A generic deep convolutional neural network framework for prediction of receptor–ligand interactions—NetPhosPan: application to kinase phosphorylation prediction

Fenoy, Emilio; Gonzalez-Izarzugaza, Jose Maria; Jurtz, Vanessa ; Brunak, Søren; Nielsen, Morten

Published in:
Bioinformatics

Link to article, DOI:
10.1093/bioinformatics/bty715

Publication date:
2019

Document Version
Peer reviewed version

A generic Deep Convolutional Neural Network framework for prediction of Receptor-ligand Interactions. NetPhosPan; Application to Kinase Phosphorylation prediction.

Emilio Fenoy[1], Jose M. G. Izarzugaza[2], Vanessa Jurtz[2], Søren Brunak[3] and Morten Nielsen[1,2,*]

[1] Instituto de Investigaciones Biotecnologicas, Universidad Nacional de San Martin, San Martin, B 1650 HMP, Buenos Aires, Argentina.

[2] Department of Bio and Health Informatics, Technical University of Denmark, 2800 Kongens Lyngby, Denmark.

[3] Novo Nordisk Foundation Center for Protein Research, Faculty of Health and Medical Sciences, University of Copenhagen, DK-2200 Copenhagen, Denmark.

*To whom correspondence should be addressed.

Abstract

Motivation: Understanding the specificity of protein receptor-ligand interactions is pivotal for our comprehension of biological mechanisms and systems. Receptor protein families often have a certain level of sequence diversity that converges into fewer conserved protein structures, allowing the exertion of well-defined functions. T and B cell receptors of the immune system and protein kinases that control the dynamic behaviour and decision processes in eukaryotic cells by catalysing phosphorylation represent prime examples. Driven by the large sequence diversity, the receptors within such protein families are often found to share specificities although divergent at the sequence level. This observation has led to the notion that prediction models of such systems are most effectively handled in a receptor-specific manner.

Results: We show that this approach in many cases is suboptimal, and describe an alternative improved framework for generating models with pan-receptor predictive power for receptor protein families. The framework is based on deep artificial neural networks and integrates information from individual receptors into a single pan-receptor model, leveraging information across multiple receptor-specific data sets allowing predictions of the receptor specificity for all members of a given protein family including those described by limited or no ligand data. The approach was applied to the protein kinase superfamily, leading to the method NetPhosPan. The method was extensively validated and benchmarked against state-of-the-art prediction methods and was found to have unprecedented performance in particularly for kinase domains characterized by limited or no experimental data.

Availability and Implementation: The method is freely available to non-commercial users and can be downloaded at http://www.cbs.dtu.dk/services/NetPhospan-1.0.

Contact: mniel@bioinformatics.dtu.dk

Supplementary information: Supplementary data are available at Bioinformatics online.

1 Introduction

A plethora of biological functions and processes are controlled by receptor-ligand interactions. Conventionally, extensive lab-work combined with bioinformatics and machine-learning efforts have, in many cases, been employed to arrive at predictive models capturing the specificity of a given receptor-ligand system. The outcome of such analyses has revealed that receptors within a given protein family have evolved to obtained highly divergent specificities. This has primed the notion that each receptor represents a unique biological problem that best can be described in isolation. One such example is protein kinases that phosphorylate residues in other proteins; a process involved in the control of dynamic behaviour and decision processes in eukaryotic cells. More than 510 different protein kinases have been annotated within the human proteome (Manning, et. al. 2002), and many of these have been functionally characterized to have very different specificities. This high sequence and specificity diversity make it a very costly undertaking to experimentally characterizing a receptor superfamily, priming the need for bioinformatics models capable of predicting receptor specificities. An essential prerequisite for the development of bioinformatics prediction models is the availability of high quality and abundant data characterizing the given problem at hand. This is also the case for receptor-ligand systems. For the kinase system, data that covers both the phosphorylation site and phosphorylating kinase is scarce and only a minor fraction of the ~500 kinases described in the human genome have been characterized with phosphosite information (most of these with no more than 30 phosphosites identified, see Supplementary figure 1). This data scarcity makes the training of machine learning methods for prediction of phosphorylation sites challenging since the volume of data is critical to achieving an acceptable performance for such data-driven approaches.

Earlier work has thus typically either taken a receptor-generic (describing the average specificity of the protein family, adapted for instance in the early kinases phosphorylation method NetPhos (Blom, et. al. 2004), where the data available at the time made it impossible to create kinase-specific models) or receptor-specific (describing each receptor on its own) approach to address the issue of characterizing the specificities of different receptor protein families including kinases (examples include kinase phosphorylation (Blom, et. al. 1999), SH2 (Hjerrild, et. al. 2004), SH3 (Gao, et. al. 2010) and PDZ (Obenauer, et. al. 2003) domains). Work within the Major Histocompatibility Complex (MHC) receptor system of the immune system has revealed that this approach in many cases suboptimal, and that the characterization of receptor specificities within large protein families can benefit from taking a holistic approach, integrating information from multiple individual receptor data sets into a single pan-receptor framework (Hoof, et. al. 2009 and Karosiene, et. al. 2013).

Here, we illustrate how such a pan-receptor approach can readily be extended to any receptor protein superfamily system, using the protein kinase family as an example. We here use the term receptor to describe protein binding domains capable of binding small peptide fragments, and we will interchangeably use the term kinase and kinase domain to refer to the catalytic protein kinase receptor. Protein phosphorylation requires the physical binding of a protein kinase to its ligand. After recognition of conserved linear protein segments (motifs), kinases catalyze the transfer of a phosphate from ATP to a hydroxyl group in a serine, threonine or tyrosine in the target protein. Most protein kinases share a common phylogenetic origin that is evidenced by a common structural core, the protein kinase domain. This common structural framework allows for the accommodation of a considerable degree of structural variance, as well as a wide range of sequence divergence and substrate specificities while preserving the conserved basic catalytic mechanism.

Decades of studies and the recent use of high throughput mass spectrometry have identified thousands of in vivo phosphorylation sites. Given this, prediction of kinase specificities becomes increasingly amenable for machine learning approaches. Inspired by the earlier work within the peptide-MHC system (Nielsen, et. al. 2007), we propose here a framework to develop pan-receptor prediction

models for any protein receptor superfamily using protein kinases as an example. By integrating properties contained within the kinase domain protein sequence with kinase-specific ligands into a machine-learning framework, we expect the approach to leverage information between different receptor molecules and thus enabling accurate predictions also in situations where limited or even no ligand data is available for a given receptor sequence.

Earlier work has described frameworks for predicting binding specificities for novel receptors within a given protein family by use of inference from receptors with experimentally characterized binding specificity driven by sequence similarity of substrate-determining residues of the receptor protein sequences (Creixell, et. al. 2015 and Zhang, et. al. 2009). Applying such approaches to the kinase system is however not trivial due to the vast sequence diversity of the kinase domain receptors, making alignment of kinase domain sequences unto a common framework highly challenging and error-prone. To solve this problem, we apply an alignment-free convolutional neural networks (CNN) machine-learning approach. Integrating this CNN representation of the receptor with the ligand information in a conventional feed-forward network, allow us to implement a predictor for peptide phosphorylation with true pan-receptor power, allowing kinase-specific predictions of peptide phosphorylation for any kinase domain of known protein sequence. We develop the method from data in the public domain and perform exhaustive benchmarking including cross- and leave-one-kinase-out validation, mutation analysis, performance comparison to other state-of-the-art methods, and identification of specificity defining positions in a given kinase domain.

2 Materials and Methods

2.1 Dataset construction

A data set of 21-mer peptides was retrieved from PhosphoSitePlus (Hornbeck, et. al. 2015) and Phospho.ELM (Dinkel, et. al. 2011). By training several generic phosphorylation-site predictors using datasets of peptides with lengths ranging from 7 to 27, an optimal performance for peptides of length 21 was observed and this length was maintained in this work without further optimization. Each peptide reports a known phosphorylation site in its central position (11th position). Peptides are accompanied by information on the specific kinases catalysing the phosphorylation reactions. Only phosphorylations attributed to the 478 human, eukaryotic, protein kinase (ePK) (Manning, et. al. 2002) were selected, leaving out 40 kinases corresponding to the 13 atypical kinase families (aPK). Filtering out receptors represented with less than 10 data points, we arrive at a final positive dataset containing 10,344 peptides (originating from both kinases and other proteins) covering 154 different protein kinases. In figure 1, we show how the protein kinases in the datasets are distributed over the phylogeny of the (ePK) superfamily confirming a broad coverage of the taxonomy. Only one branch belonging to the "Other" group (shown in blue, in between CKI and TK groups) is left uncovered.

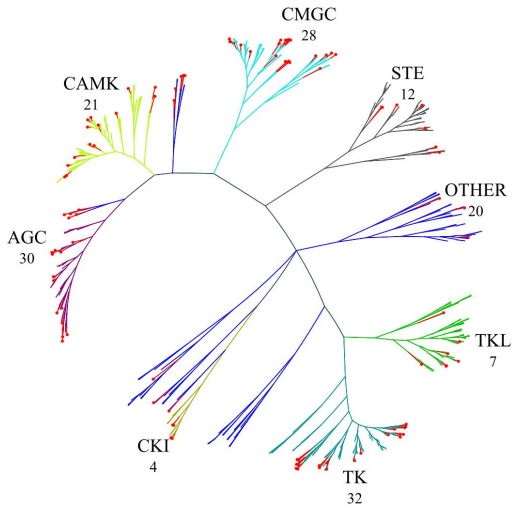

Figure 1. Phylogeny of the eukaryotic protein kinase (ePK) superfamily. Each of the major groups is shown in different colours. Under the group names are denoted the number of kinases included in the training set that belong to that group (high lighted as red leaves with circles). The tree structure was retrieved from KinBase[1] and plotted using HyperTree (Bingham, et. al. 2000).

A negative set was constructed by compiling all 21-mer peptides with a central S/T/Y from the source proteins of the positive dataset not reported as phosphorylated in any of the two databases. This negative dataset consisted of

122,974 negative peptides also covering the 154 different kinases. The kinase domain of each kinase was retrieved from KinBase (Manning, et. al. 2002). The whole dataset was divided into five partitions to fit a typical five-fold cross-validation scheme (CV) where 4/5 of the data are used for training and the remaining 1/5 is used for testing. To address data redundancy, peptides with identity >= 50% were grouped in the same partition. Peptides with that shared identity with elements in more than one partition were left out of the data set (for details on the data and data partitioning see Figure 2).

The amino acid sequence of both peptides and protein kinase domains were represented using Blosum encoding (Henikoff, et. al. 1992); each amino acid was encoded as a vector of twenty elements from the corresponding row of a Blosum50 substitution matrix. This encoding scheme allows the data to retain information about the relation between the different amino acids.

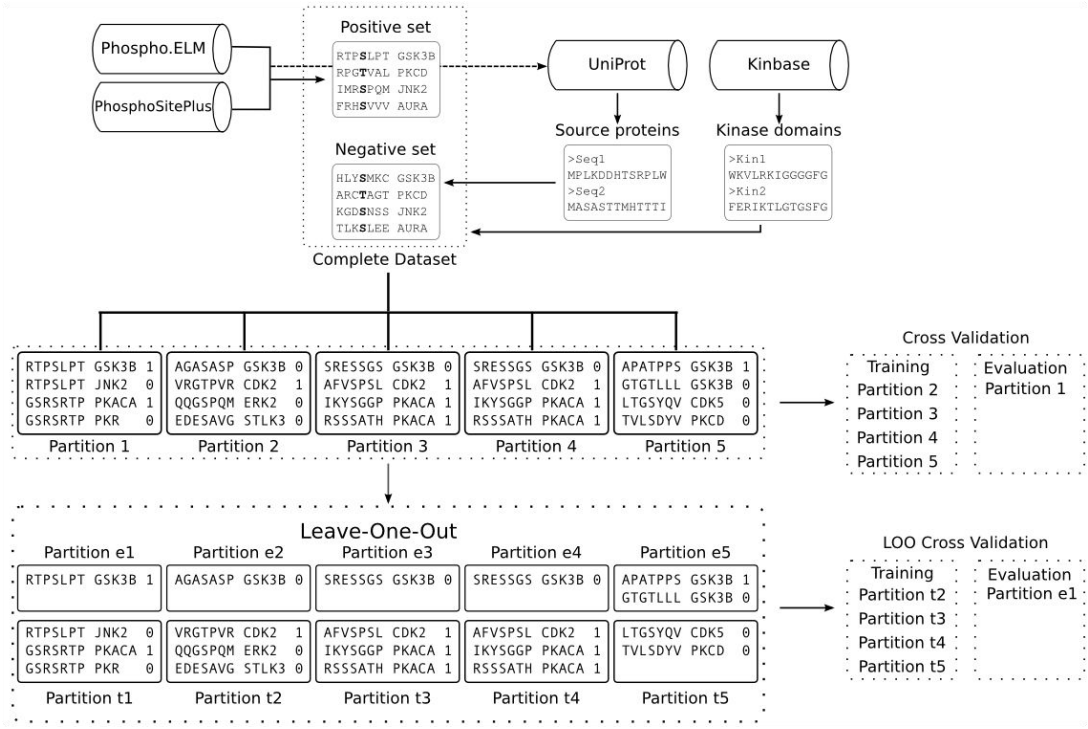

Figure 2. Dataset construction and Cross-validation set-up. Positive examples were taken from PhosphoSitePlus and Phospho.ELM, the negative set was constructed with the potential phosphosites in the positives source proteins not reported as such. The dataset was divided in 5 partitions grouping peptides with 50 or more %identity. The partitions were used in a typical five-fold cross-validation experiment, where four partitions are used in turn to train the network and the remaining one to evaluate its performance. In the Leave-One-Out experiment, every data point from a given kinase domain is removed from all partitions. The cross-validation is next performed using four partitions devoid of the data from the kinase domain in question as training data and evaluated on the kinase domain specific data from the remaining partition. This set-up ensures to exclude not only information from the left out kinase but also from similar peptides in the network training.

2.2 Artificial Neural Networks

2.2.1 Kinase-specific predictors

Kinase-specific predictors were trained individually for each kinase domain as a baseline to compare improvements induced using different strategies and architectures. To train these models, data were split by protein kinase, maintaining the original partitioning in the training set (see above). Networks were trained using stochastic gradient descent

back-propagation for 100 epochs in a typical cross-validation scheme without early stopping. In this way, the left-out evaluation data was independent of the training (see figure 2). The architecture was shared among all the kinase-specific networks having 420 input neurons (corresponding to 21 ligand positions each encoded with 20 digits), 50 hidden neurons and one output. Other more complex architectures and hyper-parameter settings were investigated but did not demonstrate a consistent performance gain (data not shown).

2.2.2 The pan-receptor kinase predictor

A single pan-receptor predictor was developed to allow prediction of phosphorylations specific to any kinase domain of know protein sequence. Inspired by the pan-specific predictor developed for peptide MHC binding (Hoof, et. al. 2009, Karosiene, et. al. 2013), the pan-receptor kinase predictor was constructed combining the sequence of the complete kinase domain of the kinase attributed the phosphorylation with the peptide sequence for each data point. This was achieved by representing the kinase domain as a 1D-convolutional layer followed by a pooling layer. The convolution was obtained using 40 filters of size 3, 5 and 7 amino acids leading to a total of 120 filters. The pooling layer performs a global max-pooling, selecting the maximum value for each filter and these maximum values were then used as input together with the encoded peptide sequence (encoded with 420 input values as for the kinase-specific networks) in a feed-forward layer with 90 neurons (see figure 3c) connected to the output neuron. The model was trained in a five-fold cross-validation scheme using stochastic gradient descent back-propagation for 100 epochs, with a batch size of 20, a learning rate of 0.05 and a sigmoid activation function for all neurons including the output neuron. This means that the model output is quantitative (and not qualitative/binary). Weights were initialized using a Gaussian distribution with a mean of 0 and a standard deviation of 0.01. All hyper-parameters, except the number of filters, were chosen from experience in previous works (Jurtz, et. al. 2017a, Jurtz, et. al. 2017b) and were not optimized to this particular case. The implementation was done using the Lasagne library (Dieleman et. al. 2015), which manages Theano (Al-Rfou et. al. 2016) to build and train neural networks.

2.3 Validation

2.3.1 Overall performance

The performance of the different methods was estimated from the combined test set predictions; the test predictions of all the partitions were concatenated and the area under the receiver operating characteristic curve (AUC) was calculated for each individual kinase. The comparison between different methods was done by applying a binomial test without ties, where the highest AUC determines the best performing method.

2.3.3 Leave-one-out

To estimate the performance of the pan-receptor predictor for kinase domains not included in the training, a leave-one-kinase-out (LOO) experiment was conducted. Here, a set of pan-receptor predictors was trained, each by removing all data points for a specific kinase domain in each partition. The predictor was next trained using 5-fold cross-validation on the remaining data, as described above. Each left-out partition was next predicted; i.e. left-out peptides from partition 0 were predicted from an LOO network trained on partitions 1, 2, 3 and 4, etc. (for details on the LOO experiment setup refer to figure 2). This strategy is naturally not applicable to the kinase-specific method. Here, predictions for the left-out kinase domain were performed using a nearest neighbour approach. The nearest neighbour was identified from the sequence similarity of the given kinase domain to the other receptors in the dataset using a blosum50 based alignment and translated to distance as one minus the sequence identity divided the alignment length, and the distance to the nearest neighbour was calculated as 1 – nid/alen, where nid is the

number of identical amino acids in the alignment and alen the alignment length. Next, the nearest neighbour kinase-specific model was used to make predictions for the data of the left-out domain in each partition using the network trained without its partition of origin, as was the case for the pan-receptor LOO evaluation (for details see Figure 2).

2.3.4 Receptor mutation analysis

This experiment was set up to identify positions in the kinase domain that the network identifies as important. We performed an alanine scan where every amino acid in the receptor sequence was mutated to alanine and predictions were made for a set of thousand random phosphorylatable peptides in natural protein sequences. The selection of random ligands instead of high binders was used to measure both gain and loss in predicted score in comparison with the predictions using the wild-type molecule. Comparing the predictions from the mutated with the native kinase sequence a score for each position in the kinase domain was calculated as:

$$Score = \sum |S_{wt} - S_m|$$

where S_{wt} is the prediction with the wild-type kinase domain sequence and S_m is the prediction with its mutated version. This score was mapped to the structure of the kinase to visualize the location of relevant positions.

The structures of the kinase domains used in the alanine scan assay were retrieved from the protein data bank (Berman, et. al. 2000).

2.3.5 Ligand mutation analysis

To validate the performance of the pan-receptor kinase predictor in recognizing the effect of mutations on the phosphorylation target sequence, peptides with an identified mutation that resulted in either loss or gain of phosphorylation were retrieved from Reimand *et. al.* 2013. In each case, predictions were made for both wild-type and mutated peptide. The assigned score to each pair was calculated as S_m-S_{wt} and a scatter plot of these values plotted to visualize the changes induced by the mutations.

2.3.6 Model comparisons

The performance of NetPhosPan, was compared to GPS (Xue, et. al. 2008), Musite (Gao, et. al. 2010), NetPhos-3.1 (Blom, et. al. 1999 and Blom, et. al. 2004) as well as the recently published, MusiteDeep method (Wang, et. al. 2017). These methods use different strategies to solve the same problem, GPS is a kinase-specific phosphorylation site predictor implemented in JAVA based on similarity-based clustering that covers 464 protein kinases, Musite, on the other hand, uses multiple support vector machines trained on protein disorder, local sequence similarity and amino acid frequencies around sites phosphorylated as features to address 67 kinases, NetPhos-3.1 is based on feed-forward neural networks trained with peptides containing the phosphorylated amino acid at the central position and is trained with 17 different kinase substrates, and finally, MusiteDeep is a kinase-specific predictor based on deep convolutional neural networks trained on data spanning 5 kinase families, covering 61 kinases. It takes as input 33-mers peptides with a central potential phosphosite. Default options of the standalone versions of each tool were used. The only modifications were for Musite where the model *H.sapiens.kinase.specific* was selected in order to use its human kinase-specific version. The data for the comparison was retrieved from PhosphositePlus, excluding all examples included in the training data of NetPhosPan and MusiteDeep. The latter information was obtained from github.com/duolinwang/MusiteDeep/tree/master/testdata.

Given that this data was published after the release of the first three tools used in this benchmark, we consider this a fair comparison in relation to NetPhosPan. Since GPS, Musite, MusiteDeep and NetPhos are all kinase-specific methods, only kinases shared by at least three methods were tested. The resulting subset featured seven kinases with only two shared between all four predictors. The metrics

used for comparison purposes were AUC, AUC0.1 (area under the ROC curve integrated up to a false positive rate of 10%), predictive positive value, sensitivity and specificity. The latter two requires classification of the different prediction into positive and negatives. For this default values for each method were used. P-values associated with each comparison were estimated using Bootstrap analyses. Here, one thousand re-samples of the original benchmark dataset, allowing replacement, were constructed (requiring the dataset to have at least one data point of each positive and negative class). Next, the p-value for one method outperforming the other was estimated as the percentage of re-sampled datasets where the other method had superior performance.

3 Results

Protein kinases are a prime example of a receptor superfamily displaying a very high diversity in binding site specificity. This diversity is manifested in the different binding motifs of individual kinase domains derived from kinase-specific ligand data. A few such examples are shown in figure 3A. The specificity of a kinase domain is as a first approximation encoded in its primary protein sequence. By integrating the kinase domain protein sequence with kinase-specific ligands in a machine-learning framework, one should, therefore, be able to construct a pan-receptor predictor, leveraging information between different kinase domain specificities, allowing for accurate phosphorylation predictions for any kinase of interest.

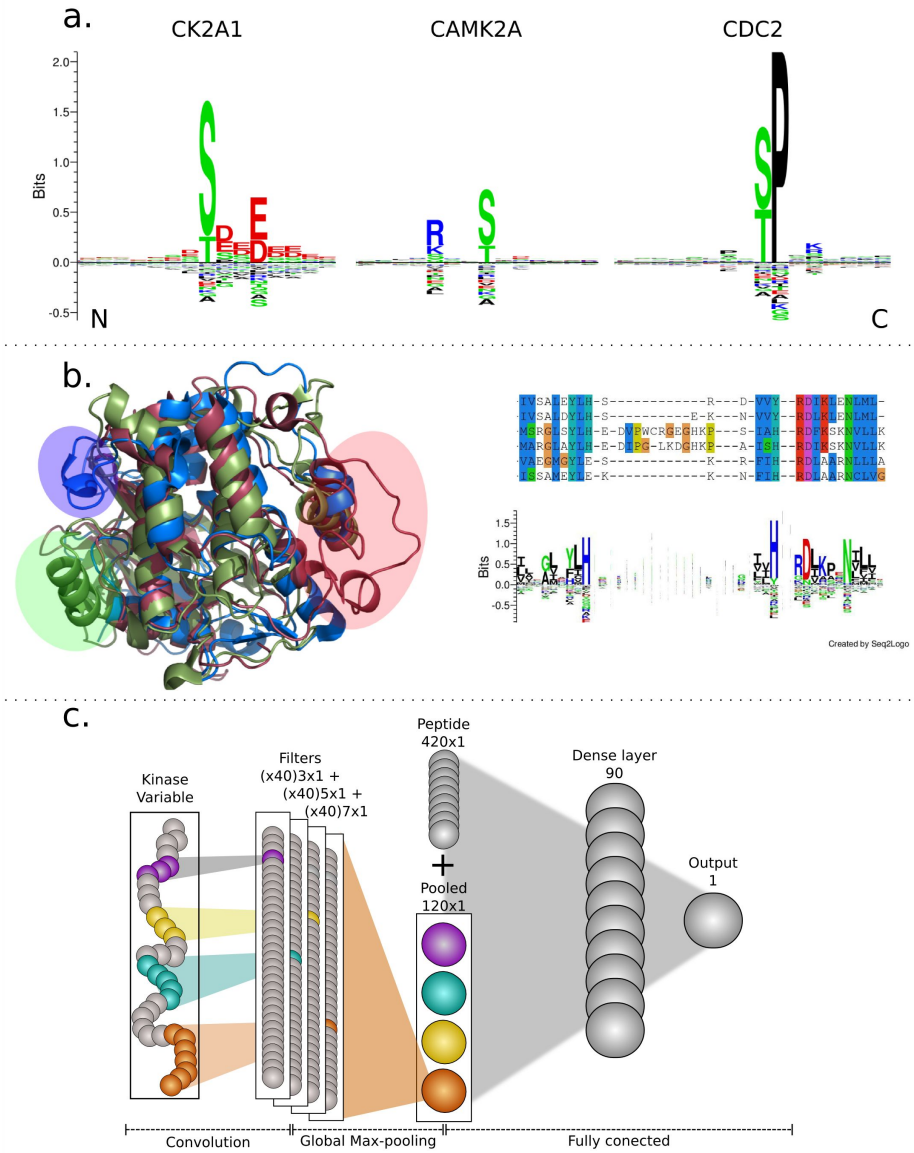

Figure 3. (a) Different kinases show specific preferences in the sequence that surround the phosphorylated amino acid, this preference is usually maintained within kinase families. (b) Illustration of the main challenge of the pan-receptor approach. The right panel shows two representations of the diversity in kinases at the sequence level, a small window taken from the alignment of 478 human kinase domains, containing the catalytic aspartic acid. Regions of insertions and deletions can be found along the entire alignment. In the logo, the thickness of each character depends on the number of gaps at the given positions. The left panel shows the structural differences of three kinases (CDC42BPB (DMPK), VRK3 (VRK), and CDK2 (CDK)). The coloured circles present a particular structure specific to each individual kinase. (c) The architecture of the artificial neural network used to train NetPhosPan. The catalytic domain is processed by a convolutional layer that applies several filters in order to extract a fixed number of features. The encoded peptide is concatenated to the convolutional layer output and used as input to a conventional feed-forward layer. The output neuron predicts a likelihood of whether the given residue in the input peptide is phosphorylated. Note, therefore, that the method is not a classifier.

3.1 NetPhosPan performance

Like earlier pan-specific prediction methods (Hoof, et. al. 2009, Karosiene, et. al. 2013, Andreatta, et. al. 2015, Creixell, et. al. 2015 and Nielsen, et. al. 2016), the pan-receptor NetPhosPan phosphorylation predictor takes into account not only information contained within the ligand, but also incorporates information from the kinase domain of the receptor. To address the challenge of sequence and structural diversity of the kinase domains illustrated in figure 3B, the receptor sequences are represented as max-pooled convolutional neural networks (CNN). In essence, these networks consist of filters (linear models) trained to capture short linear motifs in the input data. Using max-pooling conforms each convolutional filter to contribute with one value to the model. Having a fixed number of CNN filters, all receptors are transformed into an equal-sized feature space (equal to the number of filters), that in turn is combined with the ligand, allowing the method to infer the rules relating the kinase receptor sequence to its functional specificity (see figure 3C and details in Materials and Methods). As observed for the MHC system, this pan-receptor predictor should be capable of leveraging information between kinase domains and achieve improved performance compared to conventional receptor-specific methods trained on ligand data specific for single kinase domains. This in particular in situations where ligand data are either scarce or completely absent. The model was trained on 10,344 positive phosphorylated peptides covering 154 different protein kinases (for details on the data and method see Materials and Methods). Figure 4A reports the predictive performance of the model and confirms that this is indeed the case. Here, the performances of the pan-receptor and kinase-specific methods are compared (details are given in Table S1) and show that the pan-receptor method significantly outperforms the kinase-specific version. This is the case for kinase domains regardless of the amount of ligand data available. However, the gain is (as expected) most pronounced when the number of ligands is small (p-value < 0.001, binomial test comparing the performance of the two methods for kinase domains characterized by fewer than 30 positive data points). The performance of the pan-receptor predictor was further compared to a naïve sequence similarity-based predictor where for each peptide, the target value (phosphorylated or not-phosphorylated) was inferred from the target value of the most similar peptide in the training data for the given kinase. As expected, the similarity-based model had a very poor predictive performance with an average of 0.709 of AUC (compared to an average performance of the pan-specific method of 0.878, see supplementary Table S1) confirming that the data partition adequately deals with the redundancy in the ligand dataset and that the method has captured signals allowing for extrapolation beyond a simple lookup table.

3.2 Exposure analysis

A residue must be accessible to the kinase in order to be phosphorylated. NetPhosPan, however, does not explicitly incorporate any information related to the surface exposure of the residues predicted to be phosphorylated, and one could speculate that the method learns to discriminate between buried and exposed residues rather than phosphorylation. To assess to what degree this was the case, the performance was evaluated on the subset of exposed potential phosphorylation sites, predicted by NetSurfP-1.1 (Petersen, et. al. 2009). In every case, the performance loss is less than 5% (supplementary table S5), confirming that the captured signal describes the phosphorylation potential of a given residue, rather than the exposed/buried condition of each peptide.

3.3 Leave-One-Out performance

One of the objectives of NetPhosPan is to predict phosphorylation of peptides by kinases not characterized by ligand data. In order to emulate this situation, a leave-one-out (LOO) experiment was performed. Here, methods were trained using the setup described earlier for the pan-receptor method, each time excluding all data of a single kinase from the training data. Next, the data from the left-out kinase domain was used as an evaluation set

to measure the performance of the method. As a comparison, a nearest neighbour approach was used. Here, the data set of a given kinase domain was predicted using the receptor specific network trained on data from the most similar other kinase domain (for details on the LOO and nearest neighbour approaches see material and methods). The result of the LOO experiment is summarized in figure 4B and shows across all distances to the nearest neighbour a clear improved performance of the pan-receptor predictor over the nearest neighbour approach (p-value = 0.004, binomial test excluding ties over all data points).

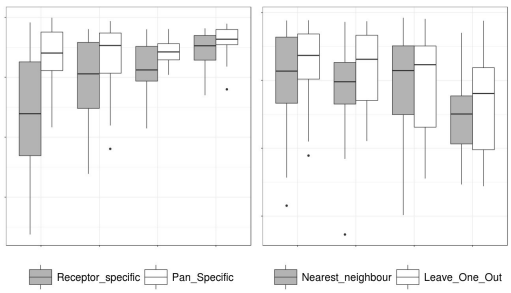

Figure 4. Performance of the different methods. The kinases were grouped according to the number of positive data points in the dataset and the AUC values of each group were plotted for each method. (a) Performance of NetPhosPan and the kinase-specific methods. The pan-receptor method outperforms the kinase-specific method in every group with a large improvement when there are only a few data points. (b) Comparison of leave-one-out and nearest neighbour methods in novel data prediction. The leave one out experiment shows an improvement when compared to the nearest neighbour.

To illustrate why the LOO outperformed the nearest neighbour method, we visualized the binding specificity of two kinases as sequence logos (figure 5). The logos were constructed using Seq2logo (Thomsen, et. al. 2012) from the top 1% highest scoring peptides from a set of 100,000 random natural peptides (each containing S or T at the central position) predicted by the two prediction methods. The two kinases were selected as one with a close nearest neighbour (JNK1), and one with a far nearest neighbour (CHK1). From the plots, it is clear that both methods succeed at predicting the specificity of JNK1. For CHK1 on the other hand, the nearest neighbour motif shows a clear deviation from the experimental data in particular at the C-terminal side of the phosphorylation site where this to a much lesser degree is the case for the LOO method. This thus confirming that the pan-receptor method is capable of leveraging information from the different kinase receptors, and in turn use this information to make accurate predictions for novel data.

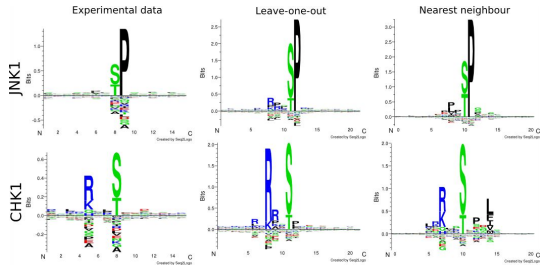

Figure 5. Sequence logo representation of kinase specificities of JNK1 and CHK1. The left column gives the logo derived from the experimental data, the central column the logo derived from the LOO predictions, and the right column the logo derived from the nearest neighbour predictions. Logos were estimated as described in the text.

In summary, NetPhosPan was demonstrated to have a consistent and high predictive performance with AUC values above 0.9 within all kinase groups (see supplementary table S1), and an average AUC over all kinases included in the benchmark of 0.878. When compared with a receptor-specific predictor, the method showed comparable performance when the training data was abundant and a significant improvement in cases when only a few data points were available. This and the robust performance showed in the leave-one-out experiment, suggests that the cross-learning between different receptors allows the method to infer ligand-receptor relations boosting its performance beyond that obtained by the receptor-specific models. Further, filtering out data points predicted as buried, we could prove that the signal contained within NetPhosPan was indeed related to phosphorylation and not just to the

identification of amino acid compositions found in exposed regions of the protein of interest.

3.4 Ligand mutation analysis

Mutations in a ligand can lead to a gain or loss of phosphorylation capacity. This event is tightly related to modifications of the phosphorylation pathways and signalling network rewiring often observed in diseases with a high mutational load such as cancer. To investigate if the NetPhosPan method can be used to predict the impact on phosphorylation of such single mutation variants, a set of ligands with experimentally confirmed alternation in phosphorylation between wild-type and mutant variant were retrieved from an independent dataset published by Reimand *et. al.* 2013. Peptides with 100% identity to any peptide in the training set were discarded. Predictions were performed using NetPhosPan for both wild-type and mutated versions of these ligands, and the predicted change was compared to the experimentally observed alternation. The results of the analysis are shown in figure 6, where ligands that lose or gain phosphorylation are analysed separately. The analysis shows that NetPhosPan to a very high degree (in particular for the ligands that lose phosphorylation) is capable of predicting loss of phosphorylation upon mutation. The relatively poor performance of NetPhosPan for predicting the events of gained phosphorylation is to a high degree explained by a very poor correlation between the motives of the specificities of the given kinase in question (as reported in both PhosphositePlus, and the training data, and derived from NetPhosPan) and the nature of the mutation that should induce the increased likelihood of phosphorylation. A striking example of this includes SPSQLSKW[P/S]GSPTSRSSDELD where the mutant variant [S] is annotated to gain phosphorylation to the kinases ERK1 and ERK2 when compared to the wildtype [P] (highlighted with void circles in figure 6). These kinases are both described to have a preference for proline (P) at the P-2 position (two amino acids to the N terminal side of the phosphosite). It is not clear what is the source of these seeming inconsistencies.

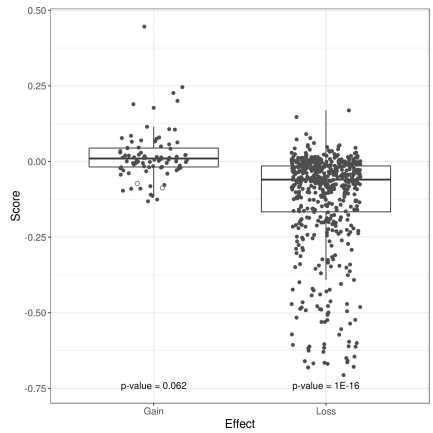

Figure 6. Mutation analysis of ligands. The score variation between wild-type and mutated peptides is plotted for both groups: peptides that lose its capacity of being phosphorylated (Loss) and peptides which become prone to interact with a kinase (Gain). A negative ΔScore means a drop in prediction score from the wild-type ligand to its mutated version corresponding to a loss of phosphorylation potential while a positive ΔScore corresponds to the opposite effect.

3.5 In-silico kinase mutations

Having demonstrated that the NetPhosPan predictor has pan-receptor potential and can be applied to gain insights to the impact of mutations in phospho-ligands, we next investigated to what degree the signal captured by the method matches properties of the protein structure and function of the kinase domains. To assess this, an alanine scan was performed. Here, each amino acid in a given kinase domain protein sequence was mutated to alanine and the mutant variant used to predict the phosphorylation of a thousand random natural peptides (containing S or T at the central position). Next, a score was calculated for each kinase domain position measuring the variations in predicted values between wild-type and mutant domain (for details see materials and methods) and this score was mapped to the domain 3D protein structure, lighting positions where the difference in prediction was high. The result of these analyses is shown in figure 7 for two kinase domains and reveals that most of the positions with the higher variation score were found in the vicinity of the active site in the kinase domain. These results suggest that

the machine-learning framework in a fully automated and unsupervised manner has captured essential biologically and functionally relevant information stored in the kinase domain sequence.

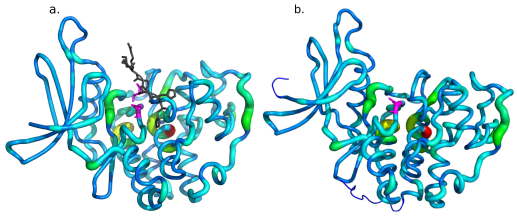

Figure 7. Alanine scan experiment. Each amino acid in the kinase domain of the kinase domain from CDK2 (a) and ROCK1 (b) molecules is coloured (in a scale from red to blue, with red being high) and enlarged using the prediction variation score. In the centre, the conserved aspartic acid, which is important for the catalytic activity of the enzyme is shown in magenta. CDK2 domain is in complex with a peptide (in dark grey), its phosphosite is highlighted in magenta. (PDB ID: 1QMZ and 2ETR, respectively).

3.6 Method comparison

To compare the predictive power of NetPhosPan, the method was challenged against a panel of five widely used phosphorylation site predictors (for details see Materials and Methods). Only kinases covered by at least three methods were included in the benchmark. The scores predicted for each potential phosphosite were used to calculate the performance values of each method. Reported phosphosites were tagged as positives while every other potential phosphorylatable amino acid was taken as negative. Since most of these methods only report predicted phosphorylation sites, every non-reported site was assigned a prediction score of zero. (for details on the benchmark and performance corresponding measures see materials and methods). Table 1 and 2 display the performance values (AUC and PPV) for each method on the different datasets. Here, AUC was calculated as described earlier, and PPV (predicted positive value) was calculated as the proportion of true positives among the top N highest scoring predictions, where N is equal to the number of positive peptides in the given data set. Supplementary tables S2, S3 and S4 give the sensitivity, specificity and AUC0.1 values of each method for each dataset.

Kinase	Training	AP	AN	GPS-3.0	NetPhos-3.1	Musite-1.0	MusiteDeep	Kinase-specific	NetPhosPan
PKACA	715	25	1667	0.720	0.854	0.599	**0.867**	0.860	0.846
CDK1	509	318	18768	0.899	0.428	0.839	0.922	0.931	**0.947***
ERK1	339	10	623	0.871	-	0.825	0.876	0.887	**0.975**
ERK2	441	15	698	0.950	-	0.855	**0.988**	0.986	0.986
PKCB	68	51	2314	-	0.735	0.736	**0.955***	0.864	0.912
P38A	182	8	626	0.827	-	-	0.836	0.911	**0.941**
CK2A1	541	17	761	0.897	-	-	**0.934**	0.849	0.874

Table 1. Comparison of the predictive performance of the different methods included in the benchmark. Training gives the number of positive ligands for the given kinase in the training data, AP and AN is the number of positive and negative data in the evaluation data set. The 6 methods included in the benchmark are GPS-3.0 ,(Xue, et. al. 2008) NetPhos-3.1 (Blom, et. al. 1999 and Blom, et. Al. 2004), Musite-1.0 (Gao, et. al. 2010), MusiteDeep (Wang, et. al. 2017), the kinase-specific method described here, and NetPhosPan. The AUC value is calculated as described in the text assigning phosphosites as positives while all other potential phosphorylatable peptides were taken as negatives. In bold, is denoted the highest performance value achieved for a given kinase. Results denoted with (*) resulted statistically significant ($p<0.05$) in a bootstrap analysis performed between the two methods (MusiteDeep and NetPhosPan) with the best overall performance.

Kinase	Training	AP	AN	GPS-3.0	NetPhos-3.1	Musite-1.0	MusiteDeep	Kinase-specific	NetPhosPan
PKACA	715	25	1667	**0.32**	0.28	0.24	0.28	**0.32**	0.28
CDK1	509	318	18768	0.20	0	0.29	0.26	0.31	**0.37***
ERK1	339	10	623	0.30	-	0.10	0.20	0.20	**0.40**
ERK2	441	15	698	0.40	-	0.53	**0.67**	0.47	0.53
PKCB	68	51	2314	-	0.18	0.31	**0.43**	0.39	0.37
P38A	182	8	626	**0.13**	-	-	**0.13**	**0.13**	0.00
CK2A1	541	17	761	0.53	-	-	**0.65***	0.53	0.12

Table 2. Positive predictive value (PPV) for each kinase in the benchmark. Displayed are the number of positive data points in the training set (Training), the number of positive and negative peptides in the benchmark (AP and AN respectively) and the PPV obtained by each method, calculated as the proportion of true positives in the subset of size AP of best scoring peptides. In bold, is denoted the highest performance value achieved for a given kinase. Results denoted with (*) resulted statistically significant ($p<0.05$) in a bootstrap analysis performed between the methods with the best overall performance.

In summary, these results demonstrated suggest a comparable or superior predictive performance of NetPhosPan compared to the other methods included in the benchmark. Note, that all the kinases included are characterized by more than 65 positive data points in the training dataset hence allowing a high performance of receptor-specific methods, and thus not fully capturing the predictive potential of NetPhosPan. This is also reflected in the fact that the kinase-specific predictor developed here and included in the benchmark demonstrates a performance comparable to that of NetPhosPan. To fully appreciate the predictive power of the proposed method, the result of this benchmark must be aligned with the results of the leave-one-out experiments described earlier (in particular figure 2B). Such an alignment suggests that the high performance of the NetPhosPan method observed here can be extrapolated to the complete space of kinases; an ability that is not covered by the other methods included in the benchmark.

4. Discussion

We have here outlined a framework enabling the development of pan-receptor models capable of predicting the ligand-binding specificity for any member of a receptor protein family with a known protein sequence. The only prerequisite for a successful application of the approach is the availability of a representative set of receptor-specific ligand data. The proposed framework was used here to develop a pan-receptor model for prediction of the ligand specificity of protein kinases.

The developed prediction method is as the name suggests pan-specific, meaning that the method can make predictions of phosphorylation for any kinases domain characterized by a protein sequence. Previous methods capable of inferring kinase domain specificity of novel kinases have been described earlier (Brinkworth, et. al. 2006 and Creixell, et. al. 2015). However, these methods all rely heavily on a multiple sequence alignment of the different kinase domains to identify substrate-determining residues and use this information to predict binding preferences for kinases with no available ligand data. Such alignment-based methodologies are highly error-prone when working with sequence diverse protein families. Usually, multiple alignments should be manually reviewed to correct errors which prevent the development of fully automatic methods that can predict over new sequences i.e. mutated versions. In contrast, is the framework underlying NetPhosPan alignment-free and allows in a fully automated manner to identify shared patterns in the kinase domains avoiding any kind of

alignment-related error. The method was benchmarked using cross-validation, leave-one-kinase-out and against other publicly available phosphorylation predictors, and was in all cases found to perform at par or better than all methods tested. In particular, in the leave-one-kinase-domain-out validation, the method was found to maintain high performance also for kinase domains characterized by limited or no ligand data point. This is a unique property of NetPhosPan since all other available methods are kinase-specific and hence limited to make predictions for the small subset of kinases characterized by a sufficiently large set of ligands.

In cancer, a common example of a complex disease, cells present hypermutated phenotypes where somatic mutations accumulate rapidly (Nebot-Bral, et. al. 2017). Furthermore, a common observation in the clinical treatment of cancer is that patients rapidly develop resistance to treatment. These resistance mechanisms often imply the generation of new specificities that rewire the canonical protein-protein network and allow for the generation of alternative escape pathways. Using a large set of experimentally peptides containing validated phosphorylation inducing and phosphorylation depleting mutations, we illustrate how NetPhosPan could be used to aid the understanding of such mutation-driven rewiring of the phosphorylation network. Comparing wild-type and mutant variants, we observed a consistent increase in the predicted phosphorylation potential for the peptides containing phosphorylation inducing mutations and likewise a decrease in phosphorylation potential for peptides containing mutants experimentally found to abolish phosphorylation. These results suggest that the proposed method can be applied to predict the acquisition of novel kinase-substrate interactions that would deviate from the canonical phosphorylation signalling pathways.

To further illustrate how NetPhosPan can be used to guide the understanding of kinase specificity, we conducted a mutation analysis on the kinase domains. Doing this, we could identify specificity defining positions in a given kinase domain. This observation thus further points to that the method also from the kinase receptor point of view, can be used to identify mutations that could modify the kinase specificity and hence potentially lead to modifications of the phosphorylation pathways and signalling network rewiring.

In conclusion, we have developed the first pan-receptor predictor of protein phosphorylation. The method has been benchmarked and demonstrated to be state of the art. The framework described for developing the method was here applied solely to kinase phosphorylation, however, this represents one of the hardest problems to describe due to its large sequence and structural diversity. Given the high success of the obtained model, we find it very likely that the framework can be readily applied to any other receptor-ligand interaction system and could, in our view, form the cornerstone for future developments of receptor-ligand prediction models related to most of the essential regulatory processes in cellular organisms.

NetPhosPan is publicly available at www.cbs.dtu.dk/services/NetPhospan-1.0.

Author Contributions

M.N. conceived and designed the project. E.F. prepared the dataset, trained and evaluated the networks, and tested the method. M.N. and E.F. analyzed the results and wrote the manuscript with the support of J.M.G.I., V.J. and S.B.

Funding

This work was supported by Novo Nordisk Foundation [NNF17OC0027594 and NNF14CC0001 to SB] and the Innovation Fund Denmark [5184-00102B to JMGI].

Competing financial interest

The authors declare no competing financial interests.

References

Andreatta, M., et. al. (2015) Accurate pan-specific prediction of peptide-MHC class II binding affinity with improved binding core identification. Immunogenetics 67, 641-650.

Al-Rfou R. et. al. (2016) Theano: A Python framework for fast computation of mathematical expressions. arXiv preprint arXiv:1605.02688.

Berman, H.M. et. al. (2000) The Protein Data Bank. Nucleic Acids Res. 28, 235-242.

Bingham, J. & Sudarsanam, S. (2000) Visualizing large hierarchical clusters in hyperbolic space. Bioinformatics 16, 660-1.

Blom, N., Gammeltoft, S. & Brunak, S. (1999) Sequence- and structure-based prediction of eukaryotic protein phosphorylation sites. J. Mol. Biol., 294, 1351-62.

Blom, N., Sicheritz-Ponten, T., Gupta, R., Gammeltoft, S. & Brunak, S. (2004) Prediction of post-translational glycosylation and phosphorylation of proteins from the amino acid sequence. Proteomics, 6, 1633-49.

Brinkworth, R.I., Munn, A.L., Kobe, B. (2006) Protein kinases associated with the yeast phosphoproteome. BMC Bioinformatics, 7:47.

Creixell, P. et. al. (2015) Unmasking Determinants of Specificity in the Human Kinome. Cell 163, 187–201.

Dieleman S. et. al . (2015) Lasagne: First release. Zenodo. 10.5281/zenodo.27878

Dinkel, H., et. al. (2011) Phospho.ELM: a database of phosphorylation sites--update 2011. Nucleic Acids Res. 39, 261-7.

Gao, J., Thelen, J.J., Dunker, A.K. & Xu, D. (2010) Musite, a Tool for Global Prediction of General and Kinase-specific Phosphorylation Sites. Mol. Cell. Proteomics, 9, 2586 –2600.

Henikoff, S. & Henikoff, J.G. (1992) Amino acid substitution matrices from protein blocks. Proc. Natl. Acad. Sci. USA 89, 10915–10919.

Hjerrild, M., et. al. (2004) Identification of phosphorylation sites in protein kinase A substrates using artificial neural networks and mass spectrometry. J. Proteome. Res., 3, 426-33.

Hoof, I., et. al. (2009) NetMHCpan, a method for MHC class I binding prediction beyond humans. Immunogenetics 61, 1-13.

Hornbeck, P.V., et. al. (2015) PhosphoSitePlus, 2014: mutations, PTMs and recalibrations. Nucleic Acids Res. 43, 512-20.

Jurtz, V. et. al. (2017a) NetMHCpan-4.0: Improved Peptide–MHC Class I Interaction Predictions Integrating Eluted Ligand and Peptide Binding Affinity Data. J Immunol 199 (9) 3360-3368.

Jurtz, V.I. et. al. (2017b) An introduction to deep learning on biological sequence data: examples and solutions. Bioinformatics 22, 3685–3690.

Karosiene, E., et. al. (2013) NetMHCIIpan-3.0, a common pan-specific MHC class II prediction method including all three human MHC class II isotypes, HLA-DR, HLA-DP and HLA-DQ. Immunogenetics 65, 711-24.

Manning, G., Whyte, D.B., Martinez, R., Hunter, T. & Sudarsanam, S. (2002) The protein kinase complement of the human genome. Science, 298, 1912-34.

Nebot-Bral, L., et. al. (2017) Hypermutated tumours in the era of immunotherapy: The paradigm of personalised medicine. Eur. J. Cancer 84, 290-303.

Nielsen, M. & Andreatta, M. (2016) NetMHCpan-3.0; improved prediction of binding to MHC class I molecules integrating information from multiple receptor and peptide length datasets. Genome Med. 8, 33.

Nielsen, M., et. al. (2007) NetMHCpan, a Method for Quantitative Predictions of Peptide Binding to Any HLA-A and -B Locus Protein of Known Sequence. PLoS ONE 2, e796..

Obenauer, J.C., Cantley, L.C. & Yaffe, M.B. (2003) Scansite 2.0: proteome-wide prediction of cell signaling interactions using short sequence motifs. Nucleic Acids Res., 31, 3635–41.

Petersen, B., Petersen, T.N., Andersen, P., Nielsen, M. and Lundegaard, C. (2009) A generic method for assignment of reliability scores applied to solvent accessibility predictions. BMC Struct. Biol.

Reimand, J., Wagih, O. & Bader, G.D. (2013) The mutational landscape of phosphorylation signaling in cancer. Sci. Rep. doi: 10.1038/srep02651.

Thomsen, M.C. & Nielsen, M. (2012) Seq2Logo: a method for construction and visualization of amino acid binding motifs and sequence profiles including sequence weighting, pseudo counts and two-sided representation of amino acid enrichment and depletion. Nucleic Acids Res. 40, 281-7.

Wang, D., et. al. (2017) MusiteDeep: a deep-learning framework for general and kinase-specific phosphorylation site prediction. Bioinformatics 33, 3909–16.

Xue, Y., et. al. (2008) GPS 2.0, a Tool to Predict Kinase-specific Phosphorylation Sites in Hierarchy. Mol. Cell. Proteomics. 7, 1598-1608.

Zhang, H., Lund, O. and Nielsen, M. (2009) The PickPocket method for predicting binding specificities for receptors based on receptor pocket similarities: application to MHC-peptide binding. Bioinformatics 25(10), 1293-9.