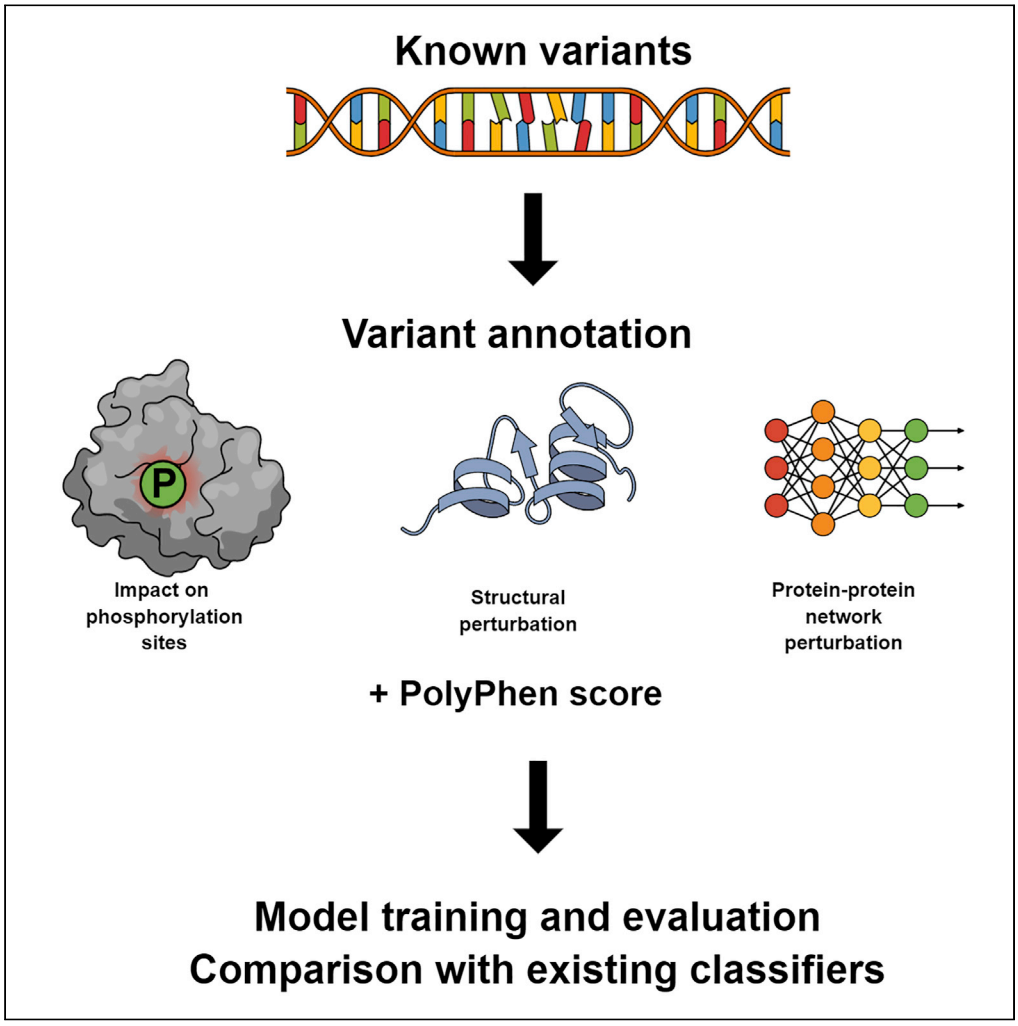
iScience

CellPress
OPEN ACCESS

Article

PhosphoEffect: Prioritizing Variants On or Adjacent to Phosphorylation Sites through Their Effect on Kinase Recognition Motifs



Stephen Cole, Sudhakaran Prabakaran

sp339@cam.ac.uk

HIGHLIGHTS

Phosphorylation site mutations contribute to pathological states such as cancer

Existing classifiers do not account for the alteration of kinase recognition motifs

Our PhosphoEffect classifier prioritizes variants better than existing ones

Cole & Prabakaran, iScience *23*, 101321
August 21, 2020 © 2020 The Author(s).
https://doi.org/10.1016/j.isci.2020.101321

iScience

CellPress
OPEN ACCESS

Article

PhosphoEffect: Prioritizing Variants On or Adjacent to Phosphorylation Sites through Their Effect on Kinase Recognition Motifs

Stephen Cole[1] and Sudhakaran Prabakaran[1,2,*]

SUMMARY

Phosphorylation sites often have key regulatory functions and are central to many cellular signaling pathways, so mutations that modify them have the potential to contribute to pathological states such as cancer. Although many classifiers exist for prioritization of coding genomic variants, to our knowledge none of them explicitly account for the alteration or creation of kinase recognition motifs that alter protein structure, function, regulation of activity, and interaction networks through modifying the pattern of phosphorylation. We present a novel computational pipeline that uses a random forest classifier to predict the pathogenicity of a variant, according to its direct or indirect effect on local phosphorylation sites and the predicted functional impact of perturbing a phosphorylation event. We call this classifier PhosphoEffect and find that it compares favorably and with increased accuracy to the existing classifier PolyPhen 2.2.2 when tested on a dataset of known variants enriched for phosphorylation sites and their neighbors.

INTRODUCTION

Mutations in the coding region of the human genome can contribute to pathological phenotypes through their molecular effect on the structure and function of proteins. Although many genetic diseases are caused by well-characterized mutations in single genes, others, such as cancer, are complex polygenic conditions resulting from the accumulation of numerous genomic mutations, with different cancers exhibiting a wide variety of mutational patterns.

In the post-genomic era, next-generation sequencing (NGS) platforms such as Illumina and IonTorrent can generate huge amounts of sequence data and identify millions of variants in a single genome (Zhang et al., 2011). This can facilitate the identification of pathogenic mutations in tumors and aid in their prediction, diagnosis, prognosis, and intervention, as well as guide in the design of novel targeted drugs (Torshizi and Wang, 2018). However, the majority of mutations in any cancer genome are benign, so-called passenger mutations, which arise by chance and may hitchhike to fixation if they co-occur with mutations beneficial to the growth of the tumor, the so-called driver mutations. Distinguishing driver from passenger mutations has been the holy grail of cancer diagnosis and treatment (Pon and Marra, 2015).

Recent developments in machine learning algorithms, which can predict the impact of mutations on protein structure and/or function and "prioritize" them based on clinical relevance have indeed come to the rescue of cancer diagnosis and treatment. These algorithms include PolyPhen (Adzhubei et al., 2010), SIFT (Sorting Intolerant From Tolerant; Sim et al., 2012), FATHMM (Functional Analysis Through Hidden Markov Models; Rogers et al., 2018), CADD (Combined Annotation Dependent Depletion; Rentzsch et al., 2018), and MutationTaster (Schwarz et al., 2014). Their predictions are based on a wide range of features including sequence context, conservation, and predicted impact on protein structure. They also take into account post-translational modifications (PTMs), such as phosphorylation, as these can regulate the structure, function, and interaction partners of a protein and play important roles in cellular signaling (Ardito et al., 2017).

Mutations that perturb PTMs—for instance, substituting a phosphorylation site for a residue that cannot be phosphorylated— can have a significant impact on the function of a protein and/or the regulation of its

[1]Department of Genetics, University of Cambridge, Downing Site, Cambridge CB2 3EH, UK

[2]Lead Contact

*Correspondence: sp339@cam.ac.uk

https://doi.org/10.1016/j.isci.2020.101321

This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

activity (Radivojac et al., 2008). Efforts have therefore been made to catalog missense variants that map to phosphorylation sites, for instance, AWESOME (A Website Exhibits SNP On Modification Event; Yang et al., 2019), ActiveDriverDB (Krassowski et al., 2018), and MIMP (Mutation Impact on Phosphorylation, Wagih et al., 2015). Similar attempts have been made for ubiquitination with some experimental validation (Martínez-Jiménez et al., 2020).

However, not all phosphorylation sites are necessarily functional, and the regulatory importance of phosphorylation sites can vary greatly. Several studies have therefore aimed to classify functional versus nonfunctional phosphorylation sites according to features such as sequence context, conservation, and structure (Beltrao et al., 2012; Xiao et al., 2016). Additionally, residues of neighboring phosphorylation sites can influence the degree of phosphorylation, since protein kinases largely recognize short linear motifs surrounding the target residue. Thus, mutations in the flanking region of phosphorylation sites have the potential to greatly influence protein structure and function by up- or downregulating the stoichiometry of phosphorylation at that site. However, as far as we are aware, existing variant classifiers do not explicitly consider these effects.

Additionally, given that protein phosphorylation frequently modifies protein-protein interaction (PPI) partners and this modification is central to many, if not most, cellular signaling pathways, another important consideration is the impact of disrupting or altering phosphorylation sites on its PPI, and indeed on the whole PPI network. The study of molecular interaction networks has become increasingly applied to human diseases such as cancer, with the effect of many known cancer mutations being understood through their effect on PPI networks (Hijazi et al., 2020; Creixell et al., 2015). But one of the major shortcomings of existing phosphorylation classifiers in our view is their failure to take such network information into account and implement them in the algorithm.

Finally, variant prioritization algorithms generally consider sequence conservation as a key feature, based on the reasoning that mutations in highly conserved regions have undergone negative selection and are thus most likely deleterious. However, this may penalize against pathogenic variants that are on or adjacent to phosphorylation sites, since these are enriched in disordered regions that are generally poorly conserved (Krassowski et al., 2018).

In light of this, we sought to develop a classifier that built upon existing algorithms for prioritizing variants but included a greater range of features describing the effect of a mutation on protein structure and function, as well as the broader PPI network, through its effect on direct and/or neighboring phosphorylation sites. We collected and annotated a set of variants of known clinical significance, then used the annotated features to train and test a random forest-based classifier. We find that the classifier, which we call PhosphoEffect, compares favorably to PolyPhen. We thus present a user-ready, open-source pipeline that can take as input a list of single amino acid point mutations and output the predicted probability that each mutation is pathogenic.

RESULTS

Enrichment of Mutations around Phosphorylation Sites

We retrieved all annotated human phosphosites from PhosphoSitePlus (Hornbeck et al., 2015), a comprehensive database of PTMs curated from UniProtKB and published experimental datasets. These were mapped onto missense mutations derived from the COSMIC (Catalogue Of Somatic Mutations In Cancer) database, a comprehensive, manually curated genomic data repository for human cancers (Tate et al., 2019). As described in the Methods section, we took a sample of 100 tumors of each of the main tumor types in the dataset of whole-genome/whole-exome sequence studies, containing a total of 282,019 single amino acid mutations. We evaluated the frequency of mutations on and around phosphorylation sites to test for enrichment (Figure 1). We found that mutations were underrepresented in phosphorylation sites (hypergeometric test, fold enrichment = 0.94, Benjamini-Hochberg (BH)-corrected $p = 1.60 \times 10^{-5}$). In contrast, mutations were significantly overrepresented at positions −4, −3, −2, −1, and +1 (fold-enrichments 1.06, 1.14, 1.03, 1.08, and 1.05 respectively, all BH-corrected p values < 0.05), whereas there was no significant enrichment in either direction at the remaining positions (Table 1).

We hypothesized that mutations on phosphorylation sites are underrepresented because they are more likely to abolish the phosphorylation site altogether and be deleterious, whereas mutations in the

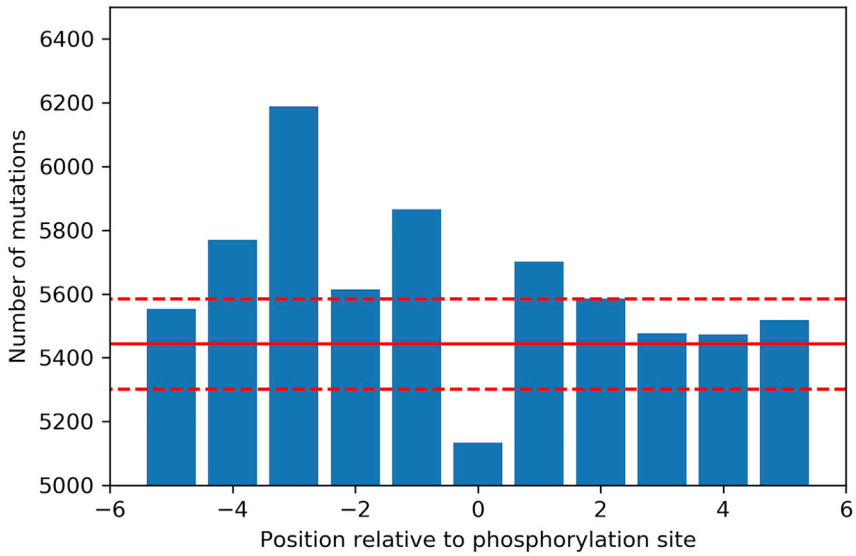


Figure 1. Enrichment of Mutations in the Vicinity of Phosphorylation Sites in Human Cancers
The solid red line shows the expected number of mutations at each site, whereas the dotted lines are the upper and lower 2.5% bounds of the corresponding hypergeometric distribution, respectively.

neighborhood of phosphorylation sites that alter the binding affinity of kinases and/or phosphatases may fine-tune the level of phosphorylation and lead to deregulated protein function, rationalizing their overenrichment in tumors.

Evaluating the Success of Existing Predictors on Variants Affecting Phosphorylation Sites

Existing predictors such as PolyPhen do not take into account the effect of mutations on residues of neighboring phosphosites nor do they explicitly consider the network-rewiring effect of mutations affecting phosphorylation. We assessed the performance of PolyPhen on three classes of variants: mutations mapping to phosphorylation sites (class 1), mutations that are not directly phosphorylated but are within five residues of a phosphorylation site (class 2), and mutations that have no phosphorylation sites in

Site Relative to Phosphosite	Fold Enrichment of Mutations	BH-Corrected p Value
−5	1.02	0.126
−4	1.06	6.86×10^{-6}
−3	1.14	2.42×10^{-8}
−2	1.03	1.79×10^{-2}
−1	1.08	3.08×10^{-8}
0	0.94	1.60×10^{-5}
1	1.05	3.54×10^{-4}
2	1.03	4.74×10^{-2}
3	1.01	0.632
4	1.01	0.662
5	1.01	0.296

Table 1. Enrichment of Mutations in the Vicinity of Phosphorylation Sites in the COSMIC Cancer Mutation Database
All hypergeometric test p -values are Benjamini-Hochberg corrected for multiple testing.

Feature	Source
Molecular weight of wild-type residue (Da)	
Molecular weight of mutant residue (Da)	
Chemical property of wild-type residue (acidic, basic, polar, nonpolar)	
Chemical property of mutant residue	
Polyphen score	Polyphen 2.2.2 (Adzhubei et al., 2010)
Secondary structure (helix, sheet, disordered, other)	RING (Piovesan et al., 2016)
Wild-type residue phosphorylated?	PhosphoSitePlus (Hornbeck et al., 2015)
Number of phosphorylated neighbors (+-5)	PhosphoSitePlus
Impact on phosphorylation level	Derived from NetPhorest (Horn et al., 2014)
Network perturbation score	Derived from iPTMnet (Huang et al., 2018) and STRING database (Szklarczyk et al., 2019)
Number of internal contacts (of neighboring phosphorylated residues)	RING
RAPDF (of neighboring phosphorylated residues)	RING

Table 2. Features Used to Train the Classifier

RAPDF, residue-specific all atom-dependent conditional probability distribution function, RING, residue interaction network generator.

the +-5 region (class 3) (Figure 2). A total of 21,037 variants of known clinical significance were retrieved from the ClinVar database as described in the Methods section. Of these, 450 were classified as class 1, 3,452 were class 2, and 17,135 were class 3. PolyPhen classifies variants as "probably damaging" (false positive rate <10%), "possibly damaging" (false positive rate 10%–20%), or benign (false-positive rate >20%). For the purpose of the following evaluation we are classifying predictions of "probably damaging" or "possibly damaging" as pathogenic.

We found that PolyPhen had the highest sensitivity on the class 3 variants, with a true positive rate of 88% indicating that 12% of pathogenic variants were incorrectly classified as benign. In contrast, 16% of pathogenic class 1 variants and 17% of pathogenic class 2 variants were classified as benign. A chi-square test indicated a significant difference between the three proportions (chi-square = 39.67, df = 2, $p = 2.43 \times 10^{-9}$). In *ad hoc* pairwise testing the only significant difference was between class 2 and class 3 (chi-square = 36.97, df = 1, BH-corrected $p = 7.19 \times 10^{-9}$, Table S1). This indicates that PolyPhen has a tendency to underestimate the pathogenicity of mutations that occur in the neighborhood of phosphorylation sites.

In regard to specificity, the false-positive rate was 25% for class 1 variants, 23% for class 2, and 26% for class 3. The differences between the groups were marginally significant (chi-square = 6.18, df = 2, p = 0.0455), and *ad hoc* pairwise testing showed a marginally significant difference between class 2 and class 3 (chi-square = 6.02, df = 1, p = 0.424, all p values BH corrected for multiple testing). However, the effect size is small and the significance of this difference, if any, is unclear.

The Influence of Mutations on Kinase Recognition Motifs

To test the hypothesis that mutations in the neighborhood of phosphorylation sites could contribute to disease by altering or creating kinase recognition motifs, and thus modifying the strength or regulation of phosphorylation sites, we needed a metric to predict the strength of phosphorylation at a site as a function of its sequence context. To this end, we used the NetPhorest classifier (Horn et al., 2014) as described in the

Hyperparameter	Value
Bootstrap	True
Criterion	Gini
Maximum leaf nodes	None
Minimum impurity decrease	1×10^{-5}
Minimum impurity split	None
Minimum samples per leaf	1
Minimum samples per split	2
Number of estimators	100

Table 3. Hyperparameters of the Random Forest Classifier
The optimal hyperparameters, selected by grid search cross-validation, for the random forest classifier are shown.

Methods section and derived a score as the sum of the probabilities that a site would be phosphorylated by each possible class of kinase.

To illustrate this, an example is shown of p53 (Figure 3), one of the most commonly mutated genes in human cancers (Muller and Vousden, 2013). The residue Ser20 is itself phosphorylated and also lies just downstream of two other phosphosites, Ser15 and Thr18. A point mutation at this position therefore has the potential to have a triple effect on the phosphorylation stoichiometry of the protein.

Figure 3 shows the NetPhorest predictions for these three residues in the wild-type and in two different point mutants: serine to tyrosine (S20Y), which can also be phosphorylated, and serine to proline (S20P), which abrogates phosphorylation at this position. Neither mutation affects the sole ATM/ATR kinase recognition motif of Ser15, so the score on this phosphorylation site would be 0 for both mutants. In contrast, the other two phosphosites are altered by both mutations. S20Y increases the summed strength for Thr18 phosphorylation from 1.03 to 1.1, so has a score of 0.07 on this site, whereas Tyr20 has a summed strength of 0.85 compared with 0.53 on Ser20, a score of 0.32. This results in a total score of 0.39 for the S20Y mutant, reflecting its impact on the strengths of direct and neighboring phosphosites.

The less conservative S20P mutant reduces the strength of Thr18 phosphorylation to 0.85 and completely abolishes the Ser20 phosphorylation as proline lacks the phosphorylatable hydroxyl group. The total score for S20P is therefore $|(0.85\text{-}1.03)| + |(0\text{-}0.53)| = 0.71$, representing a bigger impact of this mutation on the surrounding phosphorylation pattern.

This procedure was applied to the class 2 variants, as described in the previous section, to determine whether there was a correlation between the impact of a variant on local phosphorylation sites and its pathogenicity. The mean score of the pathogenic variants was 0.21, whereas that of the benign variants was 0.20 (Mann-Whitney $U = 1{,}429{,}324$, $n1 = 1{,}575$, $n2 = 1{,}877$, $p = 0.047$, Figure S1). Although only marginally significant, this supports our hypothesis that deleterious variants are slightly more likely to influence local phosphorylation patterns than benign ones, and so this impact score was used as a feature for training the classifier. Additionally, since some mutations have multiple phosphorylation sites in their flanking regions, structural and functional features of neighboring phosphosites that were used to estimate the impact of perturbing phosphorylation were weighted according to the predicted impact of the mutation on that phosphosite, as detailed in the Methods section.

Performance of the Random Forest Classifier

A random forest classifier, which we call PhosphoEffect, was trained on a training set of 2,274 benign and 2,400 pathogenic mutants (Tables S2, S3, S4, and S5), as described in the Methods section. The features of the variants which were selected for training are shown in Table 2, and the optimal hyperparameters for the classifier, chosen by grid search cross-validation, are shown in Table 3.The size of the training set was limited by the number of known pathogenic missense variants corresponding to phosphosites or their

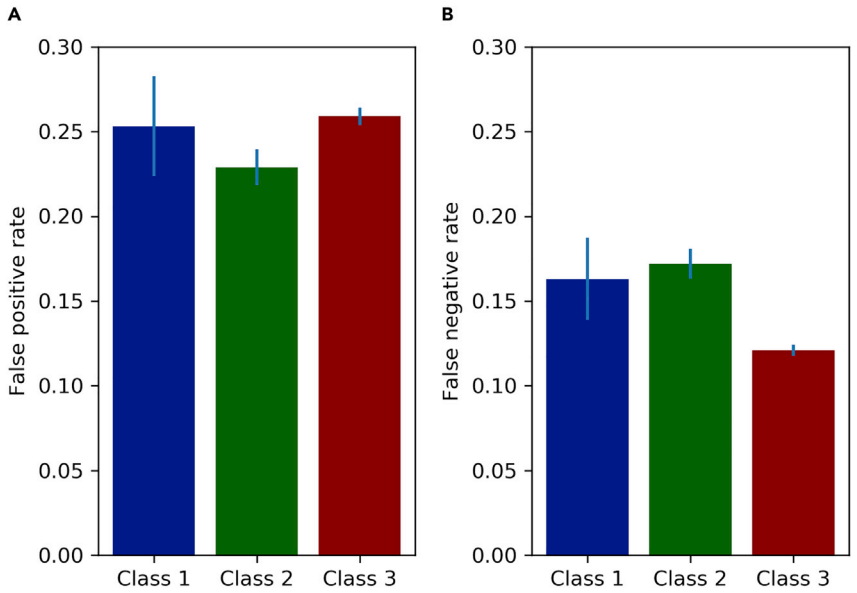


Figure 2. Performance of PolyPhen on Known Variants

The false-positive (A) and false negative (B) rates of the PolyPhen classifier on phosphorylated residues, residues neighboring phosphorylation sites, and all others. The error bars indicate standard errors of the proportion. The difference in false-positive rates between classes was significant (p=2.43×10^-9) and the difference in false-negative rates was marginally significant (p=0.0455). All p-values, including for ad-hoc pairwise testing, are in Table S1.

neighbors. The feature matrices, pre- and post-processing as described in the Methods section, can be found in Tables S6, S7, S8, S9, S10, S11, S12, and S13.

The performance of PolyPhen and PhosphoEffect were compared on a test set of 568 benign and 600 pathogenic mutants (Tables S14 and S15). PolyPhen had a true-positive rate of 81.2% and a false-positive rate of 25.4%, with an area under the receiver operating characteristic curve (AUROC) of 0.859. Using a 25% cutoff for false-positive rate for fair comparison, PhosphoEffect had a true positive rate of 86.3%, with an AUROC of 0.884 (Figure 4).

As with PolyPhen, probability scores associated with a 0%–10% false-positive rate are dubbed "probably pathogenic" and those with a 10%–20% false-positive rate are "possibly pathogenic." At this cutoff, 82.7% of pathogenic variants in the test set were assigned "possibly" or "probably" pathogenic, with a probability threshold of 0.485 for a variant to be classified as possibly or probably pathogenic.

Since the PolyPhen score was used as one of the features to train the classifier, an improvement is to be expected. When the classifier was trained in the absence of PolyPhen score as a feature, the result was an AUROC of 0.747, indicating the features used to update the prediction of variants have reasonable predictive power on their own.

Figure 5 shows the relative importance of the 18 features used to train PhosphoEffect (Table S16). Unsurprisingly, given that the aim of the classifier is to update the PolyPhen prediction based on the impact of a mutation on phosphorylation levels, the PolyPhen score is by far the most important predictor, with a feature importance of 0.54. The impact of the mutation on phosphorylation level according to NetPhorest was the second most important 0.10. This supports our hypothesis that the modification of kinase recognition motifs is an important contributor to the pathogenicity of a variant and should be explicitly included in variant prioritization algorithms.

Structural features of the variant and its neighboring phosphosites were of intermediate importance, but the network perturbation score is of virtually no importance. This is not surprising because most of the values for this score are zero, mainly due to the highly incomplete annotation of phosphorylation-dependent PPI, which are challenging to assay on a high-throughput scale and come solely from the curation of published low-throughput studies.

A

P04637			
S15 Kinase	ATM/ATR group	0.46	VEPPLSQETFS
T18 Kinase	PDHK group	0.21	PLSQETFSDLW
	ACT2/2B TGFbR2 group	0.15	PLSQETFSDLW
	CK2 group	0.13	PLSQETFSDLW
	DAPK group	0.11	PLSQETFSDLW
	GRK group	0.09	PLSQETFSDLW
	PKD group	0.09	PLSQETFSDLW
	PKC group	0.07	PLSQETFSDLW
	HIPK1/2 group	0.07	PLSQETFSDLW
	CK1 group	0.06	PLSQETFSDLW
	DMPK group	0.05	PLSQETFSDLW
S20 Kinase	CK2 group	0.26	SQETFSDLWKL
	GRK group	0.12	SQETFSDLWKL
	CK1 group	0.09	SQETFSDLWKL
	PKC group	0.06	SQETFSDLWKL

B

P04637			
S15 Kinase	ATM/ATR group	0.46	VEPPLSQETFY
T18 Kinase	PDHK group	0.18	PLSQETFYDLW
	PKC group	0.13	PLSQETFYDLW
	MAP2K group	0.13	PLSQETFYDLW
	ACT2/2B TGFbR2 group	0.12	PLSQETFYDLW
	PKD group	0.11	PLSQETFYDLW
	DAPK group	0.11	PLSQETFYDLW
	CK2 group	0.10	PLSQETFYDLW
	GRK group	0.09	PLSQETFYDLW
	CK1 group	0.07	PLSQETFYDLW
	HIPK1/2 group	0.06	PLSQETFYDLW
Y20 Kinase	MAP2K group	0.09	SQETFYDLWKL
	Src group	0.05	SQETFYDLWKL
	Eph group	0.05	SQETFYDLWKL
	Abl group	0.04	SQETFYDLWKL
	KDR FLT1 group	0.03	SQETFYDLWKL
	Met group	0.03	SQETFYDLWKL
	InsR group	0.02	SQETFYDLWKL
	EGFR group	0.02	SQETFYDLWKL
	Tec group	0.02	SQETFYDLWKL
	Syk group	0.02	SQETFYDLWKL

C

P04637			
S15 Kinase	ATM/ATR group	0.46	VEPPLSQETFP
T18 Kinase	ACT2/2B TGFbR2 group	0.13	PLSQETFPDLW
	PDHK group	0.13	PLSQETFPDLW
	DAPK group	0.11	PLSQETFPDLW
	PKD group	0.09	PLSQETFPDLW
	PKC group	0.08	PLSQETFPDLW
	CK2 group	0.08	PLSQETFPDLW
	HIPK1/2 group	0.07	PLSQETFPDLW
	CK1 group	0.06	PLSQETFPDLW
	DMPK group	0.05	PLSQETFPDLW
	TTK	0.05	PLSQETFPDLW

Figure 3. NetPhorest Outputs for p53 with Point Mutations at Ser20
(A–C) (A) Wild-type sequence, (B) Mutant S20Y, (C) Mutant S20P.

Pipeline

We present a pipeline implemented in Python (compatible with versions 3.5 and above) into which the user can input a tab-delimited text file containing a list of point mutations to query (format UniProtKB accession number, residue number, wild-type amino acid, mutant amino acid) to obtain a list of classifications along with associated probabilistic scores. A separate output file logs any errors that are raised during the running of the pipeline, for instance, if a wild-type residue in the input does not match the canonical sequence.

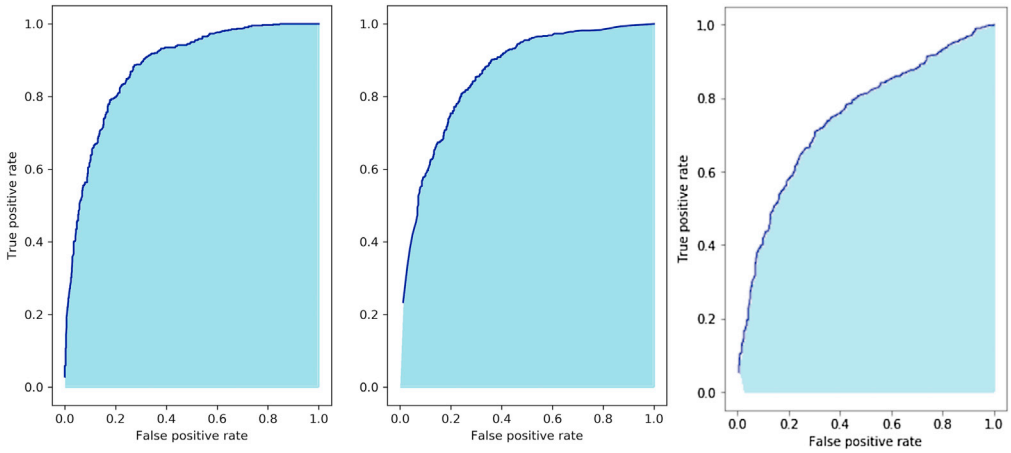


Figure 4. Performance of the PhosphoEffect Compared with PolyPhen on the Test Dataset

Receiver operating characteristic (ROC) curves for (A) PhosphoEffect classifier and (B) PolyPhen; (C) PhosphoEffect without PolyPhen feature and their respective areas under the curve are 0.884, 0.859, 0.747.

DISCUSSION

In this work we have developed a new classifier for the prioritization of missense variants according to their effect on phosphorylation sites. It takes as a base the score assigned by PolyPhen representing the probability that the variant is pathogenic and updates this score using a variety of features reflecting the impact of the mutation on kinase recognition motifs (resulting in the enhancement, disruption, or deregulation of phosphorylation sites) and the structural and functional impact of such perturbations.

On our testing set, which was enriched for mutations on or around phosphosites, PhosphoEffect incorrectly classified 17.3% of pathogenic variants as benign, compared with 18.8% for PolyPhen, with a substantially lower false-positive rate, representing a clear improvement in the accuracy of distinguishing benign and pathogenic variants.

Additionally, the impact of a mutation on local phosphorylation strength was the second most important feature in the classifier, after the PolyPhen score. This supports the inclusion of features describing the direct and indirect impact of mutations on phosphorylation sites, and such an approach could identify large numbers of clinically relevant mutations that act indirectly through modifying the strength of neighboring phosphosites, which are overlooked by existing classifiers.

In the era of personalized medicine, where patients with cancer are increasingly having their tumor genomes sequenced, the identification of novel pathogenic variants could aid in diagnosis, prognosis, and choice of treatment, as well as targeted drug design. We hope that this computational pipeline will facilitate the identification of new potentially deleterious mutations that rewire phosphorylation-dependent signaling in cancer and can be clinically validated.

Limitations of the Study

However, there remains a lot of work to be done in this area. For instance, our metric for assessing the impact of mutations on the global PPI network through modification of phosphorylation barely contributed at all to the performance of the classifier, as for almost all variants this score was zero. This is most likely due to the scarcity of data on PPI, since iPTMnet is a database curated from evidence from published studies on individual PPI and phosphorylation sites. Since one of the major roles of phosphorylation events is to alter PPIs (Nishi et al., 2011) and phosphorylation is central to many disease-related signaling pathways (Ardito et al., 2017), we do believe that the network-rewiring impact of mutations is crucial to consider when prioritizing variants. Future work would entail creating a new model to predict the dependency of a PPI on a given phosphorylation event, for instance, based on the presence of phosphoprotein binding domains and their recognition motifs (Guo et al., 2019), which could itself be trained on data from the iPTMnet database and would give a more accurate estimation of the impact of a mutation on the protein-protein

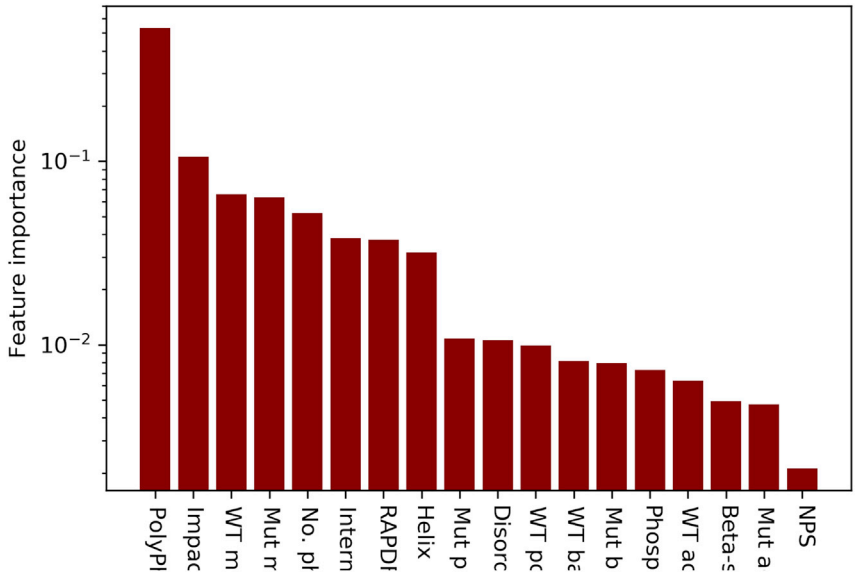


Figure 5. Feature Importances

The relative contribution of each feature on which the classifier was trained to the model predictions; note log-scale of y axis.

interaction network. Furthermore, there are many more features that could have been included in the model, such as the solvent accessibility or number of water contacts per phosphorylated residue—these may indicate how likely a phosphorylation event is to modify the structure or interactions of a residue.

Resource Availability

Lead Contact

Sudhakaran Prabakaran email: sp339@cam.ac.uk

Materials Availability

Not applicable

Data and Code Availability

All codes for this work can be obtained from https://github.com/PrabakaranGroup/PhosphoEffect-pipeline.

METHODS

All methods can be found in the accompanying Transparent Methods supplemental file.

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.isci.2020.101321.

ACKNOWLEDGMENTS

We thank the editor and the reviewers for their helpful comments that strengthened this paper. Funding: Nothing to declare.

AUTHOR CONTRIBUTIONS

S.C. performed all the analysis, interpreted data, and wrote the manuscript. S.P. designed and supervised the work, interpreted the data, and wrote the manuscript.

DECLARATION OF INTERESTS

S.P. is a cofounder of NonExomics, LLC.

Received: December 16, 2019
Revised: May 10, 2020
Accepted: June 25, 2020
Published: August 21, 2020

REFERENCES

Adzhubei, I.A., Schmidt, S., Peshkin, L., Ramensky, V.E., Gerasimova, A., Bork, P., Kondrashov, A.S., and Sunyaev, S.R. (2010). A method and server for predicting damaging missense mutations. Nat. Methods *7*, 248–249.

Ardito, F., Giuliani, M., Perrone, D., Troiano, G., and Lo Muzio, L. (2017). The crucial role of protein phosphorylation in cell signaling and its use as targeted therapy. Int. J. Mol. Med. *40*, 271–280.

Beltrao, P., Albanèse, V., Kenner, L.R., Swaney, D.L., Burlingame, A., Villén, J., Lim, W.A., Fraser, J.S., Frydman, J., and Krogan, N.J. (2012). Systematic functional prioritization of protein posttranslational modifications. Cell *150*, 413–425.

Creixell, P., Schoof, E.M., Simpson, C.D., Longden, J., Miller, C.J., Lou, H.J., Perryman, L., Cox, T.R., Zivanovic, N., Palmeri, A., et al. (2015). Kinome-wide decoding of network-attacking mutations rewiring cancer signaling. Cell *163*, 202–217.

Guo, Y., Peng, D., Zhou, J., Lin, S., Wang, C., Ning, W., Xu, H., Deng, W., and Xue, Y. (2019). iEKPD 2.0: an update with rich annotations for eukaryotic protein kinases, protein phosphatases and proteins containing phosphoprotein-binding domains. Nucleic Acids Res. *47*, D344–D350.

Hijazi, M., Smith, R., Rajeeve, V., Bessant, C., and Cutillas, P.R. (2020). Reconstructing kinase network topologies from phosphoproteomics data reveals cancer-associated rewiring. Nat. Biotechnol. *38*, 493–502.

Horn, H., Schoof, E.M., Kim, J., Robin, X., Miller, M.L., Diella, F., Palma, A., Cesareni, G., Jensen, L.J., and Linding, R. (2014). KinomeXplorer: an integrated platform for kinome biology studies. Nat. Methods *11*, 603–604.

Hornbeck, P.V., Zhang, B., Murray, B., Kornhauser, J.M., Latham, V., and Skrzypek, E. (2015). PhosphoSitePlus, 2014: mutations, PTMs and recalibrations. Nucleic Acids Res. *43*, D512–D520.

Huang, H., Arighi, C.N., Ross, K.E., Ren, J., Li, G., Chen, S.C., Wang, Q., Cowart, J., Vijay-Shanker, K., and Wu, C.H. (2018). iPTMnet: an integrated resource for protein post-translational modification network discovery. Nucleic Acids Res. *46*, D542–D550.

Krassowski, M., Paczkowska, M., Cullion, K., Huang, T., Dzneladze, I., Ouellette, B.F.F., Yamada, J.T., Fradet-Turcotte, A., and Reimand, J. (2018). ActiveDriverDB: human disease mutations and genome variation in post-translational modification sites of proteins. Nucleic Acids Res. *46*, D901–D910.

Martínez-Jiménez, F., Muiños, F., López-Arribillaga, E., Lopez-Bigas, N., and Gonzalez-Perez, A. (2020). Systematic analysis of alterations in the ubiquitin proteolysis system reveals its contribution to driver mutations in cancer. Nat. Cancer *1*, 122–135.

Muller, P.A.J., and Vousden, K.H. (2013). p53 mutations in cancer. Nat. Cell Biol. *15*, 2–8.

Nishi, H., Hashimoto, K., and Panchenko, A. (2011). Phosphorylation in protein-protein binding: effect on stability and function. Structure *19*, 1807–1815.

Piovesan, D., Minervini, G., and Tosatto, S.C. (2016). The RING 2.0 web server for high quality residue interaction networks. Nucleic Acids Res. *44*, W367–W374.

Pon, J., and Marra, M. (2015). Driver and passenger mutations in cancer. Ann. Rev. Pathol. *10*, 25–50.

Radivojac, P., Baenziger, P.H., Kann, M.G., Mort, M.E., Hahn, M.W., and Mooney, S.D. (2008). Gain and loss of phosphorylation sites in human cancer. Bioinformatics *24*, i241–i247.

Rentzsch, P., Witten, D., Cooper, G.M., Shendure, J., and Kircher, M. (2018). CADD: predicting the deleteriousness of variants throughout the human genome. Nucleic Acids Res. *47*, D886–D894.

Rogers, M.F., Shihab, H.A., Mort, M., Cooper, D.N., Gaunt, T.R., and Campbell, C. (2018). FATHMM-XF: accurate prediction of pathogenic point mutations via extended features. Bioinformatics *34*, 511–513.

Schwarz, J.M., Cooper, D.N., Schuelke, M., and Seelow, D. (2014). MutationTaster2: mutation prediction for the deep-sequencing age. Nat. Methods *11*, 361–362.

Sim, N., Kumar, P., Hu, J., Henikoff, S., Schneider, G., and Ng, P. (2012). SIFT web server: predicting effects of amino acid substitutions on proteins. Nucleic Acids Res. *40*, W425–W427.

Szklarczyk, D., Gable, A.L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., Simonovic, M., Doncheva, N.T., Morris, J.H., Bork, P., et al. (2019). STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Res. *47*, D607–D613.

Tate, J., Bamford, S., Jubb, H.C., Sondka, Z., Beare, D.M., Bindal, N., Boutselakis, H., Cole, C.G., Creatore, C., Dawson, E., et al. (2019). COSMIC: the catalogue of somatic mutations in cancer. Nucleic Acids Res. *47*, D941–D947.

Torshizi, A., and Wang, K. (2018). Next-generation sequencing in drug development: target identification and genetically stratified clinical trials. Drug Disc Today *23*, 1776–1783.

Wagih, O., Reimand, J., and Bader, G.D. (2015). MIMP: predicting the impact of mutations on kinase-substrate phosphorylation. Nat. Methods *12*, 531–533.

Xiao, Q., Miao, B., Bi, J., Wang, Z., and Li, Y. (2016). Prioritizing functional phosphorylation sites based on multiple feature integration. Sci. Rep. *6*, 24735.

Yang, Y., Peng, X., Ying, P., Tian, J., Li, J., Ke, J., Zhu, Y., Gong, Y., Zou, D., Yang, N., et al. (2019). AWESOME: a database of SNPs that affect protein posttranslational modifications. Nucleic Acids Res. *47*, D874–D880.

Zhang, J., Chiodini, R., Badr, A., and Zhang, G. (2011). The impact of next-generation sequencing on genomics. J. Genet. Genomics *38*, 95–109.

iScience, Volume *23*

Supplemental Information

PhosphoEffect: Prioritizing Variants On or Adjacent to Phosphorylation Sites through Their Effect on Kinase Recognition Motifs

Stephen Cole and Sudhakaran Prabakaran

Transparent Methods

Collection of published cancer genome mutations

A dataset of mutations in human cancers was downloaded from COSMIC (Catalogue Of Somatic Mutations In Cancer), a comprehensive, manually curated genomic data repository for human cancers (Tate et al. 2019). The dataset downloaded corresponded only to coding mutations identified in whole genome or whole exome sequencing experiments, as opposed to targeted studies, because we were interested in the relative frequencies of cancer mutations and targeted experiments are biased towards known cancer-related genes. The dataset was filtered to retain only single amino acid point mutations, giving a total of 3,874,127 mutations corresponding to 19,184 genes and 27,332 tumours.

The mutations were derived from 20 different cancer types. Since some cancer types were greatly overrepresented in the dataset due to being more comprehensively studied, we took a random stratified sample of 100 tumours from each cancer type. The resulting dataset of 2,000 tumours contained 282,019 mutations and was used for downstream analysis.

Mapping published phosphorylation sites to mutations and enrichment analysis

Annotated phosphorylation sites for the entire human genome were downloaded from PhosphoSitePlus. These were then mapped to the COSMIC dataset to determine the number of mutations corresponding to phosphorylation sites or their flanking regions (+-5 residues).

Assuming mutations were randomly distributed with respect to phosphorylation sites, the expected number of mutations mapping to any given position relative to a phosphosite should be the number of mutations in the dataset multiplied by the proportion of all sites in the proteome that are annotated as phosphorylation sites.

Since PhosphoSitePlus contains 216,597 annotated phosphosites and the summed length of all proteins encoded by unique genes (i.e. excluding splice isoforms or other variants) is 11,222,506, this proportion is 0.0193 and our dataset of 282,080 mutations should be expected to contain 5,443 mapping directly to phosphosites (and the same number mapping to positions directly up- or downstream of phosphosites, with a minor drop-off due to phosphorylations at the N- or C-termini, which was not considered in this analysis). Hypergeometric tests were performed to test for significant under- or over enrichment of mutations at positions on or neighbouring phosphosites, with Benjamini-Hochberg false discovery rate correction for multiple testing.

Collection and annotation of known variants

In order to train and test the model we collected a set of variants of known clinical significance and annotated them with a range of biophysical and functional features relating to their impact on protein phosphorylation. Human genomic variants of known pathogenic significance - either “benign” or “pathogenic” - were downloaded from ClinVar (Landrum et al. 2016), focussing only on single amino acid missense mutations. A total of 21037 mutations were retrieved (8883 benign and 12154 pathogenic). Mutations were classified as class 1 (on phosphosites), class 2 (not on a phosphosite, but within 5 residues of a phosphosite), or class 3 (not within 5 residues of a phosphosite).

In order to train our model to detect whether direct or neighbouring phosphorylation is a predictor of pathogenicity, our final subset of variants used to train the model - 3000 pathogenic and 2842 benign - consisted of a split of two-thirds class 1 and 2, and one-third class 3. While benign variants are much more common than pathogenic variants in real genomic datasets, reviews have shown that using a balanced training set for binary classifiers gives better predictive power for disease-causing mutations [Wei and Dunbrack, 2013].

Prediction of impact of mutations on phosphorylation pattern

To predict the impact of a mutation on or neighbouring a phosphorylation site on the strength of phosphorylation, we used the open-source software NetPhorest 2.1 which uses a neural network/random forest hybrid approach to estimate the likelihood of a linear motif being phosphorylated by different classes of kinases based on their known targets. The NetPhorest algorithm can be downloaded and queried from the command line, taking as input a fasta file and outputting each predicted potential phosphosite, along with a series of scores reflecting the estimated probability of the site being phosphorylated by each class of kinases, given the five residues up- and downstream of the site.

A Python script was written to download the canonical fasta file of each protein given its accession number and to edit the fasta file to incorporate a given point mutation. If one or more annotated phosphorylation sites were situated within five residues up-or downstream of the mutation, the canonical (wild-type) and edited (mutant) fasta files were queried in NetPhorest. An estimator of the change in phosphorylation strength at a given phosphosite incurred by the mutation was calculated as the absolute value of the difference of the summed strengths across all kinase classes for the phosphosite in the mutant and wild-type sequences. For mutations with multiple phosphosites in the +-5 region, the total score for the mutation is the sum of the scores for each phosphosite. This is shown in Equation 1 below:

$$\alpha_m = \sum_p \left| \sum_k (x_{k,p,WT} - x_{k,p,M}) \right|$$

Where xk,p,WT is the recognition strength of kinase k at phosphorylation site p on the wild-type sequence, and xk,p,M is the strength on mutant M.

Point mutations that occur directly on phosphorylation sites may enhance, decrease or abrogate phosphorylation. In the latter case, the impact on local phosphorylation strength will be large; however, depending on the properties of the mutant residue, the mutation may mimic constitutive phosphorylation. For instance, mutations of threonine to the basic residues aspartate glutamate can have this phosphomimetic effect due to the introduction of a negative charge; on the other hand, a tyrosine to glutamate mutation is unlikely to mimic tyrosine phosphorylation as the introduction of the negative charge will be outweighed by the loss of the bulky benzene ring (Chen and Cole, 2015). For this reason, the molecular weight and chemical properties of both the wild-type and mutant residues were included as features in the model.

Prediction of impact on protein structure

To assess the impact of a phosphomutation on the structure of the protein, we used the software Residue Interaction Network Generator 2.0 (RING 2.0; Piovesan et al. 2016), downloaded with the kind permission of the authors, to obtain biophysical features of the phosphosites. Protein Data Bank (PDB) structures for proteins were accessed using the ePDB application programmatic interface (API) best_structures tool, which ranks all the PDB structures for a given protein by quality (sequence coverage and resolution). We sought to identify the highest quality structure which included the residue of interest plus a flanking region of at least 5 residues (or up to the terminus for those within 5 residues of the N- or C-terminus); if there was no such structure available, we identified the structure with the maximum possible flanking sequence, to maximise the likelihood that the residue of interest would be found in its native structure and sequence context. The PDB files were fed into the RING software. Given that many PDB structures are of protein complexes, we considered only chain-internal contacts between amino acids, and also considered contacts between amino acid residues and water molecules.

From the RING output file, we extracted the secondary structure, degree (number of chain-internal inter-residue contacts), and the residue-specific all atom-dependent conditional probability distribution function (RAPDF), a measure of the thermodynamic stability of the residue (Samudrala and Moult, 1998). For the latter two, we calculated a weighted mean for each of the phosphosites within the +5/-5 flanking region of the residue of interest, weighted according to the predicted impact of the mutation on the strength of phosphorylation as described in the previous section.

Prediction of impact on protein-protein interaction networks

To predict how a mutation would perturb protein-protein interactions (PPIs) through its effect on phosphorylation, we used data from iPTMnet, an integrated database for post-translational modifications (Huang et al. 2018). iPTMnet, accessible through the Python API pyiptmnet, contains data curated from literature evidence of the impact of phosphorylation events on PPIs. Four types of phosphorylation-dependent association are given: increased association, decreased association, inhibited association, and unknown, where there was evidence for an effect of the phosphorylation event on the PPI, but not the direction of the effect. In many cases there were multiple sources supporting the dependency of a PPI on a given phosphorylation event, and occasionally these had different association types. This was resolved by assigning a dependency score to each phosphorylation-dependent association: inhibited association was scored as 2; increased or decreased association were scored as 1; and where different sources indicated a different association type, the arithmetic mean of these scores was used. A score of 1 was assigned when the type of association was unknown.

We used the edge betweenness centrality of a PPI in the global human PPI network as an estimator for its importance, and to assess the degree to which a mutation altering phosphorylation would perturb the PPI network. Genome-wide PPI data was downloaded from the STRING (Search Tool for the Retrieval of INteracting Genes/proteins) database (version 11, Szklarczyk et al. 2019). Associations were filtered to include only physical binding interactions (611,087), and Ensembl_PRO IDs were converted to UniProt IDs using the human ID mapping database in UniProt, excluding isoform specifications. The interaction data was used to reconstruct the global PPI network using the Python module networkx, which contains a range of classes and functions for the construction, analysis and visualisation of networks (Hagberg et al. 2008). The reconstructed network had 14,527 nodes and 312,626 edges, and from this the betweenness centrality of each edge was calculated, defined as the proportion of shortest paths between all node pairs that include that edge.

From this, a network perturbation score (NPS) was estimated for each mutation as follows: for each phosphorylation site within the +-5 region, a site-specific score was calculated as the sum of the dependency score multiplied by the edge betweenness centrality for each PPI dependent on that phosphorylation; the overall score for the mutation was calculated as the score for each phosphosite, again weighted according to the predicted impact of the mutation on the strength of phosphorylation at that site, as in Equation 2 below:

$$NPS = \frac{\sum_p (\beta_p \sum_{ip} \delta_{ip} \varepsilon_{ip})}{\sum_p \beta_p}$$

Where

$$\beta_p = \left| \sum_k (x_{k,p,WT} - x_{k,p,M}) \right|$$

is the impact of mutation M on phosphorylation site p, and ip are the protein-protein interactions dependent on phosphorylation at site p, with dependency δ and edge betweenness centrality ε. Where interactions in the iPTMnet database were absent from the STRING-derived network, the EBC was imputed as the median EBC for all interactions in the network (3.935×10^{-6}).

Model training and evaluation

Model training and evaluation was carried out using the scikit-learn package for machine learning in Python (Pedregosa et al. 2011). The final list of features which were used to train the model are shown in Table 2.

All categorical variables were encoded as one-hot dummy variables and quantitative variables were scaled to mean 0, variance 1. A random forest classifier was trained on the training set of 4674 variants, with grid search cross validation used to select the optimal hyperparameters for fitting. In the grid search method, different combinations of hyperparameters are sampled, the classifier is fit to the training set, and cross-validation is performed; the hyperparameters which result in the best performance in cross-validation are then selected. The optimal hyperparameters shown in Table 3.

Receiver operating characteristic (ROC) curves for the random forest classifier and for PolyPhen were generated by varying the probability threshold for classification as pathogenic from 0 to 1 in steps of 0.0001, and the area under the ROC curve (AUROC) was estimated using the trapezium rule.

Supplementary Figures

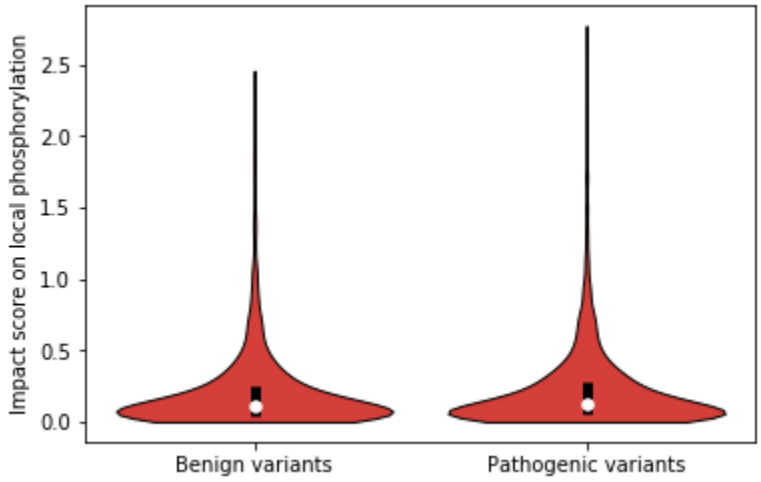


Figure S1, related to Figure 3: the impact score, derived from NetPhorest, of benign and pathogenic variants close to phosphorylation sites on the estimated strength of the phosphorylation sites according to their kinase recognition motifs.