Experimental measurement and computational prediction of bacterial Hanks-type Ser/Thr signaling system regulatory targets

Noam Grunfeld[1], Erel Levine[1,2#], Elizabeth Libby[1#*]

[1]Department of Bioengineering, Northeastern University, Boston MA USA

[2]Department of Chemical Engineering, Northeastern University, Boston MA USA

[#]These authors contributed equally to this work

*Corresponding Author: e.libby@northeastern.edu

Keywords: Ser/Thr kinase, Ser/Thr phosphatase, phosphorylation, phosphoproteomics, computational models

Abstract

Bacteria possess diverse classes of signaling systems that they use to sense and respond to their environments and execute properly timed developmental transitions. One widespread and evolutionarily ancient class of signaling systems are the Hanks-type Ser/Thr kinases, also sometimes termed "eukaryotic-like" due to their homology with eukaryotic kinases. In diverse bacterial species, these signaling systems function as critical regulators of general cellular processes such as metabolism, growth and division, developmental transitions such as sporulation, biofilm formation, and virulence, as well antibiotic tolerance. This multifaceted regulation is due to the ability of a single Hanks-type Ser/Thr kinase to post-translationally modify the activity of multiple proteins, resulting in the coordinated regulation of diverse cellular pathways. However, in part due to their deep integration with cellular physiology, to date we have a relatively limited understanding of the timing, regulatory hierarchy, the complete list of targets of a given kinase, as well as the potential regulatory overlap between the often multiple kinases present in a single organism. In this review we discuss experimental methods and curated datasets aimed at elucidating the targets of these signaling pathways, and approaches for using these datasets to develop computational models for quantitative predictions of target motifs. We emphasize novel approaches and opportunities for collecting data suitable for the creation of new predictive computational models applicable to diverse species.

Introduction

Bacteria use signaling systems to sense and respond to their environment. This enables them to survive their often-changing environments, execute properly timed developmental transitions including to virulent states, and survive stress and antibiotic treatment. Among these signaling systems are the Hanks-type Ser/Thr kinases and phosphatases[1], also termed "eukaryotic-like" (or eSTKs/eSTPs) due to their homology to eukaryotic signaling systems. Compared to eukaryotic systems that began to be characterized over 60 years ago, prokaryotic systems were only first identified in the early 1990s[2, 3].These bacterial kinases are likely evolutionarily ancient,

sharing a common ancestor with those found in eukarya and archaea[4, 5]. These signaling systems typically consist of a receptor kinase that phosphorylates targets on Ser or Thr residues and a partner phosphatase that provides reversible regulation through dephosphorylation[6]. Unlike other phosphorylation-based signaling systems such as bacterial two-component systems, in which the kinase generally regulates cellular physiology through a dedicated transcription factor (response regulator)[7], the Hanks-type bacterial Ser/Thr kinases can regulate cellular physiology more broadly through direct phosphorylation of diverse classes of proteins[8]. These target proteins are not limited to transcription factors, and often include other types of proteins such as enzymes in central metabolism, translation factors, enzymatic pathways, and structural components, in addition to cross regulation of other signaling pathways [6, 9]. In contrast to Asp phosphorylation in two-component systems, Ser/Thr phosphorylation is relatively stable, with a typically significantly longer half-life[7]. Like their homologs in eukaryotes, prokaryotic Hanks-type Ser/Thr signaling systems also use a separate phosphatase (sometimes termed eukaryotic-like phosphatases or eSTPs) to provide reversible regulation[8].

Because of their ability to regulate multiple pathways concurrently, in many bacterial species Hanks-type Ser/Thr signaling can be essential and appears to function as a kind of "master regulator" for coordinating cell growth and division, metabolism, development, and stress resistance [8, 10-13]. In several species, including clinically important pathogens, this class of kinases is known to be essential and/or regulate antibiotic resistance, making these pathways an attractive drug target[14, 15].

As the targets of these systems are diverse and demonstrably often critical for cellular physiology, there has been considerable interest in attempting to identify, characterize, and predict the regulatory targets of every known kinase. Experimental methods developed for this aim, while rapidly improving, are often highly labor intensive, especially since the list of targets can be highly growth state specific. It is therefore highly warranted to develop computational models for predicting putative targets and their properties directly from genome sequence data. Critically, such models perform the best when built on large-scale high-quality training data from robust experimental results. In this review, we will discuss the current availability of such

experimental data sets and computational models, and highlight the types of data and models that can have a significant impact on our understanding of these signaling pathways.

In order to train a computational model to predict the targets of a specific kinase, it is necessary to have significant amounts of robust experimental data on its precise phosphorylation sites. However, to date comprehensively identifying diverse phosphosites in bacteria and correctly matching them with the appropriate pathway has been challenging. The optimal scenario would include a robust method to activate the signaling pathway coupled with a reliable readout of target activation in live cells. Such methods, however, are not typically available. As discussed in this review, a multitude of experimental techniques have been used to date in diverse species, including phosphoproteomics, *in vitro* kinase assays, genetics, peptide libraries, and synthetic transcription factors (Table 1). In principle, combining these experimental methods with new computational techniques could enable a deeper understanding of bacterial physiology, including in understudied and non-domesticated species, aid in the development of new antibiotics, as well as develop new regulatory pathways for synthetic biology or industrial applications.

Experimental approaches

Phosphoproteomics

Given the diverse classes of possible regulatory targets of Hanks-type Ser/Thr kinases, whole proteome screening for phosphosites has the potential to identify lists of putative target sites that can then be matched with the appropriate pathway. Furthermore, the relative stability of Ser/Thr phosphorylation (compared to His/Asp) makes them particularly suitable for phosphoproteomics. In this technique, bacterial cultures are lysed, and proteins are digested into peptide fragments (e.g., with trypsin). This is followed by a phosphopeptide enrichment step to increase the relative proportion of phosphorylated peptides. The resulting peptides are then analyzed by mass spectrometry to identify mass shifts consistent with phosphorylation[16]. This method identifies phosphorylated peptides, regardless of mechanism. However, since

Hanks-type Ser/Thr kinases are widespread and often abundant in bacterial genomes and have many putative targets, it is reasonable to assume that a significant (or even predominant) fraction of the phosphosites identified by phosphoproteomics can be attributed to the Hanks-type signaling pathways [11]. Indeed this has been successful in identifying many possible phosphosites and pathways of interest in diverse organisms ranging from model organisms such as *Bacillus subtilis* and *Escherichia coli* to clinically relevant pathogens such as *Mycobacterium tuberculosis*, *Acinetobacter baumannii*, *Clostridium difficile*, *Staphylococcus aureus*, *Streptococcus pyogenes*, *Listeria monocytogenes*, *Bordetella pertussis*, *Streptococcus pneumoniae* among many others, and as reviewed in [17, 18]. Across bacterial species, certain pathways and proteins tend to appear consistently in all data sets, including for example translation factors, enzymes involved in central metabolism and cell wall synthesis, as well as virulence factors.

With good reason, bacterial phosphoproteomics is believed to suffer from poor coverage of the proteome. In many studies, a significant fraction of the proteome is not detected, which is a prerequisite to detecting a phosphopeptide at likely even lower abundance[19]. Therefore, the inability to detect a specific phosphosite may be due to many factors, including issues of protein abundance and stability, lack of proper pathway activation, as well as intrinsic physical and chemical differences between peptides causing them to ionize unequally or degrade. Although the size of the phosphoproteome is not known, close examination of the proteomic data sets suggests that many proteins and their possible corresponding phosphosites are not being identified. For example, it is clear that membrane proteins are currently unrepresented in the data sets, which can be at least partially attributable to technical challenges around mass spectrometry compatible solubilization [20-22]. Recently, advances in phosphoproteomics have strongly increased the sensitivity and depth of these data sets, resulting in large increases in the number of phosphosites identified across many bacterial species. For example, whereas early studies on *B. subtilis* identified ~103 phosphorylation sites on ~78 proteins [23], approximately 10 years later studies in the same bacterium identified ~1085 phosphorylations on ~488 proteins [19], providing much larger data sets. In other organisms, improvements in phosphopeptide enrichment have been shown to increase the number of phosphopeptides identified two to four-

fold for *S. pyogenes* and *L. monocytogenes*, resulting in approximately ~400 phosphorylated proteins per organism [24].

While phosphoproteomics can broadly identify phosphorylation sites, it does not in isolation directly identify the kinase or kinases responsible. To implicate a specific kinase with the phosphorylation of potential phosphosites several studies have used kinase and phosphatase mutants, kinase depletion strains, or specific kinase inhibitors to look for changes in the relative abundance of identified sites using phosphoproteomics. For some recent examples in various organisms see *M. tuberculosis*[25, 26], *B. subtilis* [27], [19] *S. aureus* [28, 29] *L. monocytogenes* [30], *S. pneumoniae*[31], and *E. coli* [32] . While this method does not rule out indirect interactions, it does help narrow the possible targets of interest and specific pathways for further study [32-34].

Phosphoproteomics is the only experimental approach that generates large-scale data sets, which are essential for training modern machine-learning models. Even without attribution to a specific pathway, experimentally confirmed phosphosites can help to pre-train a model or to find effective ways to mathematically represent phosphosites sequences. The improvement in quality and sensitivity of phosphoproteomics techniques is therefore conducive of developing better machine-learning models.

In vitro kinase assays

The gold standard approach to validating a matched kinase-substrate interaction is the *in vitro* kinase assay. In its simplest form, a purified kinase and a substrate are incubated together in the presence of ATP and magnesium to allow phosphotransfer to occur. Often these reactions are directly detected using gamma-^{32}P (or ^{33}P) ATP, phosphoprotein separation using Phos-tag gels, or less commonly, phospho-specific antibodies (α-phos-Thr or α-phos-Ser) or stains. There are some important advantages to *in vitro* kinase assays. Due to the use of purified components, the reactions can be used to determine specific residues that are phosphorylated on both the kinase (autophosphorylation) and on the substrate when combined with downstream mass spectrometry. Since the substrates are purified, this also often results in much higher coverage of the protein by mass spectrometry, aiding identification of phosphosites. This workflow has

been used successfully to identify specific target residues in a large variety of organisms. Some very recent examples include the identification of the phosphorylation sites responsible for the regulation of the protease PrkA by PrkC in *B. subtilis* [35], phosphorylation sites on the peptidoglycan hydrolase CwlA by PrkC in *C. difficile* [36], phosphorylation of GpsB by IreK in *E. faecalis*[37, 38], and the regulation of capsular polysaccharides in *Streptococcus suis* through Stk1 phosphorylation of CcpS[39] and in *S. pneumoniae* through StkP phosphorylation of CcpA[40] . Importantly, this method also allows for the matching of a specific phosphosite on a substrate with the activity of a specific kinase. Although this method is well known to be potentially prone to false positives due to unphysical interaction times or stoichiometries, there are ways to minimize this concern with time dependent concentration titrations, for example as was done systematically for the PhoB/PhoR TCS system [41]. Limiting reaction times has also been used to identify histidine kinase – response regulator specificity in TCS systems[42, 43], a technique that has been successfully used to reveal the specificity of the interaction between the Hanks-type Ser/Thr kinase PrkC and the response regulator WalR [44] in *B. subtilis* . This study demonstrated specificity for WalR by PrkC even among response regulators with highly conserved amino acid sequences around the phosphosite.

Although this method can produce the most precise and detailed results, it is important to note that there are some inherent challenges in attempting high-throughput *in vitro* kinase assays. One of the main challenges is the reliable expression and purification of an active Hanks-type kinase, as expression of these kinases can be highly toxic or difficult to purify in standard expression systems such as *E. coli*. This was encountered in a systematic attempt to purify all known Hanks-type kinases from *M. tuberculosis*[33]. Additionally, *in vitro* assays often use only the catalytic domain of the kinase, discarding its extracellular and transmembrane domains. This can strongly impact kinase activity, as seen for example with the *B. subtilis* kinase Ser/Thr YabT[45 46]. In many cases it is difficult to disentangle these effects, as less active kinases can in principle be less toxic and easier to purify. Finally, *in vitro* assays are time consuming and often require system-specific expertise. Still, to date, this remains the most robust method for pairing the activity of a given kinase with a specific phosphorylated residue on a substrate.

Once sites are identified, they can often be further validated *in vivo* using a combination of point mutants (e.g. in the phosphosite) or kinase and phoshphatase mutants to infer the connection between a given phenotype and a phosphosite. Often these validations are done using a combination of methods – for example, immunoprecipitation of a potential phosphoprotein, followed by phos-tag gel separation and/or blotting with α-phos-Thr or α-phos-Ser antibodies, or using a phospho-specific stain. Some very recent examples of the success of this workflow include the regulation of quiescence and antibiotic tolerance in *S. aureus* associated with EF-G phosphorylation[29], determination of the GpsB phosphosites responsible for cephalosporin resistance in *E. faecalis*[47], and the phosphosites on the transcriptional regulator CodY that regulate anthrax toxin production in *B. anthracis*[48].

Synthetic peptide target libraries for motif prediction

Like their eukaryotic kinase relatives, bacterial Hanks-type kinases can recognize short peptides (~13 amino acids), enabling *in vitro* screening for phosphorylation of libraries of synthesized peptides using a purified kinase (Figure 1(a)). These data can then be used to identify sequence motifs for that specific kinase, or can be used to train a more general computational model, as has been done for related eukaryotic kinases (see for example [49]). This approach has been used for nine kinases of this class found in *M. tuberculosis* to reveal kinase specific phosphopeptide motifs [33] in a combined synthetic library *in vitro* kinase assay approach. In this work, a small library (~336) of biotinylated peptides based on sites identified by phosphoproteomics was created. Each peptide in the library was incubated in the presence of radiolabeled ATP with a panel of nine purified kinases. The peptides were then bound to streptavidin coated plates, washed, and assayed for ^{33}P incorporation. This highly sensitive method found that roughly half the peptides could be phosphorylated to some degree by at least one of the nine kinases, and many peptides could be phosphorylated by most or all of them. A much smaller fraction of the library (~48 substrates) were phosphorylated by only one kinase in this assay, suggesting the identification of a kinase-substrate pair. This dataset was used to computationally predict the preferred substrate motif for the six kinases that were the most active *in vitro*. Interestingly, this strategically designed small library revealed the importance of specific residues on the

target phosphopeptides (e.g., large hydrophobic residues at the +3 and +5 positions relative to the phosphosite), demonstrating how strategically designed peptide libraries have the potential to reveal detailed information for bacterial kinase specificity. This approach requires addressing several experimental challenges, including purification of active kinases, optimization of *in vitro* assays, as well as quantitative precision of the readout.

Modular synthetic transcription factors and sensors

Many of the challenges in the *in vitro* approaches discussed above can be circumvented by *in vivo* assays. Extensive interest in measuring kinase activity for related eukaryotic kinases *in vivo* lead to the development of genetically encoded FRET-based biosensors for kinase activity that have single cell resolution [50]. These sensors have a modular design, consisting of a FRET pair of fluorophores, a short phosphorylatable substrate sequence, and a forkhead-associated domain that specifically binds phosphopeptides (Figure 1(b)). Upon phosphorylation of the substrate sequence, a conformational change occurs, resulting in a change in FRET signal. These sensors were successfully used for eukaryotic Ser/Thr kinases such as PKC[51] and Aurora B [52] among more than 20 others[50], and their modular nature proved adaptable to the bacterial Hanks-type Ser/Thr kinase PrkC from *B. subtilis* [53]. This modular design was used to swap the substrate peptide among four variants and observe sequence-specific changes in phosphorylation activity.

Prototypical two-component systems have a dedicated response regulator transcription factor. A straightforward way to assay their activity *in vivo* is to express a reporter protein from a promoter that is directly regulated by that transcription factor[54]. In contrast, Hanks-type Ser/Thr kinases are not typically the only regulators of a transcription factor[55]. Therefore, creating a transcriptional reporter for this family of kinases required the design of a synthetic transcription factor. Using the design principles of the bacterial FRET sensor and protein engineering, a modular synthetic transcription factor that specifically responds to PrkC activity in *B. subtilis* was created [53]. The design of this transcription factor relies on the ability of Hanks-type kinases to phosphorylate short substrate peptides. In this case, the substrate peptides are embedded

within LacI, the inhibitor of the *lac* operon (Figure 1(c)). When phosphorylated, these substrates can bind to a phospho-binding domain (FHA2 originally from Rad53[51]) and decrease the activity of the engineered LacI, resulting in downstream gene expression. These modular sensors have been used to demonstrate pathway activation by providing a direct and dynamic *in vivo* readout of kinase activity that can be measured in colonies on petri dishes, in bulk liquid cultures, or by microscopy in single cells.

As related sensors have been successfully used in many similar eukaryotic systems, it is likely these sensors can be further extended to bacterial systems beyond *B. subtilis* with some optimization. Since the sensitivity of the synthetic transcriptional regulator and the FRET sensor both rely on conformational changes induced by phosphorylation and subsequent binding to a phosphopeptide binding domain, extending the use of these systems to different bacterial species should be initially optimized in the context of controls. This is to minimize off target effects and sensitivity of the sensor to phosphorylation, for example by testing a specific phosphosubstrate choice using kinase and phosphatase mutant genetic backgrounds, or performing *in vitro* or *in vivo* kinase assays. As an additional consideration, the modular phosphopeptide binding domain (FHA2) used in the *B. subtilis* study has been characterized to be partially sensitive to the choice of amino acid in the +3 position relative to the phosphosite[56]. For example, better sensitivity was achieved using an I in the +3 position as a biosensor in both the PrkC study in *B. subtilis*[53] and was used for a FRET biosensor for Aurora B activity in eukaryotic cells[52]. After optimization, the modular nature of this sensor and its single-cell sensitivity could allow quantitative measurements of the specificity of a large substrate library, with the high throughput and accuracy required for training machine learning models.

Computational approaches

The availability of large data sets of experimentally verified phosphosites raises the possibility that machine learning approaches could be used to improve the curation and characterization of the phosphoproteome. The questions that can potentially be addressed by these approaches include the prediction that a specific site on a given protein can be substrate for

phosphorylation (a phosphosite); the prediction that a site is phosphorylated by a given kinase or kinases; and the quantitative prediction of the likelihood of such events, especially in quantitative comparison with other potential substrates of the same kinase. While several attempts have been made to develop such models, the success of available models is limited.

Available datasets

UniProt, the comprehensive resource for protein sequence and data [57], aims to include all known post-translational modifications for each protein in the database, including those from bacteria. For each protein in the database, UniProt identifies all known post-translationally modified (PTM) sites as well as the kinases that catalyze their modification, when these are known. For bacterial proteins, however, this information is often partial or outdated.

The development of computational approaches to the study of the posphoproteome benefits from dedicated databases. A plethora of such databases are available for eukaryotic species, organized by species, by kinase families, by experimental method, and more (for a detailed list see [58]). Broad databases used recently for training large-scale machine-learning models include dbPTM [59], PhosphoSitePlus [60], and EPSD [61]. These databases provide a comprehensive view of PTM sites by integrating data from multiple other databases. dbPTM includes PTM sites in bacterial proteins, but like UniProt discussed above, these data are often spotty and outdated.

To our knowledge, only one database that is focused on prokaryotic phosphorylation sites is actively maintained [62]. This database, dbPSP, contains almost 20,000 experimentally validated phosphosites from more than 2000 bacterial species. While the site provides reference for the source of information for every identified site, it does not explicitly identify upstream kinases or phosphatases, even when such information is available. This hinders the use of these data for development of models that link substrates with their associated regulators.

Since bacterial phosphosites exhibit a high degree of conservation [63] these databases can provide a useful starting point for proteins that have not been experimentally tested if information is available for their homologs in related species. This observation could in principle be used as a

prior for computational models, increasing the confidence that a conserved site acts as a phosphosite. However, making the quantitative connection between the degree of conservation and the level of confidence would require detailed experimental data across species for kinase families that is not currently available.

Prediction of phosphorylation targets

The tasks of identifying phosphosites in a given protein or identifying potential phosphorylation targets of a specific kinase have attracted machine learning approaches for more than two decades. Given the availability and accessibility of large data sets for eukaryotic kinase targets, most of the modeling effort has been focused on eukaryotic kinases (mostly those in mammals and yeast) [58]. Still, some efforts have been made to develop computational tools for predicting phosphorylation targets in bacteria in general [64-67] and for the *B. subtilis* Ser/Thr kinase PrkC in particular [68].

Most computational approaches use the sequence around a potential phosphosite to determine the likelihood that it is actively phosphorylated. The hypothesis behind these approaches is that a local signal near the phosphosite is necessary for recognition by the relevant kinase. To predict new phosphosites, the substrate sequences of known phosphosites are used to learn common sequence features that could be responsible for molecular recognition. Next, the sequences of candidate proteins are scanned for sites that distinctively exhibit these features. As described below, models that take this approach differ in the length of the substrate sequence they use, as well as in the use of additional information (such as structural or biochemical information).

Other approaches focus on other types of information instead or in addition to the substrate sequence, including evolutionary conservation or patterns of phosphorylation events across tissues and experimental conditions. Examples of such approaches applied to eukaryotic kinases are given below. These approaches, however, require large data sets that are only starting to become available for bacteria.

Machine learning and bacterial phosphorylation

Sequence-based approaches are typically formulated as classification problems: given a short sequence, the task is to determine whether it represents a phosphosite or not, or alternatively whether it is a substrate for a specific kinase or not (Figure 2). The success of such models can be unequivocally evaluated by measuring their ability to correctly predict phosphosites that were not part of their training data. Different implementations of this concept are distinct in two important ways: the representation of the input sequence, and the specific model used for classification. Beyond the obvious need to decide on the length of the sequence used by the model, models can be presented with the amino-acid sequence alone, or with additional information such as chemical properties of each amino acid, structural features, and more. Among the many models available for classification tasks, two approaches – Support Vector Machines (SVMs) and Random Forests – are particularly popular in the computational biology space, because they both work well with data sets that are not very large (10s or 100s of samples). In addition, the structure of these models sometimes allows identifying what sequence features were recognized by the model as the most informative for classifying them as phosphosites.

NetPhosBac, one of the earlier attempts [69], used a very small set of 140 MS-verified phosphorylation sites in *E. coli* or *B. subtilis* to train a small neural network, which only used a 13 amino-acid substrate (5 amino acids on each side of the phosphosite) as input. This model achieved a very limited success. The same data set was used, a few years later, to develop another machine learning predictor, cPhosBac [67]. The design of this model around a Support Vector Machine (SVM) was more appropriate for such a small data set and showed a mildly improved performance. A similar approach was taken in an attempt to identify targets of a single kinase, PrkC [68]. This study used as few as 36 experimentally verified phosphorylation sites as a training set. While cross-validation suggested high performance, it would be reasonable to doubt the generalizable predictive power of this model.

Finally, a recent model nicknamed MPSite[64] used a previous version of the dbPSP database mentioned above to establish a training set of more than 1700 phosphorylation substrates, an order of magnitude more than the data used in previous models. The new idea behind this model, which was built on a Random Forest classifier, was to combine multiple encodings of the 21-amino acid substrate sequences. In addition to the primary sequence, these encodings represent chemical and structural properties. The authors of MPSite showed that the combination of multiple representations significantly improve the performance of the model. This represents the current state of the art, with 81% specificity (the true-negative rate), at 41% and 62% accuracy (the fraction of correct predictions) for Phospho-serine and Phospho-threonine sites, respectively.

Lessons from eukaryotic models

As mentioned above, considerably more data are available for eukaryotic phosphorylation sites, likely at least partially due to the fact eukaryotic systems were discovered much earlier[2, 8]. These data were used to develop multiple machine learning models. Whether these models were trained on a specific kinase, specific organism, or more comprehensive eukaryotic data, they cannot be used to predict bacterial sites [69].

Many research groups developed computational tools for predicting general or kinase-specific phosphosites [67, 70-74]. Two tools that underwent multiple rounds of revisions and updates represent the current state of the art: KinasePhos[75] is built around a support vector machine (SVM) trained on 41,421 experimentally verified kinase-specific phosphorylation sites from several animals, two species of yeast, and one plant, while Group-based Prediction System [76] (GPS) integrates a logistic regression and a deep neural network trained on 490,762 sites. Both works take the approach of training a general model for predicting phosphosites, and then retraining specific models for individual kinases. KinasePhos 3.0 and GPS 6.0 include respectively 771 and 44,046 models for different kinases, kinase families, and family groups. On average, the accuracy of these models exceeds 87%, with specific models of better-studied kinases reaching up to 98%. While the updated design of these models include some modern

algorithms, what makes them truly powerful is the use very large data sets that allow optimization of feature representation, investigation of the power of different features to inform the model, and development of highly specified models [75].

Recent years saw an enormous advance in the application of deep neural network across biology, including microscopy image processing, protein folding, drug design, and more[77-79]. These models are data-hungry and work well only when provided with large sets of labeled data. On the other hand, they are insensitive to noise and can handle experimental inaccuracies relatively well. With the large expansion of available data for eukaryotic kinases, several attempts have been made to develop neural network models for the phosphosite identification [70, 80], including the recent incorporation of a deep neural network into the veteran GPS model[76] [81]. These studies report improvement in accuracy in models that are not kinase specific and are therefore built on large data sets. In addition, it has been suggested that the vast amount of data available for well-studied kinases could also be used to identify potential targets of unknown kinases, using an approach known as zero-shot learning [82].

As mentioned above, other approaches for discovery of PTM interactions and phosphosites do not rely on sequence features. For example, a recent study [83] combined data that associates kinases with conserved protein domains with protein co-expression data to express the probability that a given protein is regulated by a kinase as a function of the number of its domains known to interact with the kinase and their level of co-expression. Based on the rationale that PTM sites tend to show higher conservation than the sequence of the protein in which they reside [63], DAPPLE [80, 84] predicts the probability that a query site is phosphorylated by sequence comparison with homolog proteins.

Using a different type of conservation, a recent study [85] used phospho-proteomics data that was collected in different tissues or under different conditions to identify proteins that are co-phosphorylated, namely phosphorylated under the same set of conditions. This model then predicts that all proteins co-phosphorylated with a known target of a kinase are also modified by the same kinase. These predictions can be enhanced using knowledge about functional

interactions [86, 87]. Notably, these approaches predict that a protein may be modified by a certain kinase, but do not necessarily identify the relevant phosphosite.

The success of these models rely on the breadth of data available in the eukaryotic field. As more data emerges from bacterial system, similar techniques may be applicable to bacterial systems. Moreover, advances in transfer learning may allow to pre-train models using data from eukaryotic systems, and adapting the models to bacterial systems by fine-tuning them with smaller sets of experimental data from bacteria.

Outlook

Ever since the first bacterial Hanks-type Ser/Thr signaling pathway was identified in *M. Xanthus* [3], whole genome sequencing has demonstrated that these evolutionarily ancient pathways are widespread in bacteria. Subsequently, it was discovered that these signaling systems can perform regulation on diverse cellular processes by directly phosphorylation of proteins. Experimentally, this was largely accomplished using a combination of phosphoproteomics and targeted *in vitro* validation of individual phosphosites. Over roughly the last decade, the emergence of several additional experimental advances are poised to enable rapid progress in phosphosite identification. These include technical improvements in bacterial phosphoproteomics, leading to more comprehensive identification of phosphopeptides, and new synthetic approaches using libraries of short peptides *in vitro* enabling precise sequence specific testing of kinase-substrate interactions, and modular transcription factors *in vivo* as a method to demonstrate the effect of substrate sequence on the timing and abundance of phosphorylation. Together these advances have the potential to create much larger and higher quality data sets than were previously available and provide the tools for detailed testing of computational model predictions. However, a lack of uniformity and the limited scope of the data that associates prokaryotic kinases with their respective phosphosite targets hinders the ability to develop robust predictive models.

The use of machine learning models for prediction and discovery of kinase phosphosites has increased with the development of high-throughput experimental methods. The increasing success of models focused on eukaryotic kinases suggests that the expansion in data size and diversity will ultimately allow the development of robust models for bacterial kinases. This would require a community effort to create a well-curated, well-labeled, freely accessible database. Such effort should include standardizing data reporting for the field asking for example that experimental results include species, sites, responsible kinase, experimental method, growth condition, etc. In addition, it would be useful to associate each phosphosite with a well-defined quantitative score that indicates the strength of the observed phosphitylation activity at that site. These data could then be used to train models to identify features that distinguish high-occupancy from low-occupancy sites, as well as to distinguish between true low-occupancy sites and experimental noise. Such standardized well-curated databases would be instrumental in enabling bacterial-specific predictive models for kinase-substrate predictions.

Despite the incompatibility of models designed for eukaryotic kinases in predicting targets of bacterial Ser/Thr kinases, recent progress in transfer learning opens up the possibility of utilizing these models as a foundation for constructing dedicated models for bacterial kinases. This is particularly attractive since the current data sets from related eukaryotic systems are much larger than the bacterial ones and could therefore facilitate productive use of the much smaller bacterial data sets. To our knowledge, this has not yet been attempted, in part because of the lack of easily accessible well-curated and labeled bacterial data. In addition, lessons learned from multi-dimensional representation of query sequences, which includes structural and biochemical properties in addition to primary sequence, could be incorporated into bacteria-focused models with minimal modification.

Many factors influence whether a specific kinase phosphorylates a potential phosphosite. These include externals factors, such as environmental signals and growth conditions, and internal ones, such as the abundance of co-factors and competing targets. The computational approaches discussed here aim to identify all possible phosphosites, realizing that some of them may not be

phosphorylated under certain conditions. Moreover, it is possible that some families of targets that are not phosphorylated in the conditions used to train the computational models and will be absent from its predictions. A future challenge is to develop a computational model tasked with predicting the probability that a phosphosite is phosphorylated under given conditions. Developing such models would require detailed characterization of phosphorylation abundance across multiple conditions of different types.

Being able to identify and define the regulatory targets of a Hanks-type kinase is an important first step towards several important goals. First, these signaling pathways are involved in regulating cellular growth and survival, and the ability to follow the regulatory dynamics of these targets can expose novel physiological mechanisms. Second, since these pathways have been implicated in antibiotic resistance, knowing the targets involved can help in devising novel strategies to robustly interfere with the emergence of resistance and guide development of synergistic therapies. Finally, kinases have been empirically discovered as potentially efficient drug targets (e.g., *M. tuberculosis* PknB is an essential protein[88-90]), and the knowledge of their affected targets can reveal mechanisms of action of such drugs. In all these applications, the availability of real-time *in vivo* reporters can be instrumental in uncovering causal interactions and pathway dynamics.

Strategic use of new technical advances in experimental techniques, such as improved phosphoproteomics, new *in vivo* techniques using synthetic biology, and cheap library generation and sequencing can therefore enable large strides in our ability to generate predictive machine-learning based models for the targets of bacterial Hanks-type Ser/Thr signaling systems. This will take on particular importance as understanding bacterial physiology becomes increasingly important in industrial and medical applications on undomesticated or poorly genetically tractable strains.

Acknowledgments

We thank Jonathan Dworkin for helpful comments. This work was supported by NIH grant R35GM147429 to EAL and by NSF grant MCB1946944 to EL.

The authors declare no conflict of interest. Figure 1 was partially created with the aid of Biorender.

Figure and table legends

Table 1. Summary of current experimental techniques and their respective advantages/disadvantages for building computational models

Figure Legends

Figure 1. High-throughput and synthetic reporter of kinase activity. A) *in vitro* kinase assays typically involve purified kinases and a purified substrate – either full length protein targets (top), or short synthetic peptides (bottom). **B)** *in vivo* FRET sensors have been adapted from homologous eukaryotic systems and shown to function in bacteria. They rely on the conformational change of the protein sensor induced by a phosphorylated substrate binding to a forkhead-associated domain (FHA2), resulting in loss of efficient fluorescence energy transfer between the two fluorophores (decrease in FRET). **C)** Synthetic transcription factors have been engineered to specifically respond to Hanks-type Ser/Thr kinase activity *in vivo*. The lac repressor (LacI) was translationally fused to a forkhead-associated domain (FHA2) and a phosphorylatable substrate, creating a synthetic transcription factor. Upon phosphorylation of the substrate by a specific kinase, repression of the promoter is reduced, resulting in reporter gene expression.

Figure 2. Computational workflow from input data to testable predictions. Various types of input data from sequencing, experimental determined phosphosites, structural properties, evolutionary conservation, and co-phosphorylation patterns can be used as inputs in a computational model. The model outputs can include (but is not limited to) testable predictions about the presence of a phosphosite, associated kinase, and similarity to related systems.

References

1. Hanks SK, Quinn AM, Hunter T. The protein kinase family: conserved features and deduced phylogeny of the catalytic domains. Science. 1988;241(4861):42-52. doi: 10.1126/science.3291115. PubMed PMID: 3291115.

2. Bakal CJ, Davies JE. No longer an exclusive club: eukaryotic signalling domains in bacteria. Trends Cell Biol. 2000;10(1):32-8. doi: 10.1016/s0962-8924(99)01681-5. PubMed PMID: 10603474.

3. Munoz-Dorado J, Inouye S, Inouye M. A gene encoding a protein serine/threonine kinase is required for normal development of M. xanthus, a gram-negative bacterium. Cell. 1991;67(5):995-1006. doi: 10.1016/0092-8674(91)90372-6. PubMed PMID: 1835671.

4. Leonard CJ, Aravind L, Koonin EV. Novel families of putative protein kinases in bacteria and archaea: evolution of the "eukaryotic" protein kinase superfamily. Genome Res. 1998;8(10):1038-47. doi: 10.1101/gr.8.10.1038. PubMed PMID: 9799791.

5. Stancik IA, Sestak MS, Ji B, Axelson-Fisk M, Franjevic D, Jers C, Domazet-Loso T, Mijakovic I. Serine/Threonine Protein Kinases from Bacteria, Archaea and Eukarya Share a Common Evolutionary Origin Deeply Rooted in the Tree of Life. J Mol Biol. 2018;430(1):27-32. Epub 20171111. doi: 10.1016/j.jmb.2017.11.004. PubMed PMID: 29138003.

6. Dworkin J. Ser/Thr phosphorylation as a regulatory mechanism in bacteria. Curr Opin Microbiol. 2015;24:47-52. Epub 20150124. doi: 10.1016/j.mib.2015.01.005. PubMed PMID: 25625314; PMCID: PMC4380854.

7. Stock AM, Robinson VL, Goudreau PN. Two-component signal transduction. Annu Rev Biochem. 2000;69:183-215. doi: 10.1146/annurev.biochem.69.1.183. PubMed PMID: 10966457.

8. Pereira SF, Goss L, Dworkin J. Eukaryote-like serine/threonine kinases and phosphatases in bacteria. Microbiol Mol Biol Rev. 2011;75(1):192-212. doi: 10.1128/MMBR.00042-10. PubMed PMID: 21372323; PMCID: PMC3063355.

9. Nagarajan SN, Lenoir C, Grangeasse C. Recent advances in bacterial signaling by serine/threonine protein kinases. Trends Microbiol. 2022;30(6):553-66. Epub 20211124. doi: 10.1016/j.tim.2021.11.005. PubMed PMID: 34836791.

10. Manuse S, Fleurie A, Zucchini L, Lesterlin C, Grangeasse C. Role of eukaryotic-like serine/threonine kinases in bacterial cell division and morphogenesis. FEMS Microbiol Rev. 2016;40(1):41-56. Epub 20150930. doi: 10.1093/femsre/fuv041. PubMed PMID: 26429880.

11. Mijakovic I, Macek B. Impact of phosphoproteomics on studies of bacterial physiology. FEMS Microbiol Rev. 2012;36(4):877-92. Epub 20111128. doi: 10.1111/j.1574-6976.2011.00314.x. PubMed PMID: 22091997.

12. Macek B, Forchhammer K, Hardouin J, Weber-Ban E, Grangeasse C, Mijakovic I. Protein post-translational modifications in bacteria. Nat Rev Microbiol. 2019;17(11):651-64. Epub 20190904. doi: 10.1038/s41579-019-0243-0. PubMed PMID: 31485032.

13. Janczarek M, Vinardell JM, Lipa P, Karas M. Hanks-Type Serine/Threonine Protein Kinases and Phosphatases in Bacteria: Roles in Signaling and Adaptation to Various Environments. Int J Mol Sci. 2018;19(10). Epub 20180921. doi: 10.3390/ijms19102872. PubMed PMID: 30248937; PMCID: PMC6213207.

14. Pensinger DA, Schaenzer AJ, Sauer JD. Do Shoot the Messenger: PASTA Kinases as Virulence Determinants and Antibiotic Targets. Trends Microbiol. 2018;26(1):56-69. Epub 20170719. doi: 10.1016/j.tim.2017.06.010. PubMed PMID: 28734616; PMCID: PMC5741517.

15. Bonne Kohler J, Jers C, Senissar M, Shi L, Derouiche A, Mijakovic I. Importance of protein Ser/Thr/Tyr phosphorylation for bacterial pathogenesis. FEBS Lett. 2020;594(15):2339-69. Epub 20200617. doi: 10.1002/1873-3468.13797. PubMed PMID: 32337704.

16. Macek B, Mijakovic I. Site-specific analysis of bacterial phosphoproteomes. Proteomics. 2011;11(15):3002-11. Epub 20110704. doi: 10.1002/pmic.201100012. PubMed PMID: 21726046.

17. Richter E, Mostertz J, Hochgrafe F. Proteomic discovery of host kinase signaling in bacterial infections. Proteomics Clin Appl. 2016;10(9-10):994-1010. Epub 20160909. doi: 10.1002/prca.201600035. PubMed PMID: 27440122; PMCID: PMC5096009.

18. Lim S. A Review of the Bacterial Phosphoproteomes of Beneficial Microbes. Microorganisms. 2023;11(4). Epub 20230403. doi: 10.3390/microorganisms11040931. PubMed PMID: 37110354; PMCID: PMC10145908.

19. Ravikumar V, Nalpas NC, Anselm V, Krug K, Lenuzzi M, Sestak MS, Domazet-Loso T, Mijakovic I, Macek B. In-depth analysis of Bacillus subtilis proteome identifies new ORFs and traces the evolutionary history of modified proteins. Sci Rep. 2018;8(1):17246. Epub 20181122. doi: 10.1038/s41598-018-35589-9. PubMed PMID: 30467398; PMCID: PMC6250715.

20. Helbig AO, Heck AJ, Slijper M. Exploring the membrane proteome--challenges and analytical strategies. J Proteomics. 2010;73(5):868-78. Epub 20100121. doi: 10.1016/j.jprot.2010.01.005. PubMed PMID: 20096812.

21. Gilmore JM, Washburn MP. Advances in shotgun proteomics and the analysis of membrane proteomes. J Proteomics. 2010;73(11):2078-91. Epub 20100823. doi: 10.1016/j.jprot.2010.08.005. PubMed PMID: 20797458.

22. Alfonso-Garrido J, Garcia-Calvo E, Luque-Garcia JL. Sample preparation strategies for improving the identification of membrane proteins by mass spectrometry. Anal Bioanal Chem. 2015;407(17):4893-905. Epub 20150513. doi: 10.1007/s00216-015-8732-0. PubMed PMID: 25967148.

23. Macek B, Mijakovic I, Olsen JV, Gnad F, Kumar C, Jensen PR, Mann M. The serine/threonine/tyrosine phosphoproteome of the model bacterium Bacillus subtilis. Mol Cell Proteomics. 2007;6(4):697-707. Epub 20070110. doi: 10.1074/mcp.M600464-MCP200. PubMed PMID: 17218307.

24. Birk MS, Charpentier E, Frese CK. Automated Phosphopeptide Enrichment for Gram-Positive Bacteria. J Proteome Res. 2021;20(10):4886-92. Epub 20210902. doi: 10.1021/acs.jproteome.1c00364. PubMed PMID: 34473931; PMCID: PMC8491273.

25. Carette X, Platig J, Young DC, Helmel M, Young AT, Wang Z, Potluri LP, Moody CS, Zeng J, Prisic S, Paulson JN, Muntel J, Madduri AVR, Velarde J, Mayfield JA, Locher C, Wang T, Quackenbush J, Rhee KY, Moody DB, Steen H, Husson RN. Multisystem Analysis of Mycobacterium tuberculosis Reveals Kinase-Dependent Remodeling of the Pathogen-Environment Interface. mBio. 2018;9(2). Epub 20180306. doi: 10.1128/mBio.02333-17. PubMed PMID: 29511081; PMCID: PMC5845002.

26. Zeng J, Platig J, Cheng TY, Ahmed S, Skaf Y, Potluri LP, Schwartz D, Steen H, Moody DB, Husson RN. Protein kinases PknA and PknB independently and coordinately regulate essential Mycobacterium tuberculosis physiologies and antimicrobial susceptibility. PLoS Pathog. 2020;16(4):e1008452. Epub 20200407. doi: 10.1371/journal.ppat.1008452. PubMed PMID: 32255801; PMCID: PMC7164672.

27. Ravikumar V, Shi L, Krug K, Derouiche A, Jers C, Cousin C, Kobir A, Mijakovic I, Macek B. Quantitative phosphoproteome analysis of Bacillus subtilis reveals novel substrates of the kinase PrkC and phosphatase PrpC. Mol Cell Proteomics. 2014;13(8):1965-78. Epub 20140105. doi: 10.1074/mcp.M113.035949. PubMed PMID: 24390483; PMCID: PMC4125730.

28. Prust N, van der Laarse S, van den Toorn HWP, van Sorge NM, Lemeer S. In-Depth Characterization of the Staphylococcus aureus Phosphoproteome Reveals New Targets of Stk1. Mol Cell Proteomics. 2021;20:100034. Epub 20210111. doi: 10.1074/mcp.RA120.002232. PubMed PMID: 33444734; PMCID: PMC7950182.

29. Huemer M, Mairpady Shambat S, Hertegonne S, Bergada-Pijuan J, Chang CC, Pereira S, Gomez-Mejia A, Van Gestel L, Bar J, Vulin C, Pfammatter S, Stinear TP, Monk IR, Dworkin J, Zinkernagel AS. Serine-threonine phosphoregulation by PknB and Stp contributes to quiescence and antibiotic tolerance in Staphylococcus aureus. Sci Signal. 2023;16(766):eabj8194. Epub 20230103. doi: 10.1126/scisignal.abj8194. PubMed PMID: 36595572.

30. Kelliher JL, Grunenwald CM, Abrahams RR, Daanen ME, Lew CI, Rose WE, Sauer JD. PASTA kinase-dependent control of peptidoglycan synthesis via ReoM is required for cell wall stress responses, cytosolic survival, and virulence in Listeria monocytogenes. PLoS Pathog.

2021;17(10):e1009881. Epub 20211008. doi: 10.1371/journal.ppat.1009881. PubMed PMID: 34624065; PMCID: PMC8528326.

31. Ulrych A, Fabrik I, Kupcik R, Vajrychova M, Doubravova L, Branny P. Cell Wall Stress Stimulates the Activity of the Protein Kinase StkP of Streptococcus pneumoniae, Leading to Multiple Phosphorylation. J Mol Biol. 2021;433(24):167319. Epub 20211021. doi: 10.1016/j.jmb.2021.167319. PubMed PMID: 34688688.

32. Sultan A, Jers C, Ganief TA, Shi L, Senissar M, Kohler JB, Macek B, Mijakovic I. Phosphoproteome Study of Escherichia coli Devoid of Ser/Thr Kinase YeaG During the Metabolic Shift From Glucose to Malate. Front Microbiol. 2021;12:657562. Epub 20210406. doi: 10.3389/fmicb.2021.657562. PubMed PMID: 33889145; PMCID: PMC8055822.

33. Prisic S, Dankwa S, Schwartz D, Chou MF, Locasale JW, Kang CM, Bemis G, Church GM, Steen H, Husson RN. Extensive phosphorylation with overlapping specificity by Mycobacterium tuberculosis serine/threonine protein kinases. Proc Natl Acad Sci U S A. 2010;107(16):7521-6. Epub 20100405. doi: 10.1073/pnas.0913482107. PubMed PMID: 20368441; PMCID: PMC2867705.

34. Kobir A, Poncet S, Bidnenko V, Delumeau O, Jers C, Zouhir S, Grenha R, Nessler S, Noirot P, Mijakovic I. Phosphorylation of Bacillus subtilis gene regulator AbrB modulates its DNA-binding properties. Mol Microbiol. 2014;92(5):1129-41. Epub 20140429. doi: 10.1111/mmi.12617. PubMed PMID: 24731262.

35. Zhang A, Lebrun R, Espinosa L, Galinier A, Pompeo F. PrkA is an ATP-dependent protease that regulates sporulation in Bacillus subtilis. J Biol Chem. 2022;298(10):102436. Epub 20220828. doi: 10.1016/j.jbc.2022.102436. PubMed PMID: 36041628; PMCID: PMC9512850.

36. Garcia-Garcia T, Poncet S, Cuenot E, Douche T, Giai Gianetto Q, Peltier J, Courtin P, Chapot-Chartier MP, Matondo M, Dupuy B, Candela T, Martin-Verstraete I. Ser/Thr Kinase-Dependent Phosphorylation of the Peptidoglycan Hydrolase CwlA Controls Its Export and Modulates Cell Division in Clostridioides difficile. mBio. 2021;12(3). Epub 20210518. doi: 10.1128/mBio.00519-21. PubMed PMID: 34006648; PMCID: PMC8262956.

37. Iannetta AA, Minton NE, Uitenbroek AA, Little JL, Stanton CR, Kristich CJ, Hicks LM. IreK-Mediated, Cell Wall-Protective Phosphorylation in Enterococcus faecalis. J Proteome Res. 2021;20(11):5131-44. Epub 20211021. doi: 10.1021/acs.jproteome.1c00635. PubMed PMID: 34672600; PMCID: PMC10037947.

38. Minton NE, Djoric D, Little J, Kristich CJ. GpsB Promotes PASTA Kinase Signaling and Cephalosporin Resistance in Enterococcus faecalis. J Bacteriol. 2022;204(10):e0030422. Epub 20220912. doi: 10.1128/jb.00304-22. PubMed PMID: 36094306; PMCID: PMC9578390.

39. Tang J, Guo M, Chen M, Xu B, Ran T, Wang W, Ma Z, Lin H, Fan H. A link between STK signalling and capsular polysaccharide synthesis in Streptococcus suis. Nat Commun.

638 2023;14(1):2480. Epub 20230429. doi: 10.1038/s41467-023-38210-4. PubMed PMID: 37120581;
639 PMCID: PMC10148854.

640 40. Kant S, Sun Y, Pancholi V. StkP- and PhpP-Mediated Posttranslational Modifications
641 Modulate the S. pneumoniae Metabolism, Polysaccharide Capsule, and Virulence. Infect
642 Immun. 2023;91(4):e0029622. Epub 20230306. doi: 10.1128/iai.00296-22. PubMed PMID:
643 36877045; PMCID: PMC10112228.

644 41. Gao R, Stock AM. Probing kinase and phosphatase activities of two-component systems
645 in vivo with concentration-dependent phosphorylation profiling. Proc Natl Acad Sci U S A.
646 2013;110(2):672-7. Epub 20121224. doi: 10.1073/pnas.1214587110. PubMed PMID: 23267085;
647 PMCID: PMC3545780.

648 42. Skerker JM, Prasol MS, Perchuk BS, Biondi EG, Laub MT. Two-component signal
649 transduction pathways regulating growth and cell cycle progression in a bacterium: a system-
650 level analysis. PLoS Biol. 2005;3(10):e334. Epub 20050927. doi: 10.1371/journal.pbio.0030334.
651 PubMed PMID: 16176121; PMCID: PMC1233412.

652 43. Fisher SL, Kim SK, Wanner BL, Walsh CT. Kinetic comparison of the specificity of the
653 vancomycin resistance VanSfor two response regulators, VanR and PhoB. Biochemistry.
654 1996;35(15):4732-40. doi: 10.1021/bi9525435. PubMed PMID: 8664263.

655 44. Libby EA, Goss LA, Dworkin J. The Eukaryotic-Like Ser/Thr Kinase PrkC Regulates the
656 Essential WalRK Two-Component System in Bacillus subtilis. PLoS Genet. 2015;11(6):e1005275.
657 Epub 20150623. doi: 10.1371/journal.pgen.1005275. PubMed PMID: 26102633; PMCID:
658 PMC4478028.

659 45. Bidnenko V, Shi L, Kobir A, Ventroux M, Pigeonneau N, Henry C, Trubuil A, Noirot-
660 Gros MF, Mijakovic I. Bacillus subtilis serine/threonine protein kinase YabT is involved in spore
661 development via phosphorylation of a bacterial recombinase. Mol Microbiol. 2013;88(5):921-35.
662 Epub 20130502. doi: 10.1111/mmi.12233. PubMed PMID: 23634894; PMCID: PMC3708118.

663 46. Shi L, Cavagnino A, Rabefiraisana JL, Lazar N, Li de la Sierra-Gallay I, Ochsenbein F,
664 Valerio-Lepiniec M, Urvoas A, Minard P, Mijakovic I, Nessler S. Structural Analysis of the
665 Hanks-Type Protein Kinase YabT From Bacillus subtilis Provides New Insights in its DNA-
666 Dependent Activation. Front Microbiol. 2018;9:3014. Epub 20190108. doi:
667 10.3389/fmicb.2018.03014. PubMed PMID: 30671027; PMCID: PMC6333020.

668 47. VanZeeland NE, Schultz KM, Klug CS, Kristich CJ. Multisite Phosphorylation Regulates
669 GpsB Function in Cephalosporin Resistance of Enterococcus faecalis. J Mol Biol.
670 2023;435(18):168216. Epub 20230728. doi: 10.1016/j.jmb.2023.168216. PubMed PMID: 37517789;
671 PMCID: PMC10528945.

672 48. Gangwal A, Sangwan N, Dhasmana N, Kumar N, Keshavam CC, Singh LK, Bothra A,
673 Goel AK, Pomerantsev AP, Leppla SH, Singh Y. Role of serine/threonine protein phosphatase

PrpN in the life cycle of Bacillus anthracis. PLoS Pathog. 2022;18(8):e1010729. Epub 20220801. doi: 10.1371/journal.ppat.1010729. PubMed PMID: 35913993; PMCID: PMC9371265.

49. Hutti JE, Jarrell ET, Chang JD, Abbott DW, Storz P, Toker A, Cantley LC, Turk BE. A rapid method for determining protein kinase phosphorylation specificity. Nat Methods. 2004;1(1):27-9. doi: 10.1038/nmeth708. PubMed PMID: 15782149.

50. Gonzalez-Vera JA, Morris MC. Fluorescent Reporters and Biosensors for Probing the Dynamic Behavior of Protein Kinases. Proteomes. 2015;3(4):369-410. Epub 20151104. doi: 10.3390/proteomes3040369. PubMed PMID: 28248276; PMCID: PMC5217393.

51. Violin JD, Zhang J, Tsien RY, Newton AC. A genetically encoded fluorescent reporter reveals oscillatory phosphorylation by protein kinase C. J Cell Biol. 2003;161(5):899-909. Epub 20030602. doi: 10.1083/jcb.200302125. PubMed PMID: 12782683; PMCID: PMC2172956.

52. Fuller BG, Lampson MA, Foley EA, Rosasco-Nitcher S, Le KV, Tobelmann P, Brautigan DL, Stukenberg PT, Kapoor TM. Midzone activation of aurora B in anaphase produces an intracellular phosphorylation gradient. Nature. 2008;453(7198):1132-6. Epub 20080507. doi: 10.1038/nature06923. PubMed PMID: 18463638; PMCID: PMC2724008.

53. Zheng CR, Singh A, Libby A, Silver PA, Libby EA. Modular and Single-Cell Sensors of Bacterial Ser/Thr Kinase Activity. ACS Synth Biol. 2021;10(9):2340-50. Epub 20210831. doi: 10.1021/acssynbio.1c00250. PubMed PMID: 34463482; PMCID: PMC8498941.

54. Scharf BE. Summary of useful methods for two-component system research. Curr Opin Microbiol. 2010;13(2):246-52. doi: 10.1016/j.mib.2010.01.006. PubMed PMID: 20138001.

55. Wright DP, Ulijasz AT. Regulation of transcription by eukaryotic-like serine-threonine kinases and phosphatases in Gram-positive bacterial pathogens. Virulence. 2014;5(8):863-85. doi: 10.4161/21505594.2014.983404. PubMed PMID: 25603430; PMCID: PMC4601284.

56. Durocher D, Taylor IA, Sarbassova D, Haire LF, Westcott SL, Jackson SP, Smerdon SJ, Yaffe MB. The molecular basis of FHA domain:phosphopeptide binding specificity and implications for phospho-dependent signaling mechanisms. Mol Cell. 2000;6(5):1169-82. doi: 10.1016/s1097-2765(00)00114-3. PubMed PMID: 11106755.

57. UniProt C. UniProt: the universal protein knowledgebase in 2021. Nucleic Acids Res. 2021;49(D1):D480-D9. doi: 10.1093/nar/gkaa1100. PubMed PMID: 33237286; PMCID: PMC7778908.

58. Xue Y, Gao X, Cao J, Liu Z, Jin C, Wen L, Yao X, Ren J. A summary of computational resources for protein phosphorylation. Curr Protein Pept Sci. 2010;11(6):485-96. doi: 10.2174/138920310791824138. PubMed PMID: 20491621.

59. Li Z, Li S, Luo M, Jhong JH, Li W, Yao L, Pang Y, Wang Z, Wang R, Ma R, Yu J, Huang Y, Zhu X, Cheng Q, Feng H, Zhang J, Wang C, Hsu JB, Chang WC, Wei FX, Huang HD, Lee TY. dbPTM in 2022: an updated database for exploring regulatory networks and functional associations of protein post-translational modifications. Nucleic Acids Res. 2022;50(D1):D471-D9. doi: 10.1093/nar/gkab1017. PubMed PMID: 34788852; PMCID: PMC8728263.

60. Hornbeck PV, Kornhauser JM, Latham V, Murray B, Nandhikonda V, Nord A, Skrzypek E, Wheeler T, Zhang B, Gnad F. 15 years of PhosphoSitePlus(R): integrating post-translationally modified sites, disease variants and isoforms. Nucleic Acids Res. 2019;47(D1):D433-D41. doi: 10.1093/nar/gky1159. PubMed PMID: 30445427; PMCID: PMC6324072.

61. Lin S, Wang C, Zhou J, Shi Y, Ruan C, Tu Y, Yao L, Peng D, Xue Y. EPSD: a well-annotated data resource of protein phosphorylation sites in eukaryotes. Brief Bioinform. 2021;22(1):298-307. doi: 10.1093/bib/bbz169. PubMed PMID: 32008039.

62. Shi Y, Zhang Y, Lin S, Wang C, Zhou J, Peng D, Xue Y. dbPSP 2.0, an updated database of protein phosphorylation sites in prokaryotes. Sci Data. 2020;7(1):164. Epub 20200529. doi: 10.1038/s41597-020-0506-7. PubMed PMID: 32472030; PMCID: PMC7260176.

63. Macek B, Gnad F, Soufi B, Kumar C, Olsen JV, Mijakovic I, Mann M. Phosphoproteome analysis of E. coli reveals evolutionary conservation of bacterial Ser/Thr/Tyr phosphorylation. Mol Cell Proteomics. 2008;7(2):299-307. Epub 20071015. doi: 10.1074/mcp.M700311-MCP200. PubMed PMID: 17938405.

64. Hasan MM, Rashid MM, Khatun MS, Kurata H. Computational identification of microbial phosphorylation sites by the enhanced characteristics of sequence information. Sci Rep. 2019;9(1):8258. Epub 20190604. doi: 10.1038/s41598-019-44548-x. PubMed PMID: 31164681; PMCID: PMC6547684.

65. Wang M, Wang T, Wang B, Liu Y, Li A. A Novel Phosphorylation Site-Kinase Network-Based Method for the Accurate Prediction of Kinase-Substrate Relationships. Biomed Res Int. 2017;2017:1826496. Epub 20171012. doi: 10.1155/2017/1826496. PubMed PMID: 29312990; PMCID: PMC5660750.

66. Liu Y, Wang M, Xi J, Luo F, Li A. PTM-ssMP: A Web Server for Predicting Different Types of Post-translational Modification Sites Using Novel Site-specific Modification Profile. Int J Biol Sci. 2018;14(8):946-56. Epub 20180522. doi: 10.7150/ijbs.24121. PubMed PMID: 29989096; PMCID: PMC6036757.

67. Li Z, Wu P, Zhao Y, Liu Z, Zhao W. Prediction of serine/threonine phosphorylation sites in bacteria proteins. Adv Exp Med Biol. 2015;827:275-85. doi: 10.1007/978-94-017-9245-5_16. PubMed PMID: 25387970.

68. Zhang QB, Yu K, Liu Z, Wang D, Zhao Y, Yin S, Liu Z. Prediction of prkC-mediated protein serine/threonine phosphorylation sites for bacteria. PLoS One. 2018;13(10):e0203840.

743 Epub 20181002. doi: 10.1371/journal.pone.0203840. PubMed PMID: 30278050; PMCID:
744 PMC6168130.

745 69. Miller ML, Soufi B, Jers C, Blom N, Macek B, Mijakovic I. NetPhosBac - a predictor for
746 Ser/Thr phosphorylation sites in bacterial proteins. Proteomics. 2009;9(1):116-25. doi:
747 10.1002/pmic.200800285. PubMed PMID: 19053140.

748 70. Wang D, Zeng S, Xu C, Qiu W, Liang Y, Joshi T, Xu D. MusiteDeep: a deep-learning
749 framework for general and kinase-specific phosphorylation site prediction. Bioinformatics.
750 2017;33(24):3909-16. doi: 10.1093/bioinformatics/btx496. PubMed PMID: 29036382; PMCID:
751 PMC5860086.

752 71. Obenauer JC, Cantley LC, Yaffe MB. Scansite 2.0: Proteome-wide prediction of cell
753 signaling interactions using short sequence motifs. Nucleic Acids Res. 2003;31(13):3635-41. doi:
754 10.1093/nar/gkg584. PubMed PMID: 12824383; PMCID: PMC168990.

755 72. Blom N, Sicheritz-Ponten T, Gupta R, Gammeltoft S, Brunak S. Prediction of post-
756 translational glycosylation and phosphorylation of proteins from the amino acid sequence.
757 Proteomics. 2004;4(6):1633-49. doi: 10.1002/pmic.200300771. PubMed PMID: 15174133.

758 73. Gnad F, Ren S, Cox J, Olsen JV, Macek B, Oroshi M, Mann M. PHOSIDA
759 (phosphorylation site database): management, structural and evolutionary investigation, and
760 prediction of phosphosites. Genome Biol. 2007;8(11):R250. doi: 10.1186/gb-2007-8-11-r250.
761 PubMed PMID: 18039369; PMCID: PMC2258193.

762 74. Gnad F, Gunawardena J, Mann M. PHOSIDA 2011: the posttranslational modification
763 database. Nucleic Acids Res. 2011;39(Database issue):D253-60. Epub 20101116. doi:
764 10.1093/nar/gkq1159. PubMed PMID: 21081558; PMCID: PMC3013726.

765 75. Ma R, Li S, Li W, Yao L, Huang HD, Lee TY. KinasePhos 3.0: Redesign and expansion of
766 the prediction on kinase-specific phosphorylation sites. Genomics Proteomics Bioinformatics.
767 2022. Epub 20220630. doi: 10.1016/j.gpb.2022.06.004. PubMed PMID: 35781048.

768 76. Chen M, Zhang W, Gou Y, Xu D, Wei Y, Liu D, Han C, Huang X, Li C, Ning W, Peng D,
769 Xue Y. GPS 6.0: an updated server for prediction of kinase-specific phosphorylation sites in
770 proteins. Nucleic Acids Res. 2023;51(W1):W243-W50. doi: 10.1093/nar/gkad383. PubMed PMID:
771 37158278; PMCID: PMC10320111.

772 77. Sapoval N, Aghazadeh A, Nute MG, Antunes DA, Balaji A, Baraniuk R, Barberan CJ,
773 Dannenfelser R, Dun C, Edrisi M, Elworth RAL, Kille B, Kyrillidis A, Nakhleh L, Wolfe CR, Yan
774 Z, Yao V, Treangen TJ. Current progress and open challenges for applying deep learning across
775 the biosciences. Nat Commun. 2022;13(1):1728. Epub 20220401. doi: 10.1038/s41467-022-29268-7.
776 PubMed PMID: 35365602; PMCID: PMC8976012.

78. Eraslan G, Avsec Z, Gagneur J, Theis FJ. Deep learning: new computational modelling techniques for genomics. Nat Rev Genet. 2019;20(7):389-403. doi: 10.1038/s41576-019-0122-6. PubMed PMID: 30971806.

79. Gupta R, Srivastava D, Sahu M, Tiwari S, Ambasta RK, Kumar P. Artificial intelligence to deep learning: machine intelligence approach for drug discovery. Mol Divers. 2021;25(3):1315-60. Epub 20210412. doi: 10.1007/s11030-021-10217-3. PubMed PMID: 33844136; PMCID: PMC8040371.

80. Luo F, Wang M, Liu Y, Zhao XM, Li A. DeepPhos: prediction of protein phosphorylation sites with deep learning. Bioinformatics. 2019;35(16):2766-73. doi: 10.1093/bioinformatics/bty1051. PubMed PMID: 30601936; PMCID: PMC6691328.

81. Lin M, Xiao D, Geddes TA, Burchfield JG, Parker BL, Humphrey SJ, Yang P. SnapKin: a snapshot deep learning ensemble for kinase-substrate prediction from phosphoproteomics data. bioRxiv. 2021:2021.02.23.432610. doi: 10.1101/2021.02.23.432610.

82. Deznabi I, Arabaci B, Koyuturk M, Tastan O. DeepKinZero: zero-shot learning for predicting kinase-phosphosite associations involving understudied kinases. Bioinformatics. 2020;36(12):3652-61. doi: 10.1093/bioinformatics/btaa013. PubMed PMID: 32044914; PMCID: PMC7320620.

83. Qin GM, Li RY, Zhao XM. PhosD: inferring kinase-substrate interactions based on protein domains. Bioinformatics. 2017;33(8):1197-204. doi: 10.1093/bioinformatics/btw792. PubMed PMID: 28031187.

84. Trost B, Maleki F, Kusalik A, Napper S. DAPPLE 2: a Tool for the Homology-Based Prediction of Post-Translational Modification Sites. J Proteome Res. 2016;15(8):2760-7. Epub 20160713. doi: 10.1021/acs.jproteome.6b00304. PubMed PMID: 27367363.

85. Ayati M, Wiredja D, Schlatzer D, Maxwell S, Li M, Koyuturk M, Chance MR. CoPhosK: A method for comprehensive kinase substrate annotation using co-phosphorylation analysis. PLoS Comput Biol. 2019;15(2):e1006678. Epub 20190227. doi: 10.1371/journal.pcbi.1006678. PubMed PMID: 30811403; PMCID: PMC6411229.

86. Yilmaz S, Ayati M, Schlatzer D, Cicek AE, Chance MR, Koyuturk M. Robust inference of kinase activity using functional networks. Nat Commun. 2021;12(1):1177. Epub 20210219. doi: 10.1038/s41467-021-21211-6. PubMed PMID: 33608514; PMCID: PMC7895941.

87. Ayati M, Yilmaz S, Chance MR, Koyuturk M. Functional characterization of co-phosphorylation networks. Bioinformatics. 2022;38(15):3785-93. doi: 10.1093/bioinformatics/btac406. PubMed PMID: 35731218; PMCID: PMC9344848.

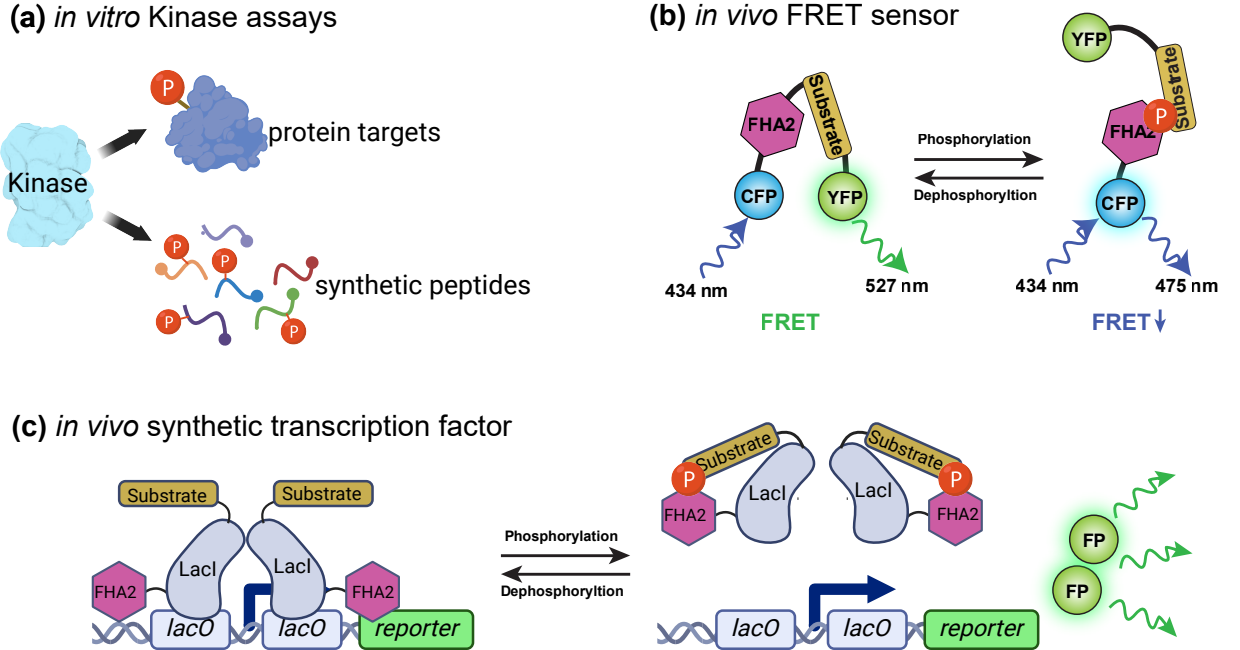
88. Chawla Y, Upadhyay S, Khan S, Nagarajan SN, Forti F, Nandicoori VK. Protein kinase B (PknB) of Mycobacterium tuberculosis is essential for growth of the pathogen in vitro as well as

812 for survival within the host. J Biol Chem. 2014;289(20):13858-75. Epub 20140404. doi:
813 10.1074/jbc.M114.563536. PubMed PMID: 24706757; PMCID: PMC4022859.

814 89. Fernandez P, Saint-Joanis B, Barilone N, Jackson M, Gicquel B, Cole ST, Alzari PM. The
815 Ser/Thr protein kinase PknB is essential for sustaining mycobacterial growth. J Bacteriol.
816 2006;188(22):7778-84. Epub 20060915. doi: 10.1128/JB.00963-06. PubMed PMID: 16980473;
817 PMCID: PMC1636329.

818 90. Sassetti CM, Boyd DH, Rubin EJ. Genes required for mycobacterial growth defined by
819 high density mutagenesis. Mol Microbiol. 2003;48(1):77-84. doi: 10.1046/j.1365-2958.2003.03425.x.
820 PubMed PMID: 12657046.

821


(a) in vitro Kinase assays
Kinase
P
protein targets
synthetic peptides
(b) in vivo FRET sensor
FHA2
Substrate
CFP
YFP
434 nm
527 nm
FRET
Phosphorylation
Dephosphoryltion
P
434 nm
475 nm
FRET↓
(c) in vivo synthetic transcription factor
Substrate
LacI
FHA2
lacO
reporter
Phosphorylation
Dephosphoryltion
FP

Input: experimental data

Primary Sequence

IPEDDEATKAIPIIT

Chemical, Structural Properties

HHPPPPHPPHHHHHP
==----==========

Evolutionary Conservation

Co-phosphorylation patterns

Model: machine learning

- Support vector machine (SVM)
- Random Forest (RF)
- Deep Neural Network (DNN)

Results: testable predictions

Phosphosite?
(binary classification: yes/no)

Targeted by which kinase?
In what conditions or cell types?
(multi-class classification)

Similar to what other substrates?
(Conservation analysis, few-shot learning)