:** 24291004 , **Uniprot protein pair:** P04150, P31749, **Interaction type:** Phosphorylation
Corresponding abstract	*Glucocorticoid resistance ..[truncated display]....* we identify the AKT1 kinase as a major negative regulator of the NR3C1 glucocorticoid receptor protein activity driving glucocorticoid resistance in T-ALL. Mechanistically, **AKT1 impairs glucocorticoid-induced gene expression by direct phosphorylation of NR3C1** at position S134 and blocking glucocorticoid-induced NR3C1 translocation to the nucleus. Moreover, we demonstrate that loss of PTEN and consequent AKT1 activation can effectively block glucocorticoid-induced apoptosis and induce resistance to glucocorticoid therapy. Conversely, pharmacologic inhibition of AKT with MK2206 effectively restores glucocorticoid-induced NR3C1 translocation .... *in vitro and in vivo*.
Genes and normalised NCBI ids	• **Start-End, Gene Mention, NCBI gene Id** • 137 - 141, AKT1, 207 • 186 - 191, NR3C1, 2908 • 294 - 298, AKT1, 207 • 375 - 380, NR3C1, 2908 • 434 - 439, NR3C1, 2908 • 508 - 512, PTEN, 5728 • 528 - 532, AKT1, 207 • 748 - 753, NR3C1, 2908
NCBI to Uniprot map	• **NCBI gene: Uniprot identifiers** • 2908: P04150, E5KQF5, E5KQF6, F1D8N4, B7Z7I2 • 207: P31749, B0LPE5, B3KVH4 • 5728: P60484, F6KD01
Normalised abstract for pair P04150, P31749	*Glucocorticoid resistance ....[truncated display]....* we identify the P31749 kinase as a major negative regulator of the P04150 glucocorticoid receptor protein activity driving glucocorticoid resistance in T-ALL. Mechanistically, **P31749 impairs glucocorticoid-induced gene expression by direct phosphorylation of P04150** at position S134 and blocking glucocorticoid-induced P04150 translocation to the nucleus. Moreover, we demonstrate that loss of P60484 and consequent P31749 activation can effectively block glucocorticoid-induced apoptosis and induce resistance to glucocorticoid therapy. Conversely, pharmacologic inhibition of AKT with MK2206 effectively restores glucocorticoid-induced P04150 translocation ..... *in vitro and in vivo*.
Negative sample pairs	• P04150, P60484 • P31749, P60484

Once we pre-process the data, the steps involved in dataset transformation is illustrated in Table 1. We first identify all the gene mentions in an abstract and their corresponding Uniprot protein identifier [5], as detailed in section *Gene mentions and protein identifiers*. We apply noise reduction techniques to remove samples where the abstract may not describe the PPI relationship, as detailed in section *Noise reduction*. We then split the dataset into train, test and validation sets such that they are stratified by interaction type and have unique PubMed ids in each set to avoid test set leakage resulting from random splits identified in the AIMed dataset [22]. Negative training samples are generated as detailed in section *Negative sample generation*.

Gene mentions and protein identifiers

To prepare the training data, the Uniprot identifiers mentioned in IntAct need to match the Uniprot identifiers that can be inferred from the abstract text by applying named entity recognition and normalisation NLP techniques, *gene mention* $\rightarrow$ *Uniprot accession number*. GNormPlus[27] identifies gene names and normalises them to the NCBI gene identifier [28]. The NCBI gene identifier is a gene-level identifier that needs to be converted to a Uniprot accession code which is a protein identifier. A single NCBI gene can be associated with multiple Uniprot accession numbers either due to biological reasons such as alternative splicing where a single gene results in different protein isoforms or due to technical reasons related to Uniprot accession number changes [5]. For a given abstract A, and Uniprot identifiers of the two PPI participants IU_1, IU_2 annotated in the IntAct database as $< A, IU_1, IU_2 >$, we perform the following:

1. Use GNormPlus to identify gene name mentions $[g_1, g_2..g_m]$ in abstract A, and to normalise gene names to corresponding NCBI gene ids, $NG = [ng_1, ng_2, .., ng_m]$.
2. Given a NCBI gene id, $ng_i \in NG$, we obtain a list of corresponding Uniprot accession numbers $U = [u_1, u_2, ..u_n]$.
3. If IU_1 or IU_2 exists in U, then we have found a matching IntAct Uniprot identifier IU_x annotated against the abstract, where $IU_x \in [IU_1, IU_2]$. Map NCBI gene id ng_i to Uniprot identifier IU_x. If neither IU_1 nor IU_2 exist in U, *i.e.* the Uniprot identified in the abstract is not annotated against the abstract, then simply return the first Uniprot identifier u_1.

This gene mention identification to Uniprot identifier mapping process is illustrated through an example in Table 1.

Noise reduction

In order to extract PTM-PPI relationships mentioned in the abstract, we aim to retain only those training samples where the PTM-PPI is described in the abstract. Hence we exclude IntAct PTM-PPI triplets from the training data if either participant's Uniprot identifiers does not exist in the normalized abstract. This is to minimize the false positive noise in the distant supervised dataset, *i.e.* removing samples where the relationship is not described in the abstract, based on the assumption that, if the proteins are not explicitly mentioned in the abstract then it is highly likely (unless protein names are not recognized by the NER tool) that the abstract does not describe the relationship between those proteins. We also remove PPIs where the stemmed interaction type (phosphoryl, dephosphoryl, methyl, acetyl, deubiquitin, ubiquitin, demethyl) is not mentioned in the abstract.

Negative sample generation

Since knowledge bases focus on relationships that exist, distant supervision-based approaches require negative training examples to be created. In our dataset, negative samples are protein pairs that are mentioned in the abstract but do not have a function referencing that paper annotated against the pair in Intact. In order to generate negative samples, we identify protein mentions from the abstracts using GNormPlus [27], which normalizes mentions to NCBI gene IDs which are then converted to Uniprot identifiers as described in section *Gene mentions and protein identifiers*. If a given protein pair $< p1, p2 >$, where $p1$ and $p2$ are mentioned in the abstract, but is not annotated against any of the 7 types of PPI relationship within the abstract according to the Intact database, then it is assumed that $< p1, p2 >$ form negative samples for that abstract, see example in Table 1.

It is important to emphasize that a negative sample does not mean that a given PPI relationship does not exist, but rather that the abstract does not describe such a

relationship. It could also be a noisy negative sample, *i.e.* the abstract describes the functional relationship between pair, but it is simply not captured in the annotation.

Training BioBERT for PPI extraction

We fine-tuned the pretrained BioBERT v1.1 [10], based on BERT-base-Cased, which is pretrained on a large collection of PubMed abstracts. We applied fine-tuning to adapt BioBERT to the PTM-PPI extraction task, as a multi-class classification problem, and assumed there is at most one type of PTM relationship between a protein pair. For full details on training configuration, see supplementary section A.1

The original BioBERT is trained using a sentence as a training sample for the language modelling task, while we work with complete abstracts. To utilize an entire abstract as a single training sample, we feed all the sentences within the abstract including the full stop (end-of-sentence marker, the period ".") separating each sentence. BioBERT has a maximum input sequence length of 512 units, this includes the sub word units to fit the entire abstract. Sub word units are concatenated to represent words and limit the vocabulary size, *e.g.*, a word such as Immunoglobulin is tokenized into 7 units (I, ##mm, ##uno, ##g, ##lo, ##bul, ##in) where ## are special symbols to indicate continuation [29]. Therefore, an abstract with 250 words can result in a much longer sequence once it is tokenised. In our training data, post tokenisation, 90% of the normalised abstracts are under the maximum limit of 512 sub units. For more detailed distribution see supplementary Table A10. Therefore, we simply truncate sequences longer than 510 (to accommodate the mandatory marker [CLS] and [SEP] tokens that BERT requires [29]) and feed the input abstracts.

Uncertainty estimation

In order to improve the probability estimate associated with each prediction, we use an ensemble of 10 PPI-BioBERT models (referred to as PPI-BioBERT-x10), all trained with exactly the same hyperparameters and training data but capturing

slightly different models due to Bernoulli (binary) dropout [30] layers in the network. The use of ensembles of models to improve uncertainty estimates, rather than just improving overall accuracy, has been shown to be effective in the computer vision task of image classification Lakshminarayanan *et al.* [6], hence we follow that approach. Using the ensemble of the 10 models, the predicted confidence $\hat{p_j}$ and predicted class $\hat{y_j}$ for the input x_j is:

$$\hat{p_j} = p_j(y = c) = \frac{1}{M}\Sigma_{i=1}^{M} p(y_{\theta_i} = c|x_j, \theta_i) \tag{1}$$

$$\hat{y_j} = argmax_c p(y = c) = \frac{1}{M}\Sigma_{i=1}^{M} p(y_{\theta_i} = c|x, \theta_i) \tag{2}$$

In addition to averaging confidence, we also capture the standard deviation of the predicted confidence across the ensemble as:

$$std(\hat{p_j}) = \sqrt{\frac{1}{M}\Sigma_{i=1}^{M}(p(y_{\theta_i} = c|x_j, \theta_i) - \hat{p_j})^2} \tag{3}$$

where M is the number of models in the ensemble; θ_i is model i; y_{θ_i} is the output predicted by model i; x is the input; c is the predicted class.

We measure the ability of the ensemble in confidence calibration using Expected Calibration Error (ECE) used previously by Guo et.al. [7] in computer vision. ECE is defined as:

$$accu(B_k) = \frac{1}{|B_k|}\Sigma_{j \in B_k} 1(\hat{y_j} = y_j) \tag{4}$$

$$conf(B_k) = \frac{1}{|B_k|}\Sigma_{j \in B_k} \hat{p_j} \tag{5}$$

$$ECE = \Sigma_{k=1}^{K} \frac{|B_k|}{n} |accu(B_k) - conf(B_k)| \tag{6}$$

where the confidence score is divided into K equally spaced bins B_k, n in the number of samples, $accu(B_k)$ is the accuracy of predictions whose confidence fall into bin

B_k and $conf(B_k)$ is the average confidence of predictions that confidence fall into bin B_k. The best calibrated model will have zero ECE.

Large scale PPI function extraction from PubMed abstracts

We extracted the complete collection of 18 million abstracts available in PubMed as of April 2019. We applied GNormPlus [27] to recognize the proteins and normalize them to UniProt identifiers. We then apply the PPI-BioBERT-x10 model to extract PTM-PPI relationships from the entire PubMed abstracts. In order to extract high quality PTM-PPI relationships, we only select predictions that have low variation in the confidence score and high prediction confidence.

Results

IntAct-based dataset

The final dataset includes the Uniprot identifiers of the participants, the normalized Abstract (the protein names normalised to Uniprot identifiers), all the proteins identified through NER in the abstract and the class label (the type of PTM interaction). After pre-processing and applying noise reduction to the original dataset with approximately over 3,000 interactions in IntAct, we are left with a total of 279 positive and 1579 negative samples across training, test and validation set. The distribution of interaction types and positive/negative samples is shown in Table 2, with phosphorylation forming almost 75% of the positive samples in the training set. The positive sample rate is between 14% - 17% in the train, test and validation set. Interaction types such as such acetylation, deubiquitination and ubiquitination, have less than 10 positive samples in total across training, test and validation.

Confidence calibration on the test and validation sets

The performance on the test and validation sets using the ensemble BioBERT model (PPI-BioBERT-x10) is detailed in Table 3. The test and validation sets have an F1-

Table 2 Train/Test/ Val Positive samples for each interaction type. NOTE The negative samples count against a interaction type is solely to indicate how many were derived from abstracts describing the interaction type. All negative samples belong to a single class "other / Negative".

	Train		Val		Test		Total	
	-	+	-	+	-	+	-	+
acetylation	31	5	2	1	9	1	42	7
dephosphorylation	167	28	36	10	33	6	236	44
deubiquitination	4	2	0	0	0	0	4	2
methylation	22	10	5	1	22	4	49	15
phosphorylation	862	139	118	21	227	44	1207	204
ubiquitination	30	5	5	1	5	1	40	7
Total	1116	189	166	34	296	56	1578	279

micro score of 41.3 and 58.6, respectively, across all interactions. The confusion matrix is shown in Figure 1.

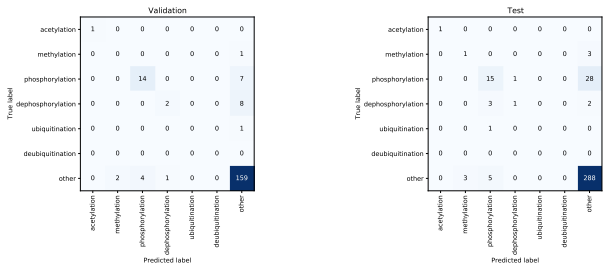

Figure 1 Confusion matrix on the test and validation set on PPI-BioBERT-x10.

We visualise the ensemble model's confidence calibration using reliability diagrams, similar to Guo *et al.* [7], as shown in Figure 2. We observe that the predicted confidence range for each interaction type is proportional to the percentage of training samples, *i.e.*, the negative samples have the highest proportion and over 80% of the negative sample predictions have a confidence score between 0.9 and 1.0, whereas acetylation, which has less than 1% of the training samples, has prediction confidence of less than 0.6. This relationship between the distribution of predicted confidence and the proportion of training samples for a given class label (interaction type) is consistent in the train, validation and test set as shown in Figure 2. This highlights 2 main limitations of measuring calibration error using ECE: **(a)** it penalizes model calibration regardless of the class imbalance in the training dataset

Table 3 The performance of ensemble PPI-BioBERT-x10 on the test and validation set. **ECE** is the expected calibration error. **SD** denotes the average standard deviation within the ensemble.

Dataset	Interaction	P	R	F1	ECE	SD	Support
Test	acetylation	100.00	100.00	100.00	0.49	0.25	1
Test	dephosphorylation	50.00	16.67	25.00	0.67	0.40	6
Test	methylation	25.00	25.00	25.00	0.60	0.28	4
Test	phosphorylation	62.50	34.09	44.12	0.79	0.26	44
Test	ubiquitination	0.00	0.00	0.00	-	-	1
Test	ECE	-	-	-	0.75	-	31
Test	average SD	-	-	-	-	0.28	31
Test	macro avg	47.50	35.15	38.82	-	-	56
Test	micro avg	58.06	32.14	41.38	-	-	56
Val	acetylation	100.00	100.00	100.00	0.53	0.16	1
Val	dephosphorylation	66.67	20.00	30.77	0.61	0.37	10
Val	methylation	0.00	0.00	0.00	0.53	0.29	1
Val	phosphorylation	77.78	66.67	71.79	0.78	0.26	21
Val	ubiquitination	0.00	0.00	0.00	-	-	1
Val	ECE	-	-	-	0.73	-	24
Val	average SD	-	-	-	-	0.28	24
Val	macro avg	48.89	37.33	40.51	-	-	34
Val	micro avg	70.83	50.00	58.62	-	-	34

when high quality predictions can be available at lower confidence scores for classes with lower proportion of samples **(b)** it penalises both over- and under-calibration equally. Hence, we find it challenging to rely on ECE and/or average ensemble confidence to extract high quality predictions, especially for interaction types that have a much lower proportion of training samples.

We therefore inspected the variation in the confidence score predicted by each model for a given sample as shown in Figure 3. Intuitively, if for a given input, all the models within the ensemble consistently predict with similar confidence then we can rely on the ensemble confidence better compared to high variation confidence scores. When we compare the variation in confidence scores of the correct *vs.* incorrect predictions, the incorrect predictions do not seem to have low variation in predicted confidence, as shown in Figure 3, except for 3 incorrect phosphorylation predictions. We apply a heuristic to select interaction-wise confidence and standard deviation thresholds where we select the 50^{th} percentile as the cut-off for both

confidence and confidence standard deviation based on the training set. This results in retaining 19% of positive predictions (6 out of 31) in the test set. Since our training data is noisy, we manually verify all predictions that have low variation in the test and validation set. On manual verification, we find all the predictions using this heuristic are in fact correct, including 3 incorrectly labelled ones, as shown in Table 4, indicating that a combination of low variation and relative high confidence has the potential to be robust against noisy labels.

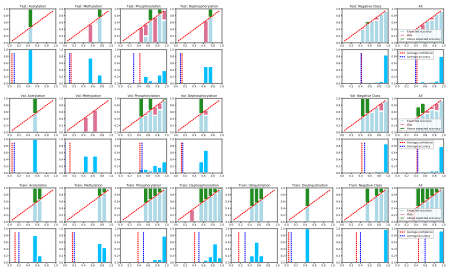

Figure 2 Confidence reliability on the test and validation set using PPI-BioBERT-x10. The average predicted confidence score across the ensemble, for a given input, falls into any one of the 10 equally spaced confidence bins.

Table 4 Manually verified relationships in the validation and test set with ensemble prediction standard deviation less than interaction-wise threshold and predicted confidence greater than interaction-wise threshold, taking the precision to 100.0

PubMed	Phrases describing relationships in the abstract	Label	Prediction
	Validation		
12150926	**mTOR (P42345)**-catalyzed phosphorylation of **4EBP1 (Q13541)** in vitro	phosphorylation	phosphorylation
15733869	**SGK (O00141)** physically associates with **CREB (P16220)** and **SGK (O00141)** phosphorylates it on serine 133	phosphorylation	phosphorylation
15557335	**Src (P12931)** phosphorylated **Alix (Q8WUM4)** at a C-terminal	phosphorylation	phosphorylation
15527798	Phosphorylation of **MDM2 (Q00987)** by the protein kinase **AKT (P31749)**	phosphorylation	phosphorylation
10864201	radiation-induced phosphorylation of **p53 (P04637)** protein at serine 15, largely mediated by **ATM (Q13315)** kinase	phosphorylation	phosphorylation
24548923	**Akt (P31749)**-mediated phosphorylation of Carma1 **(Q9BXL7)**	phosphorylation	phosphorylation
19407811	**BubR1 (O60566)** forms a complex with **PCAF (Q92831)** and is acetylated at lysine 250	acetylation	acetylation
	Test		
21920476	**coilin (P38432)** is phosphorylated in Ser184 by both VRK1 (Q99986)	phosphorylation	phosphorylation
19424295	Previous studies of **cofilin (P23528)** have shown that it is phosphorylated primarily by the LIM domain kinases **Limk1 (P53667)**	phosphorylation	phosphorylation
22726438	**FGFR2 (P21802)** phosphorylates tyrosine residues on **Grb2 (P62993)**	phosphorylation	phosphorylation
11154276	**Akt (P31749)** decreased **ASK1 (Q99683)** kinase activity stimulated by both oxidative stress and overexpression in 293 cells by phosphorylating a consensus **Akt (P31749)** site at serine 83 of **ASK1 (Q99683)**	phosphorylation	phosphorylation
25605758	phosphorylation of **Rab5b (P61020)** by LRRK2 **(Q5S007)** also exhibited	phosphorylation	phosphorylation
20856200	binding of **AKT (P31749)** (tail region) to **Vim (P08670)** (head region) results in **Vim (P08670)** Ser39 phosphorylation	phosphorylation	phosphorylation
21986944	**MAK (P20794)** associates with **CDH1 (P12830)** (FZR1 (Q9UM11), fizzy/cell division cycle 20 related 1) and phosphorylates **CDH1 (P12830)**	Negative	phosphorylation
15862297	**HDM2 (Q00987)** phosphorylation by Chk2 (O96017) doubles in the presence of **p53 (P04637)**	Negative	phosphorylation
21887822	**Hsp70 (P34932)** is phosphorylated by **Plk1 (P53350)**	Negative	phosphorylation

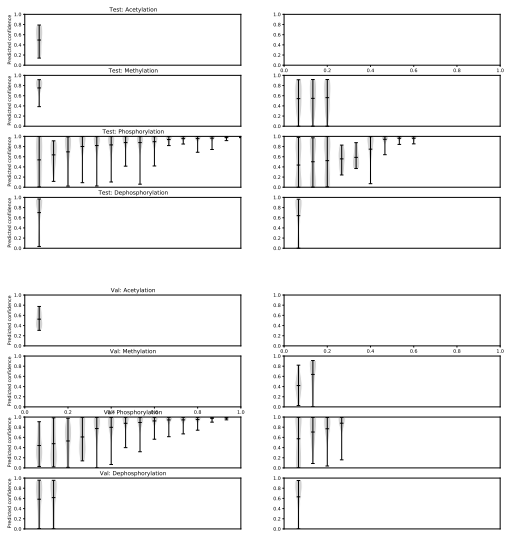

Figure 3 Confidence scores predicted by each model within the ensemble for each sample on the test and validation set using PPI-BioBERT-x10 where **left)** shows the standard deviation of correct predictions and **right)** standard deviation of incorrect predictions. Incorrect predictions do not have low variation in the confidence score except for the 3 test samples in phosphorylation which on manual verification are in fact correct.

Large scale PTM-PPI extraction from PubMed abstracts

We extracted approximately 1.6 million PPIs (546507 unique PTM-PPI triplets) from 18 million abstracts using the PPI-BioBERT-x10 model, see Table 5. The PTM-wise prediction confidence range during large scale extraction is shown in Figure 4, and confidence range is similar to training confidence ranges, where PTMs with lower training samples have lower confidence ranges. After applying interaction-wise probability and confidence standard deviation thresholds, we retain 5708 PPIs across 5 interaction types. From the 5708 predictions (4584 unique PTM-PPI triplets), we randomly select 30 predictions per PTM type for human verification. Despite the high precision (100.0%; see Table 4) on the test set after applying thresholds we find that during the large scale prediction the precision on the randomly sampled subset (subset in supplementary Table A14) drops to 33.7 as shown in Table 6. With the estimated precision of 33.7 %, of the 5708 ≈1900 have the potential to be correct PTM-PPI triplets.

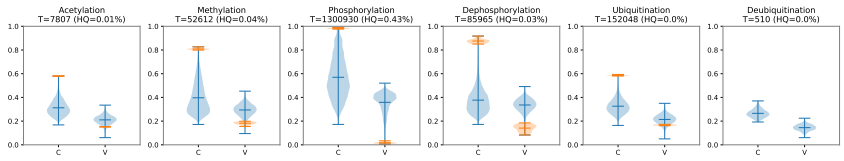

Figure 4 Distribution of PPI-BioBERT-x10 model prediction confidence (C) and confidence standard deviation(V) during large scale prediction. T is the total number of predictions and, HQ is the percentage of high quality predictions (the orange region). High quality predictions are selected using a threshold of both high confidence and low standard deviation based on the train set confidence distribution.

Table 5 The results of large scale prediction. **All** predictions indicate all the predictions from PPI-BioBERT-x10. **U**nique represents unique PTM-PPI triplet predictions. **HQ** represents high quality PTM-PPI after thresholding. **HQ MA** represents high quality PTM-PPI available in multiple abstracts.

PTM	All	All (U)	HQ	HQ (U)	HQ MA	HQ MA (U)
acetylation	7807	6113	1	1	0	0
dephosphorylation	85965	50004	29	29	1	1
deubiquitination	510	460	0	0	0	0
methylation	52612	29914	20	18	4	2
phosphorylation	1300930	381157	5654	4532	1659	537
ubiquitination	152048	78859	4	4	0	0
Total	1599872	546507	5708	4584	1664	540

We find that percentage of precision error per PTM does not seem to be correlated to the number of training samples. For instance, methylation only has 10 training samples but has the highest precision (57%, 11 out of 19), whereas phosphorylation has the highest number of training samples but only 20% (6 out of 30) are correct.

Table 6 Human evaluation on randomly sampled subset (30 interactions per PTM, unless there are fewer predictions) selected after thresholding on average confidence and standard deviation.

	acety.	dephosph.	methy.	phosph.	ubiquit.	Total
Correct	0	11	11	6	0	28
Incorrect - DNA Methylation	0	0	2	0	0	2
Incorrect - NER	0	2	1	3	0	6
Incorrect - No trigger word	0	1	0	2	4	7
Incorrect - Opposite type	0	1	0	0	0	1
Incorrect - Relationship not described	0	14	4	19	0	37
Not - sure	1	0	1	0	0	2
Total	1	29	19	30	4	83

We analyse the incorrect predictions further and categories them into:

- **Incorrect - DNA Methylation:** The abstract describes DNA methylation instead of protein methylation. This type of error mainly affects methylation. For instance from the input abstract PubMed 19386523, the prediction [Q01196 (RUNX1), Methylation , Q06455(ETO)] whereas the abstract describes *"RUNX1-ETO fusion gene on DNA methylation"*
- **Incorrect - NER:** NER has either not identified proteins mentions or not normalised them correctly.
- **Incorrect - No trigger word:** The abstract does not even mention the trigger word describing the predicted PTM. For instance,the prediction [P48431 (Sox2), Ubiquitination, Q01860 (Oct3/4)] where the input abstract PubMed 22732500 does not mention the trigger *ubiquit.*, but the abstract mainly describes *"knockdown of Sox2 and Oct3/4 gene expression in HCC cells can reduce carboplatin-mediated increases in sphere formation and increase cellular sensitivity to chemotherapy"*.
- **Incorrect - Opposite PTM:** The abstract mentions the complementary PTM, *e.g.* (phosphorylation, dephosyphorylation), (ubiquitination, deubiquitinaion). For instance, the prediction [Q15746 (myosin light chain kinase), dephosyphorylation, Q7Z406 (myosin)] from the abstract PubMed 2967285, the abstract describes phosphorylation - *"phosphorylation of the dephosphorylated brain myosin with myosin light chain kinase and casein kinase II"*
- **Incorrect - Relationship not described:** This is effectively any other type of error that doesn't fall into any of the above categories. This category has the largest percentage of error (70%, 37 out of 53 incorrect predictions). An example prediction in this category is [P0870 (IL3), Phosphorylation, P36888 (FLT3)] from the PubMed abstract 10720129 doesn't describe phosphorylation PTM between IL3 and FLT3 but rather *"Somatic mutation of the FLT3 gene, in which the juxtamembrane domain has an internal tandem duplica-*

tion, is found in 20% of human acute myeloid leukemias and causes constitutive tyrosine phosphorylation of the products. In this study, we observed that the transfection of mutant FLT3 gene into an IL3-dependent murine cell line, 32D, abrogated the IL3-dependency. "

- **Incorrect - Not related to PPI:** Here prediction is from an abstract that is not even related to PPI. This is the worst form of false positive error. Example of this is a prediction [P16070 (CD44), phosphorylation, P60568(IL2)] from PubMed abstract 7763733.
- **Unsure:** In this scenario, the human reviewer is unable to decide if the abstract describes the PTM-PPI by solely reading the abstract.Example of this is a prediction [P01019 (Angiotensin II), phosphorylation, P30556(AT1)] from PubMed abstract 9347311.

In order to reduce the false positive predictions, we further filter the high confidence and low variation predictions to a subset of predictions that are available in at least 2 abstracts to select 1659 predictions (537 unique PTM-PPI) and verify a randomly selected subset (30 per PTM) of predictions. We find that the precision on a randomly sampled subset (available as supplementary Table A13) now improves to 58.8% as shown in Table 7. The intuition behind this improvement is if the same PTM-PPI prediction can be inferred from multiple papers, the probability of the prediction being right is higher.

Table 7 Include multiple abstracts filter: Human evaluation on randomly sampled subset selected after thresholding on average confidence and standard deviation, with the additional condition that these predictions are present in multiple abstracts. We select 30 interactions per PTM, unless there are fewer predictions.

	methyl.	phosphoryl.	Total
Correct	4	16	20
Incorrect - NER	0	2	2
Incorrect - Not related to PPI	0	1	1
Incorrect - relationship not described	0	7	7
Not - sure	0	4	4
Total	4	30	34

We also compare the recall of the predictions with those PTM-PPI consolidated in iPTMnet [20][1] as shown in Table 8. Using PPI-BioBERT-x10, we have identified 37.1% (3270 out of 8805) of iPTMnet PTM-PPIs, with 815 of these found in the high confidence region. Only a total of 358 PTM-PPI in iPTMnet were sourced directly from literature using text mining, specifically the rule-based text mining method RLIMS+[17], compared to the 3270 (815 after confidence thresholding) identified by PPI-BioBERT-x10. Hence, the recall of PPI-BioBERT-x10 in relation to the PTMs captured in iPTMnet is substantially higher than the RLIMS+ method, by over 200% after confidence thresholding, demonstrating the robustness of our machine learning-based approach.

Table 8 Comparison of PPI-BioBERT-x10 predictions with iPTMnet. Of all PTM-PPI entries in iPTMnet *(iP Total)*, *iP Unique* represents the subset of unique entries. Of the unique PTM-PPIs the subset that has associated UniProt identfiers is in column *iP Uniprots*. *iP RLIMS* is the number of unique PPI-PTM sourced from RLIMS+ . The number of all the PPI-BioBERT-x10 predictions that can be recalled in iPTMnet is in *Ours*. *Ours HQ* represents the PPI-BioBERT-x10 predictions after confidence thresholding.

PTM	iP Total	iP Unique	iP Uniprots	Ours	Ours HQ	iP RLIMS
acetylation	141	73	12	0	0	0
methylation	7	4	4	0	0	0
phosphorylation	21050	8949	8805	3270	815	358
ubiquitination	2	1	0	0	0	0

Discussion

Noise in distant supervision

There are two types of noise in our distantly supervised dataset: **(a)** false positive noise and **(b)** false negative noise. False positive noise occurs where a given PPI relationship [$Protein1$, *PTM function*, $Protein2$] may not be described in the abstract but is labelled as describing the relationship. False negative noise occurs where the relationship [$Protein1$, *PTM function*, $Protein2$] is in fact described in the abstract but is not labelled as describing the relationship.

Our simple heuristic removes noisy training samples, more specifically false positive noise, *i.e.*, if the normalised Uniprot identifiers of the participating entities

[1] `https://research.bioinformatics.udel.edu/iptmnet_data/files/current/ptm.txt`

Table 9 Results of noise levels, after noise reduction, in training data to verify if the PPI relationship is described in the abstract. The training data is randomly sampled (10 samples per interaction type unless the number of available training samples is lower) and verified by a human annotator.

Interaction Type	Correct	Not - sure	Total
acetylation	4	1	5
dephosphorylation	6	4	10
deubiquitination	1	1	2
methylation	4	6	10
phosphorylation	6	4	10
ubiquitination	2	3	5
Total	23	19	42

are not mentioned in the abstract and the abstract is not likely to describe the relationship (unless NER has failed to detect the protein names and normalise them). We manually inspected the quality of training data after noise reduction using a randomly sampled subset, as detailed in Table 9. While the noise reduction heuristic has been fairly effective in removing false positive noise, in 45% of the cases the human reviewer (author A.E. with basic knowledge on PPIs and PTMs) was unable to decide whether the PPI relationship is described or not in the abstract. An example of a sample marked as unsure, is phosphorylation between c-Jun N-terminal kinase (JNK) and c-Jun (IntAct entry `https://www.ebi.ac.uk/intact/interaction/EBI-7057279`) in the abstract PubMed 19527717. This shows the complexity of the task beyond language, the need for prior context or assumed domain knowledge on how PPIs interact to interpret the abstracts for manual annotation. In the test set, we found that a combination of high confidence and low variation in confidence has been able detect some of the incomplete annotation (false negatives) as shown in Table 4.

One of the main challenges is that it is difficult to quantify noise without manual verification. Hence, machine learning approaches that rely on distant supervised data need to manually verify label quality, at the very least, on randomly subset of the data. This requires a user interface (UI) that makes it easy for humans to review the samples effectively. For this work we built a UI (see Figure 5) using Amazon

SageMaker Ground Truth, to assist in verification of both the labelled and the large scale prediction samples.

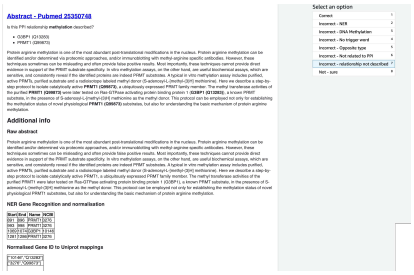

Figure 5 UI built using Amazon SageMaker Ground Truth for verifying PTM-PPI.

Generalisability of PTM-PPI extraction

Generalisability in machine learning is a broadly discussed issue. Futoma *et al.* [23] in their work in a clinical context describe how important it is to understand how, when, and why a machine learning system works. In this paper, we primarily investigated the use of confidence calibration as an indicator of when a prediction is valid. We hypothesised that using the average ensemble confidence, a state-of-the-art uncertainty estimation approach by Lakshminarayanan *et al.* [6], with our extension to also include low variation to counteract the effects of class imbalance (where each PTM has high difference in the number of training samples) should be able to minimise incorrect predictions. While this assumption seems to hold true in the test and validation tests, it does not seem to be the case during large scale application as shown in Table 6. Previous papers on PPI extraction using machine learning [13, 31, 25] have evaluated the efficacy of their method on a test set, and have used the performance on the the test set as an indicator of generalisability. The method used to create a test set and its impact on assessing generalisability has been studied in previous papers [22, 32]. For instance, we [22] have identified data leakage using random splits with 10-fold cross validation on the PPI interaction AIMed dataset where the random split results in over 70% tri-

gram word overlap between train and test folds resulting in inflated performance on the test set. The substantially lower real world performance compared to the test set, specifically in the context of BERT models, has also been reported in previous papers [32]. In addition, previous studies in computer vision have also shown neural networks can provide high confidence prediction even when the model is wrong despite robust performance on the test set [7, 33]. In particular a type of model uncertainty, epistemic uncertainty [33], where the model doesn't have sufficient information to make a decision yet has made overconfident predictions continues to be a challenge despite applying machine learning methods. More specifically, in the context of this paper the need for diverse and effective negative samples is pertinent, given the number of false positive predictions despite relatively high confidence. The creation of effective negative samples and generalisability remain open research questions in need of further study.

Human augmentation of PTM-PPI extraction

The prediction quality can be improved by providing more training samples [34, 33], which requires manual curation. However, manual curation is difficult, time consuming and not cost effective [4, 3]. This becomes a chicken and egg problem: automation is meant to speed up curation and reduce manual curation effort, however we need large amounts of manually curated training data so that the models can produce sufficiently reliable predictions.

Triaging journals alone takes up approximately 15% of the curation time [35] to select relevant papers, while the rest of the tasks such as named entity recognition and normalisation, detecting relationships take up the rest 85%. In our research, we surface PTM-PPI triplets where the entities are normalised. Hence, our solution supports human augmentation [36], *i.e.* using machine learning to make it much easier for humans to curate. This is one of the major advantages of machine learning compared to rule based systems, as adding more data and retraining can improve

the model without manually rewriting the rules. The second advantage of using confidence thresholding is the ability to adjust the thresholds per PTM enabling faster curation of PTM-PPIs with reduced representation, such as for relatively rare interaction types like deubiquitination. In order to take advantage of machine learning, we need to ensure human annotators have: **(a)** convenient user interface so the curator can quickly accept or reject the prediction and **(b)** provide a high hit ratio of correct PPIs using confidence thresholding so that humans can add more PPIs to the knowledge bases faster. This will enable continuous retraining the model with more diverse training data (both positive and negative samples) so the model can evolve and further reduce manual curation effort over time.

Conclusion

We created a distant supervised training dataset for extracting PTM-PPIs, including annotation for phosphorylation, dephosphorylation, methylation, demethylation, ubiquitination, and acetylation from PubMed abstracts by leveraging the IntAct database.

Using this dataset we trained an ensemble model, PPI-BioBERT-x10, and applied an approach by Lakshminarayanan *et al.* [6] to improve confidence calibration of the neural network by averaging the confidence scores predicted by the ensemble. We find that the predicted confidence range for each PTM is proportional to the number of training samples available and the predicted confidence range is consistent between train, test validation and large scale predictions. As a result, we extended the work by Lakshminarayanan *et al.* [6] using a threshold per PTM by combining average confidence with confidence standard deviation to improve confidence calibration and counteract the effects of class imbalance during calibration, resulting in 100% precision (retaining 19% of the positive predictions) in the test set.

We applied PPI-BioBERT-x10 to text-mine 18 million PubMed abstracts, extracting 1.6 million PTM-PPI predictions ($\approx$ 540,000 unique PTM-PPI) and shortlisting

approximately 5,700 (≈ 4,500 unique PTM-PPI) PTM-PPIs with relative high confidence and low variation. However, on manual review of a randomly sampled subset we find that the precision of prediction drops to 33.7% and generalisability aspect of knowing when a prediction is correct remains a challenge. By selecting predictions that appear in multiple abstracts (≈ 1600 triplets), the precision on a randomly sampled subset improves to 58.8%. We also find that PPI-BioBERT-x10 is able to identify over 200% more PTM-PPIs compared to RLIMS+[17] a rule based system, revealing the advantage of PPI-BioBERT-x10 with regards to recall compared to a manually crafted rule based system. We also propose that to decrease manual curation effort and to improve the prediction quality over time requires using an effective user interface along with confidence thresholding to allow humans curators to easily accept or reject predictions and continuously retrain the model.

In this work, we studied the advantages and challenges of adopting deep learning-based text mining for a new task, PTM-PPI triplet extraction, using distantly supervised data. This work highlights the need for effective confidence calibration, the importance of generalisability beyond test set performance and how real world performance can vary substantially compared to the test set. Further study includes the use of effective adversarial samples to improve the robustness of machine learning models for practical use.

Appendix

A.1 BioBERT Training details

We use the following settings to train BioBERT v1.1 based on BERT Base cased using PyTorch 1.4.0. We use gradient accumulation to train with an effective batch size of 64. We train on a Amazon SageMaker P3 instance using a single GPU for approximately 1 hours with early stopping patience.

- Loss function - cross entropy loss
- Optimiser - Adam

- Optimiser Learning rate - .00001,
- Optimiser Weight decay - 0.01,
- Batch size - 8
- Gradient accumulation steps - 8
- Epochs - 1000
- Early stopping patience epochs - 20

The full source code is here `https://github.com/elangovana/large-scale-ptm-ppi`

A.2 BioBERT Token lengths

Here we present the distribution of length of the tokens, as a result of using BERT tokeniser, to tokenise the abstract, see Table A10.

Table A10 The distribution of abstract lengths as a result of using BioBERT tokeniser. Approximately 90% of the unique normalised abstracts is under the max limit of 512.

	Train	Test	Validation
count	1305.00	352.00	200.00
mean	378.21	353.21	375.32
std	95.53	102.59	78.16
min	174.00	175.00	194.00
0%	174.00	175.00	194.00
10%	256.00	209.20	241.90
20%	285.00	250.20	323.00
30%	323.00	296.30	335.00
40%	344.60	323.00	361.80
50%	377.00	348.00	379.00
60%	402.00	376.00	389.00
70%	426.00	414.70	420.90
80%	464.00	447.00	467.00
90%	511.00	496.90	472.00
max	612.00	538.00	496.00

A.3 Performance using lowest loss

The performance on test and validation set where the models in the ensemble are selected based on lowest loss on the validation set. The Table A11 shows the F-score, the reliability diagrams are in Figure A6 and the Figure A7 shows the corresponding confusion matrix.

Table A11 The performance of ensemble PPI-BioBERT-x10 on the test and validation set where the models are selected based on lowest loss on validation set.

	Test				Validation			
InteractionType	Precision	Recall	F1-score	support	Precision	Recall	F1-score	support
acetylation	0.00	0.00	0.00	1	0.00	0.00	0.00	1
dephosphorylation	0.00	0.00	0.00	6	0.00	0.00	0.00	10
methylation	0.00	0.00	0.00	4	0.00	0.00	0.00	1
phosphorylation	66.67	36.36	47.06	44	56.25	42.86	48.65	21
ubiquitination	0.00	0.00	0.00	1	0.00	0.00	0.00	1
macro avg	13.33	7.27	9.41	56	11.25	8.57	9.73	34
micro avg	66.67	28.57	40.00	56	56.25	26.47	36.00	34

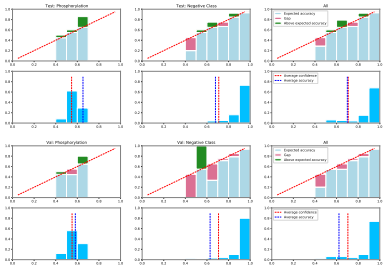

Figure A6 Confidence reliability on the test and validation set using PPI-BioBERT-x10 where the models are selected based on lowest loss on validation set.

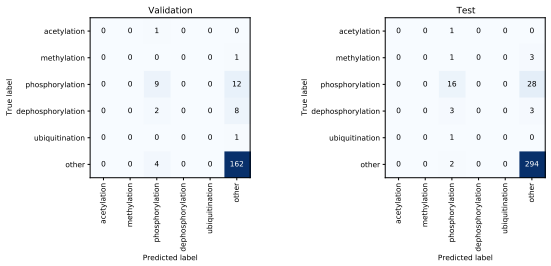

Figure A7 Confusion matrix using PPI-BioBERT-x10 where the models are selected based on lowest loss on validation set.

A.4 Human verified results

Acknowledgements

Funding

This work is supported by Australian Research Council grants DP190101350 (KV) and LP160101469 (KV,YL).

Abbreviations

NER: Named entity recognition

PPI: Protein Protein Interaction

Table A12 The distribution of unique normalised protein counts in the abstracts in the test set.

Interaction	count	mean	std	min	25%	50%	75%	max
acetylation	6	3.50	1.38	2.00	3.00	3.00	3.75	6.00
demethylation	2	5.00	0.00	5.00	5.00	5.00	5.00	5.00
dephosphorylation	21	3.62	1.69	2.00	2.00	3.00	5.00	6.00
deubiquitination	1	2.00	nan	2.00	2.00	2.00	2.00	2.00
methylation	11	2.82	0.40	2.00	3.00	3.00	3.00	3.00
phosphorylation	125	3.56	1.95	1.00	2.00	3.00	5.00	9.00
ubiquitination	2	5.00	1.41	4.00	4.50	5.00	5.50	6.00

Table A13 Human verification of randomly sampled subset of 30 PTM-PPI per PTM type. The samples were chosen after confidence calibration and the predictions were present in multiple abstracts within the confidence and standard deviation thresholds.

PubmedId	Participant1 Uniprot	Participant1 Name	Participant2 Uniprot	Participant2 Name	Prediction	Human result
19877273	P23443	S6K1	P42345	mTOR	phosphorylation	Correct
28539327	P01375	TNF	Q13546	RIPK1	phosphorylation	Incorrect - relationship not described
9148953	P06493	Cdc2	P30291	Wee1	phosphorylation	Correct
18028023	P31749	Akt	P42345	mammalian target of rapamycin	phosphorylation	Incorrect - relationship not described
14572154	P31749	PKB/Akt	P60484	PTEN	phosphorylation	Not - sure
1327869	O00757	FBPase-2	Q16875	PFK-2	phosphorylation	Incorrect - relationship not described
22855742	O14757	Checkpoint kinase 1	Q13535	ATR	phosphorylation	Correct
16477614	P23560	Brain-derived neurotrophic factor	P28482	extracellular-signal regulated kinase	phosphorylation	Incorrect - relationship not described
15289331	P31749	Akt	P49841	glycogen synthase kinase-3beta	phosphorylation	Correct
23399841	P01583	IL-1	P10145	CXCL8	phosphorylation	Not - sure
16873552	P12931	Src	P41240	C-terminal Src kinase	phosphorylation	Correct
15994312	Q05397	focal adhesion kinase	Q14289	proline-rich tyrosine kinase-2	phosphorylation	Correct
17190911	P35568	insulin receptor substrate-1	P42336	PI3-K	phosphorylation	Correct
1840422	P17677	B-50	Q92686	neurogranin	phosphorylation	Incorrect - NER
27633668	P20042	eukaryotic translation initiation factor 2	Q9BQI3	heme-regulated inhibitor	phosphorylation	Incorrect - relationship not described
12112022	P31749	PKBalpha	Q14289	Protein kinase B	phosphorylation	Incorrect - relationship not described
23337506	P01138	nerve growth factor	P04629	tropomyosin receptor kinase receptor	phosphorylation	Correct
12759443	O14920	IKK beta	Q04206	p65	phosphorylation	Not - sure
29162743	P16220	CREB	P40763	signal transducer and activator of transcripti...	phosphorylation	Incorrect - NER
9347311	P01019	Angiotensin II	P30556	AT1	phosphorylation	Not - sure
21325496	P24941	cyclin-dependent kinase 2	P38936	p21	phosphorylation	Correct
7763733	P16070	CD44	P60568	IL2	phosphorylation	Incorrect - Not related to PPI
26659448	Q02763	Tek	Q15389	angiopoietin-1	phosphorylation	Correct
20110283	P17676	CCAAT/enhancer binding protein beta	P49841	glycogen synthase kinase 3beta	phosphorylation	Correct
15970650	Q15746	myosin light chain kinase	Q7Z406	myosin	phosphorylation	Correct
29496905	P23443	S6K1)	P35568	insulin receptor substrate-1	phosphorylation	Correct
17402366	P01562	IFN-alpha	P42224	STAT1	phosphorylation	Correct
18198129	P12931	Src	Q05397	FAK	phosphorylation	Correct
17053882	P31749	Akt	P42336	phosphatidylinositol 3-kinase	phosphorylation	Incorrect - relationship not described
24360952	P12931	Src	Q05397	Focal adhesion kinase	phosphorylation	Correct
27411844	P04637	p53	Q9NQR1	SETD8	methylation	Correct
20614940	P04637	p53	Q96KQ7	G9a	methylation	Correct
24151879	P04637	p53	Q96KQ7	G9a	methylation	Correct
25554733	P04637	p53	Q9NQR1	SETD8	methylation	Correct

PTM: Post Translational Modification

ECE: Expected Calibration Error

NLP: Natural Language Processing

BERT: Bidirectional Encoder Representations from Transformers

Availability of data and materials

We publish our source code, training, test and validation dataset and the predicted ≈5700 PPIs on

`https://github.com/elangovana/large-scale-ptm-ppi`.

The Amazon SageMaker Ground Truth UI code is available at

`https://github.com/elangovana/ppi-sagemaker-groundtruth-verification`.

Ethics approval and consent to participate

Not applicable

Competing interests

We certify that there are no financial and non-financial competing interests regarding the materials discussed in the manuscript.

Consent for publication

Not applicable

Authors' contributions

AE carried out computational experiments and was responsible for the design, data analysis, data interpretation, developing software, discussion and drafting and editing the manuscript. AE, MD, KV conceptualised the study. MD, YL, DP and KV supervised AE, were involved in the design and execution of the study, supported data interpretation, and contributed to writing the manuscript. All authors read and approved the final manuscript.

Author details

[1]School of Computing and Information Systems, The University of Melbourne, Melbourne, Australia. [2]The Walter and Eliza Hall Institute of Medical Research, Melbourne, Australia. [3]Department of Clinical Pathology, Faculty of Medicine, Dentistry & Health Sciences The University of Melbourne, Melbourne, Australia. [4]School of Computing Technologies RMIT University, Melbourne, Australia.

References

1. Orchard, S., Ammari, M., Aranda, B., Breuza, L., Briganti, L., Broackes-Carter, F., Campbell, N.H., Chavali, G., Chen, C., Del-Toro, N.: The mintact project—intact as a common curation platform for 11 molecular interaction databases. Nucleic Acids Research **42**(D1), 358–363 (2013)
2. Mishra, G.R., Suresh, M., Kumaran, K., Kannabiran, N., Suresh, S., Bala, P., Shivakumar, K., Anuradha, N., Reddy, R., Raghavan, T.M.: Human protein reference database—2006 update. Nucleic Acids Research **34**, 411–414 (2006)
3. Orchard, S., Ammari, M., Aranda, B., Breuza, L., Briganti, L., Broackes-Carter, F., Campbell, N.H., Chavali, G., Chen, C., del-Toro, N., Duesbury, M., Dumousseau, M., Galeota, E., Hinz, U., Iannuccelli, M., Jagannathan, S., Jimenez, R., Khadake, J., Lagreid, A., Licata, L., Lovering, R.C., Meldal, B., Melidoni, A.N., Milagros, M., Peluso, D., Perfetto, L., Porras, P., Raghunath, A., Ricard-Blum, S., Roechert, B., Stutz, A., Tognolli, M., van Roey, K., Cesareni, G., Hermjakob, H.: The MIntAct project—IntAct as a common curation platform for 11 molecular interaction databases. Nucleic Acids Research **42**(D1), 358–363 (2013). doi:10.1093/nar/gkt1115. https://academic.oup.com/nar/article-pdf/42/D1/D358/3585170/gkt1115.pdf
4. Orchard, S., Hermjakob, H.: Shared resources, shared costs—leveraging biocuration resources. Database **2015** (2015). doi:10.1093/database/bav009. bav009. https://academic.oup.com/database/article-pdf/doi/10.1093/database/bav009/7298544/bav009.pdf
5. Consortium, T.U.: UniProt: a worldwide hub of protein knowledge. Nucleic Acids Research **47**(D1), 506–515 (2018). doi:10.1093/nar/gky1049. https://academic.oup.com/nar/article-pdf/47/D1/D506/27437297/gky1049.pdf
6. Lakshminarayanan, B., Pritzel, A., Blundell, C.: Simple and scalable predictive uncertainty estimation using deep ensembles. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems, vol. 30, pp. 6402–6413. Curran Associates, Inc., ??? (2017). https://proceedings.neurips.cc/paper/2017/file/9ef2ed4b7fd2c810847ffa5fa85bce38-Paper.pdf

7. Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: Proceedings of the 34th International Conference on Machine Learning - Volume 70. ICML'17, pp. 1321–1330. JMLR.org, ??? (2017)
8. Craven, M., Kumlien, J.: Constructing biological knowledge bases by extracting information from text sources. In: Proceedings of the Seventh International Conference on Intelligent Systems for Molecular Biology, pp. 77–86. AAAI Press, ??? (1999)
9. Mintz, M., Bills, S., Snow, R., Jurafsky, D.: Distant supervision for relation extraction without labeled data. In: Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP, pp. 1003–1011. Association for Computational Linguistics, Suntec, Singapore (2009). https://www.aclweb.org/anthology/P09-1113
10. Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C.H., Kang, J.: Biobert: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics **36**(4), 1234–1240 (2019). doi:10.1093/bioinformatics/btz682
11. Bunescu, R., Ge, R., Kate, R.J., Marcotte, E.M., Mooney, R.J., Ramani, A.K., Wong, Y.W.: Comparative experiments on learning information extractors for proteins and their interactions. Artificial Intelligence in Medicine **33**(2), 139–155 (2005)
12. Pyysalo, S., Ginter, F., Heimonen, J., Björne, J., Boberg, J., Järvinen, J., Salakoski, T.: Bioinfer: a corpus for information extraction in the biomedical domain. BMC Bioinformatics **8**(1), 50 (2007)
13. Hsieh, Y.-L., Chang, Y.-C., Chang, N.-W., Hsu, W.-L.: Identifying protein-protein interactions in biomedical literature using recurrent neural networks with long short-term memory. In: Proceedings of the Eighth International Joint Conference on Natural Language Processing (volume 2: Short Papers), pp. 240–245
14. Peng, Y., Lu, Z.: Deep learning for extracting protein-protein interactions from biomedical literature. In: BioNLP 2017, pp. 29–38. Association for Computational Linguistics. `https://www.aclweb.org/anthology/W17-2304http://dx.doi.org/10.18653/v1/W17-2304`
15. Franceschini, A., Szklarczyk, D., Frankild, S., Kuhn, M., Simonovic, M., Roth, A., Lin, J., Minguez, P., Bork, P., von Mering, C., Jensen, L.J.: STRING v9.1: protein-protein interaction networks, with increased coverage and integration. Nucleic Acids Research **41**(D1), 808–815 (2012). doi:10.1093/nar/gks1094. https://academic.oup.com/nar/article-pdf/41/D1/D808/3617210/gks1094.pdf
16. Tudor, C.O., Ross, K.E., Li, G., Vijay-Shanker, K., Wu, C.H., Arighi, C.N.: Construction of phosphorylation interaction networks by text mining of full-length articles using the eFIP system. Database **2015** (2015). doi:10.1093/database/bav020. bav020. https://academic.oup.com/database/article-pdf/doi/10.1093/database/bav020/16975967/bav020.pdf
17. Torii, M., Arighi, C.N., Li, G., Wang, Q., Wu, C.H., Vijay-Shanker, K.: Rlims-p 2.0: A generalizable rule-based information extraction system for literature mining of protein phosphorylation information **12**(1), 17–29 (2015). doi:10.1109/TCBB.2014.2372765
18. Szklarczyk, D., Gable, A.L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., Simonovic, M., Doncheva, N.T., Morris, J.H., Bork, P., Jensen, L.J., Mering, C.: STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Research **47**(D1), 607–613 (2018). doi:10.1093/nar/gky1131. https://academic.oup.com/nar/article-pdf/47/D1/D607/27437323/gky1131.pdf
19. Chen, Q., Panyam, N.C., Elangovan, A., Verspoor, K.: BioCreative VI Precision Medicine Track system performance is constrained by entity recognition and variations in corpus characteristics. Database **2018** (2018). doi:10.1093/database/bay122. bay122.

https://academic.oup.com/database/article-pdf/doi/10.1093/database/bay122/27329364/bay122.pdf

20. Huang, H., Arighi, C.N., Ross, K.E., Ren, J., Li, G., Chen, S.-C., Wang, Q., Cowart, J., Vijay-Shanker, K., Wu, C.H.: iPTMnet: an integrated resource for protein post-translational modification network discovery. Nucleic Acids Research **46**(D1), 542–550 (2017). doi:10.1093/nar/gkx1104. https://academic.oup.com/nar/article-pdf/46/D1/D542/23162331/gkx1104.pdf
21. Hornbeck, P.V., Zhang, B., Murray, B., Kornhauser, J.M., Latham, V., Skrzypek, E.: PhosphoSitePlus, 2014: mutations, PTMs and recalibrations. Nucleic Acids Research **43**(D1), 512–520 (2014). doi:10.1093/nar/gku1267. https://academic.oup.com/nar/article-pdf/43/D1/D512/17437800/gku1267.pdf
22. Elangovan, A., He, J., Verspoor, K.: Memorization vs. generalization : Quantifying data leakage in NLP performance evaluation. In: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, pp. 1325–1335. Association for Computational Linguistics, Online (2021). https://www.aclweb.org/anthology/2021.eacl-main.113
23. Futoma, J., Simons, M., Panch, T., Doshi-Velez, F., Celi, L.A.: The myth of generalisability in clinical research and machine learning in health care. The Lancet Digital Health **2**(9), 489–492 (2020). doi:10.1016/S2589-7500(20)30186-2
24. Li, G., Wu, C., Vijay-Shanker, K.: Noise reduction methods for distantly supervised biomedical relation extraction. In: BioNLP 2017, pp. 184–193. Association for Computational Linguistics, Vancouver, Canada, (2017). doi:10.18653/v1/W17-2323. https://www.aclweb.org/anthology/W17-2323
25. Zhang, H., Guan, R., Zhou, F., Liang, Y., Zhan, Z.-H., Huang, L., Feng, X.: Deep residual convolutional neural network for protein-protein interaction extraction. IEEE Access **7**, 89354–89365 (2019). doi:10.1109/ACCESS.2019.2927253
26. Orchard, S., Kerrien, S., Abbani, S., Aranda, B., Bhate, J., Bidwell, S., Bridge, A., Briganti, L., Brinkman, F.S., Cesareni, G.: Protein interaction data curation: the international molecular exchange (imex) consortium. Nature Methods **9**(4), 345 (2012)
27. Wei, C.-H., Kao, H.-Y., Lu, Z.: Gnormplus: an integrative approach for tagging genes, gene families, and protein domains. BioMed Research International **2015** (2015)
28. Brown, G.R., Hem, V., Katz, K.S., Ovetsky, M., Wallin, C., Ermolaeva, O., Tolstoy, I., Tatusova, T., Pruitt, K.D., Maglott, D.R., Murphy, T.D.: Gene: a gene-centered information resource at NCBI. Nucleic Acids Research **43**(D1), 36–42 (2014). doi:10.1093/nar/gku1055. https://academic.oup.com/nar/article-pdf/43/D1/D36/7314925/gku1055.pdf
29. Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pp. 4171–4186. Association for Computational Linguistics. https://www.aclweb.org/anthology/N19-1423http://dx.doi.org/10.18653/v1/N19-1423
30. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: A simple way to prevent neural networks from overfitting. Journal of Machine Learning Research **15**(56), 1929–1958 (2014)
31. Yadav, S., Ekbal, A., Saha, S., Kumar, A., Bhattacharyya, P.: Feature assisted stacked attentive shortest dependency path based bi-lstm model for protein-protein interaction. Knowledge-Based Systems (2018)
32. McCoy, R.T., Min, J., Linzen, T.: BERTs of a feather do not generalize together: Large variability in generalization across models with similar test set performance. In: Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP, pp. 217–227. Association for Computational Linguistics, Online (2020). doi:10.18653/v1/2020.blackboxnlp-1.21.

https://aclanthology.org/2020.blackboxnlp-1.21

33. Hüllermeier, E., Waegeman, W.: Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. Machine Learning **110**(3), 457–506 (2021)
34. Kendall, A., Gal, Y.: What uncertainties do we need in bayesian deep learning for computer vision? In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems, vol. 30. Curran Associates, Inc., ??? (2017). https://proceedings.neurips.cc/paper/2017/file/2650d6089a6d640c5e85b2b88265dc2b-Paper.pdf
35. Mottin, L., Pasche, E., Gobeill, J., Rech de Laval, V., Gleizes, A., Michel, P.-A., Bairoch, A., Gaudet, P., Ruch, P.: Triage by ranking to support the curation of protein interactions. Database **2017** (2017). doi:10.1093/database/bax040. bax040. https://academic.oup.com/database/article-pdf/doi/10.1093/database/bax040/19234641/bax040.pdf
36. Raisamo, R., Rakkolainen, I., Majaranta, P., Salminen, K., Rantala, J., Farooq, A.: Human augmentation: Past, present and future. International Journal of Human-Computer Studies **131**, 131–143 (2019). doi:10.1016/j.ijhcs.2019.05.008. 50 years of the International Journal of Human-Computer Studies. Reflections on the past, present and future of human-centred technologies

Figures

Tables

Additional Files

Additional file 1 — Sample additional file title

Additional file descriptions text (including details of how to view the file, if it is in a non-standard format or the file extension). This might refer to a multi-page table or a figure.

Table A14 Human verification, without taking into account predictions in multiple abstracts, of randomly sampled subset of 30 PTM-PPI per PTM type. The samples were chosen after confidence calibration.

Pubmed Id	Participant1 Uniprot	Participant1 Name	Participant2 Uniprot	Participant2 Name	Prediction	Human result
22732500	P48431	Sox2	Q01860	Oct3/4	ubiquitination	Incorrect - No trigger word
21492879	P48023	Fas ligand	Q8IXH7	Th1	ubiquitination	Incorrect - No trigger word
19428371	O75943	RAD24	Q99638	RAD9	ubiquitination	Incorrect - No trigger word
17168739	P02786	transferrin receptor	P09544	IRP	ubiquitination	Incorrect - No trigger word
7789533	P19338	C23	Q9Y2Q3	glutathione S-transferase	phosphorylation	Incorrect - No trigger word
19729590	P23443	S6 kinase	P35568	insulin receptor substrate 1	phosphorylation	Incorrect - relationship not described
23478265	O00571	DEAD box protein 3	Q9UHD2	TANK-binding kinase 1	phosphorylation	Incorrect - relationship not described
16365045	O75581	low density lipoprotein receptor-related prote...	P49841	GSK3beta	phosphorylation	Incorrect - relationship not described
19306938	O60500	nephrin	Q9Y5K6	CD2-associated protein	phosphorylation	Incorrect - relationship not described
10720129	P08700	IL3	P36888	FLT3	phosphorylation	Incorrect - relationship not described
27456486	P20701	LFA-1	Q05397	Protein tyrosine kinase 2	phosphorylation	Incorrect - relationship not described
2780290	P19338	nucleolin	P50613	Protein kinase	phosphorylation	Incorrect - NER
17108171	P04629	TrkA	P08138	p75 neurotrophin receptor	phosphorylation	Incorrect - relationship not described
10884023	P16220	Ca2+/cAMP responsive element binding protein	P18509	pituitary adenylate cyclase-activating polypep...	phosphorylation	Incorrect - relationship not described
27118568	O14974	myosin phosphatase-targeting subunit 1	P61586	Rho A	phosphorylation	Incorrect - relationship not described
17353368	Q13094	SH2 domain-containing leukocyte protein of 76 kD	Q92918	hematopoietic progenitor kinase 1	phosphorylation	Correct
20122754	P05067	APP	P45983	JNK	phosphorylation	Correct
23023514	P42574	caspase 3	Q16555	Collapsin response mediator protein-2	phosphorylation	Incorrect - relationship not described
7852364	P78356	PtdIns(4)P 5-kinase	Q9Y2I7	Fab1p	phosphorylation	Incorrect - relationship not described
26464283	P00533	EGFR	P35354	cyclooxygenase-2	phosphorylation	Correct
17211494	Q06124	PTPN11	Q13480	Grb2 associated binder 1	phosphorylation	Incorrect - No trigger word
29514920	P30291	Wee1	Q01850	Cdr2	phosphorylation	Correct
17045592	P49841	GSK-3beta	Q00535	cdk5	phosphorylation	Incorrect - relationship not described
22633971	P35548	Msx2	Q13950	Runx2	phosphorylation	Incorrect - relationship not described
9852063	P42681	tyrosine kinase	P63252	Kir2.1	phosphorylation	Correct
19879273	P31749	Akt	P35222	beta-catenin	phosphorylation	Correct
3014310	P04271	S-100	P50613	protein kinase	phosphorylation	Incorrect - relationship not described
9883577	Q15049	myosin light chain	Q32MK0	MLC kinase	phosphorylation	Incorrect - NER
29162743	P16220	CREB	P40763	signal transducer and activator of transcripti...	phosphorylation	Incorrect - NER
11567986	P62993	Grb2	Q92835	SH2-containing inositol 5'-phosphatase	phosphorylation	Incorrect - relationship not described
30158515	Q00535	CDK5	Q15078	p35	phosphorylation	Incorrect - relationship not described
1707345	O00459	p85 beta	P27986	p85 alpha	phosphorylation	Incorrect - relationship not described
25684187	P01588	Erythropoietin	P17302	connexin43	phosphorylation	Incorrect - relationship not described
25446109	P04792	HSP27	P45983	JNK	phosphorylation	Incorrect - relationship not described
21080372	P08670	eukaryotic translation initiation factor 4A-I ...	Q99873	PRMT1	methylation	Correct
26503212	Q8NCA5	FAM98A	Q99873	PRMT1	methylation	Correct
24412544	Q96L73	NSD1	Q9BQQ3	p65	methylation	Incorrect - relationship not described
25749972	O75164	KDM4A	P42345	mTOR	methylation	Not - sure
25554733	P04637	p53	Q9NQR1	SETD8	methylation	Correct
20615470	Q15047	KMT1E	Q8WTS6	KMT7	methylation	Incorrect - relationship not described
24151879	P04637	p53	Q96KQ7	G9a	methylation	Correct
23469257	O15047	Set1	O75934	Dam1	methylation	Correct
18472002	P04637	p53	P61978	hnRNP K	methylation	Correct
19386523	Q01196	RUNX1	Q06455	ETO	methylation	Incorrect - DNA Methylation
20614940	P04637	p53	Q96KQ7	G9a	methylation	Correct
27411844	P04637	p53	Q9NQR1	SETD8	methylation	Correct
20231378	P11274	B cell antigen receptor	P11912	Igalpha	methylation	Incorrect - NER
24129573	P01562	IFN-	Q01628	Interferon-induced transmembrane protein 3	methylation	Incorrect - relationship not described
26391684	P49757	Numb	Q9NQR1	SET8	methylation	Correct
25350748	Q13283	G3BP1	Q99873	PRMT1	methylation	Correct
26126536	Q99873	PRMT1	Q9HAU4	Smurf2	methylation	Correct
27563394	P35270	SPR	Q8TEK3	disruptor of telomeric silencing 1-like protein	methylation	Incorrect - DNA Methylation
25748791	Q8WWM7	ataxin-2-like	Q99700	ataxin-2	methylation	Incorrect - relationship not described
2967285	Q15746	myosin light chain kinase	Q7Z406	myosin	dephosphorylation	Incorrect - Opposite type
9417127	P01375	tumor necrosis factor-alpha	P29350	SHP-1	dephosphorylation	Incorrect - relationship not described
6245872	P10145	NaF	P50613	protein kinase	dephosphorylation	Incorrect - NER
11264356	Q15181	protein phosphatase 1	Q9BY44	eIF-2alpha	dephosphorylation	Correct
9079814	P35670	WC1	P60568	interleukin (IL)-2	dephosphorylation	Incorrect - relationship not described
8083198	P18031	PTP-1B	P50391	PP-1	dephosphorylation	Incorrect - relationship not described
12853978	P00533	epidermal growth factor receptor	P42574	caspase-3	dephosphorylation	Incorrect - relationship not described
12138178	P42224	Stat1	P42226	Stat6	dephosphorylation	Incorrect - relationship not described
10391142	P01308	insulin	P06213	insulin receptor	dephosphorylation	Incorrect - relationship not described
9119896	P49840	glycogen synthase kinase-3alpha (kinase FA/GSK...	Q9Y2R2	protein tyrosine phosphatase	dephosphorylation	Incorrect - NER
16620785	P31749	Akt	P49840	glycogen synthase kinase-3alpha/beta	dephosphorylation	Correct
7683660	P01308	Insulin	P19338	nucleolin	dephosphorylation	Incorrect - relationship not described
10405762	O14494	PAP2	O43688	hPAP2c	dephosphorylation	Incorrect - relationship not described
10405762	O14494	PAP2	O14495	hPAP2b	dephosphorylation	Incorrect - relationship not described
10496881	P06493	p34cdc2	P30307	cdc25-C	dephosphorylation	Correct
11773439	P17706	TC-PTP	P42229	signal transducer and activator of transcripti...	dephosphorylation	Correct
15823043	P19634	Na(+)/H(+) exchanger isoform 1	Q15181	PP1	dephosphorylation	Correct
15269244	P06213	insulin receptor	P18031	protein-tyrosine phosphatase (PTP) 1B	dephosphorylation	Correct
11495355	P21128	P11	P60903	P10	dephosphorylation	Incorrect - relationship not described
11707519	P16220	cAMP response element-binding protein	P27361	extracellular-signal related kinases 1 and 2	dephosphorylation	Incorrect - No trigger word
1321126	P06213	insulin receptor	P10586	LAR	dephosphorylation	Correct
9989818	P04637	p53	P28749	p107/E2F	dephosphorylation	Incorrect - relationship not described
17046078	P12931	Src	P29350	SHP1	dephosphorylation	Incorrect - relationship not described
9380758	P13569	cystic fibrosis transmembrane conductance regu...	P35813	PP2Calpha	dephosphorylation	Correct
24691491	P20042	eukaryotic translation initiation factor 2	P63000	Rac1	dephosphorylation	Incorrect - relationship not described
11495355	P21128	P11	Q15257	PP2A	dephosphorylation	Incorrect - relationship not described
9989818	P28749	p107/E2F	Q15257	PP2A	dephosphorylation	Correct
11222372	Q15257	PP2A	Q92934	BAD	dephosphorylation	Correct
8550570	Q15181	PP-1	Q15257	phosphatase-2A	dephosphorylation	Correct
14731392	P25440	Brd2	Q92831	PCAF	acetylation	Not - sure