SAPH-ire TFx: A Machine Learning Recommendation Method and Webtool for the Prediction of Functional Post-Translational Modifications

Short Title: Improving PTM functional prioritization through neural networks

Nolan English[1,2] and Matthew Torres[1,2,*]

[1]School of Biological Sciences, Georgia Institute of Technology, Atlanta, GA 30332

[2]Quantitative Biosciences Program, Georgia Institute of Technology, Atlanta, GA 30332

* To Whom Correspondence Should be Addressed:

Matthew Torres
Associate Professor
School of Biological Sciences
Engineered Biosystems Building, 4009
Georgia Institute of Technology
950 Atlantic Drive
Atlanta, Georgia 30332
mtorres35@gatech.edu

ABSTRACT

Post-translational Modifications (PTMs), chemical or proteinaceous covalent alterations to the side chains of amino acid residues in proteins, are a rapidly expanding feature class of significant importance in cell biology. Due to a high burden of experimental proof and the lack of effective means for experimentalists to prioritize PTMs by functional significance, currently less than ~2% of all PTMs have an assigned biological function. Here, we describe a new artificial neural network model, SAPH-ire TFx for the functional prediction of experimentally observed eukaryotic PTMs. Unlike previous functional PTM prioritization models, SAPH-ire TFx is optimized with both receiver operating characteristic (ROC) and recall metrics that maximally capture the range of diverse feature sets comprising the functional modified eukaryotic proteome. The tool was through systematic evaluation of input features, model architectures, training procedures, and interpretation metrics using a 2018 training dataset of 430,750 PTMs containing 7,480 PTMs with literature-supported evidence of biological function. The resulting model was used to classify an expanded 2019 dataset of 512,015 PTMs (12,867 known functional) containing 102,475 PTMs unencountered in the original dataset. Model output from the 2019 extended dataset was benchmarked against pre-existing prediction models, revealing superior performance in classification of functional and/or disease-linked PTM sites. Finally, a dynamic web interface provides customizable graphical and tabular visualization of PTM and SAPH-ire TFx data within the context of all modifications within a protein family, exposing several metrics by which important functional PTMs can be identified for investigation.

AUTHOR SUMMARY

The modification of proteins after they are translated is an important process that can control the structure and function of the proteins on which they occur. Hundreds of different types of modification happen at some point during the lifetime of every protein in eukaryotic cells and play an essential role in cellular processes such as cell division, cell communication, gene regulation. Using current state-of-the-art detection tools, the rate at which post-translational modifications are detected now far surpasses the rate at which they can be investigated for functionality. Furthermore, not all modifications detected are functional, making it difficult to determine into which modifications one should invest experimental effort. Here, we describe a new computational tool – SAPH-ire TFx – capable of predicting functional modification sites from large-scale datasets, and consequently focus experimental effort towards only those modifications that are likely to be biologically significant. We show that the tool performs well across multiple datasets within which known functional modifications are scattered; and we show that the tool outperforms prior functional prioritization tools. Finally, we also provide a user-friendly web tool for experimentalists to investigate SAPH-ire TFx output for proteins and protein families of interest.

INTRODUCTION

Post-translational modifications (PTMs) alter the chemistry of amino acid residues within translated proteins and thereby have the potential to expand the function and regulatory control of proteins beyond the limits of the genome (1). By definition PTMs modify the structure of proteins, and can consequently alter their function through changes in biomolecular interactions, protein localization and/or protein stability (2,3). PTMs can act on long or short timescales that allow for dynamic control and response of a cellular proteome to changing environments or cellular phases that ultimately shape cellular phenotype (4). Concomitantly, disruption to either the amino acid or modification of highly functional PTM sites can lead to cellular dysfunction and disease (5–7).

Since the advent of high-throughput protein sequencing by mass spectrometry, hundreds of different types of PTM have been found to occur on the sidechains of most of the 20 amino acids (8,9). However, the rate at which PTM data are generated – a parallel process involving hundreds of independent labs – far surpasses the rate at which it is being curated and/or processed for interpretation – a task undertaken by a much smaller set of labs and institutions (10,11). Today, somewhere between 700K to 1M PTMs represented by nearly 300 unique types are, more or less, accessible in the public domain. Although, such estimates change drastically every year as a result of new technologies that improve detection of any one PTM in the proteome (12). Of these, ~116,000 correspond to human phosphorylation sites – representing about half of the number of phosphosites estimated for the human proteome (13). Thus, the total number of curated PTMs is likely an underestimate of PTMs that exist in the collective eukaryotic proteome.

With the exponential increase in PTM data over the last 15 years, a major question is whether all PTMs detected (*accurately*) are functionally important – a question not easily answered due to the extremely high burden of experimental proof needed to prove functionality, which involves significant time, cost, and specific expertise for any given protein. This is compounded by unnecessary redundancy in experimental effort, and the tendency of most labs

107 not to report non-functional results, and the longstanding and widely accepted premise that not
108 all PTMs are functionally equivalent. Indeed, this underlying dilemma tends to promote a
109 perspective that the study of PTMs is risky and quite possibly not worth the effort, especially in
110 light of the exponential increase in data content that can appear overwhelming to a protein
111 experimentalist. PTM site features such as evolutionary conservation (14–16), PTM co-
112 localization (17–19), protein structural constraints (20,21), and other single features, have been
113 shown to be mildly predictive for function and offer useful means of filtering PTM data for likely
114 important modifications. More recently, machine learning models that incorporate multiple single
115 features have been proven to have significantly greater predictive power in comparison to single
116 features alone (22–24). In general, these models provide a priority score that is effective for rank
117 ordering PTM function potential, and are shown to be successful in identifying new functional
118 PTMs across the eukaryotic proteome, but unfortunately do little to confidently *recommend* PTMs
119 for experimental analysis due to very poor recall of true positive data.

120 Here we describe the development and application of a neural network model, SAPH-ire
121 TFx, optimized for the recommendation of functional PTMs in whole sequence protein families.
122 We used this model to prioritize the potential for function for 512,015 unique PTMs (8,373 known
123 functional) across 38,231 proteins and 763 organisms. This is the first model of its type in which
124 recall has been used for optimization, and we show that this results in significantly improved
125 predictive properties compared to previous versions of SAPH-ire and other functional prioritization
126 models. We also describe a new web-based resource and API that provides an interactive and
127 customizable tool designed to enable protein experimentalists with SAPH-ire TFx-generated PTM
128 recommendations and contextual visualization of PTMs in the context of the protein families in
129 which they are members.

130

131 **METHODS**

132 **PTMs and multiple sequence alignment**

133 SAPH-ire is PTM agnostic and includes 56 different PTM types (PTMtype), the bulk of
134 which correspond to phosphorylation, ubiquitination, acetylation, methylation, N-linked
135 glycosylation, and sumoylation (**Table S1**). PTMs were collected from multiple sources including
136 PhosphositePlus (25), SysPTM (26), and dbPTM (27). Each PTM was mapped to UniProt
137 identifiers (UID) and validated by matching the native position (NP) and residue (res) of the
138 curated PTM to UniProt sequences verified for 100% sequence identity using BLAST (28).
139 Isoforms, although rare in the PTM dataset, were also included. The final PTM dataset contains
140 629,740 unique PTMs (identified by UID-NP-res-PTMtype).

141 UID entries were mapped to whole sequence protein families using InterPro (29) followed
142 by multiple sequence alignment of family-linked UniProt sequences using MUSCLE with default
143 parameters (30). Families with fewer than 2 members containing at least 1 PTM per member were
144 excluded. This process resulted in a final 512,015 PTMs mapped to 8,039 families (**Table S2**)
145 containing 38,231 UIDs representing 763 eukaryotic organisms (**Table S3**).

146

147 **Feature selection**

148 SAPH-ire features were derived from Modified Alignment Positions (MAPs) corresponding
149 to family alignment positions that harbor at least one PTM, as described previously in detail
150 (20,22,23). A total of eleven features are extracted from 319,981 MAPs (containing the 512,015
151 PTMs) for inclusion in neural network models described below (**Table S4**). The number of unique
152 PTMs observed in the alignment position (PTM count; PTMc), the PTM residue conservation
153 within the alignment position (PRC), the predicted disorder of the modified residue (Dis), and the
154 number of unique modified residues within the alignment position (Modified residue count; MRc)
155 all provide the model with an evolutionary conservation-based perspective on the MAP – a feature
156 that has been shown to be an effective in the past (14). The next group of features provide
157 information on the local environment of the PTM by providing the number of neighboring PTMs

158 (Neighbor PTMc; NPTMc), neighboring MAPs (Neighbor count, Nc), and neighboring MAPs with
159 functional impact (Known neighbor count, KNc). Neighboring residue context has been shown to
160 be an effective predictive feature in the past by us and others (15,18). Lastly, the number of
161 sources that have observed the PTM (Observation source count; OBSrc), a normalized version
162 of this feature that takes family membership into account (OBSrc_pf), and the number of UniProt
163 entries associated with the MAP (Alignment position member count, AP-MEM) are utilized in this
164 model for the first time here.
165

166 **Model implementation, cost function optimization, training, and model selection**

167 The SAPH-ire TFx neural network model and modified cost function (defined below) were
168 implemented in Tensorflow using the estimator API (31). PTMs and MAPs were processed into
169 features using Python 3.7.3 and Pandas 0.24.0 (32). PTMs with at least two sources corroborating
170 evidence of biological functional, defined by PhosphositePlus (25), were treated as the positive
171 class with all others being treated as negative. MAPs with a single source were treated as
172 negative because they lacked independent confirmation of functional significance and because
173 their inclusion weakened model performance. The cost for false negatives (misclassified known
174 functional PTMs) was weighted at a 4 to 1 ratio to false positives (PTMs with unknown function
175 classified as functional) to reflect the goal of recommending unstudied PTMs for research. Below
176 this 4/1 ratio the performance suffered, and above it there was no significant improvement while
177 the model began to exhibit signs of overfitting (data not shown).

178 Neural network models of various structure (in terms of connectivity, activation function,
179 cost function weighting, etc.) were generated in batches of 100 or 200 depending on architecture
180 complexity. The training set for these models were bootstrapped in order to over-represent the
181 positive class to avoid sample distribution biasing (33). Models were trained using a 33% holdback
182 rate, with this holdback being used to evaluate batches of models against the same evaluation

183 set. Model selection relied on Receiver Operating Characteristic (ROC) and Recall summarization
184 metrics integrated by an Fzero score defined as:

185 $$Fzero = 2 * \frac{auroc * recall}{auroc + recall}$$

186 Optimal models were defined as those with the greatest Fzero score.

187

188 **Pathogenic enrichment analysis**

189 Genetic mutations and the curated interpretation of their significance to disease was
190 collected from Clinvar (34). The proteins affected by genetic mutations were filtered for single
191 nucleotide polymorphisms (SNPs) that alter PTM sites present within the SAPH-ire dataset. Then,
192 these sites were aggregated by SAPH-ire MAP and separated into one of four categories
193 according the NIH classification: Benign, Likely Benign, Pathogenic, or Likely Pathogenic. Only
194 Benign and Pathogenic categories were used for analysis.

195

196 **Graphical and statistical data analyses**

197 Graphical and statistical data analyses were achieved using a combination of R (35), and
198 JMP 14.1 (SAS Institute Inc.).

199

200 **SAPH-ire web-tool implementation**

201 The results of the SAPH-ire model have been made accessible through Web and
202 Application Programming Interface available at (https://saphire.biosci.gatech.edu). Briefly, SAPH-
203 ire input and output data were uploaded into a Mongo database made accessible through API
204 endpoints created using the Eve package in Python (https://docs.python-eve.org/en/stable/). The
205 Web interface, which also pulls from the API as its data source, was built in Vue.js relying on
206 Vuetify for design elements and the open source Plotly.js library for delivering graphical summary
207 of the data.

208

209 **RESULTS**

210 **Optimizing SAPH-ire TFx model architecture and performance through recall**

211 Previous models (including our own) have employed the use of quantitative features in
212 isolation (15,16) or in combination (20,22–24,36)) to rank PTMs according to potential for function.
213 In the process of re-evaluating this approach, we concluded that rank-based models are
214 incongruous with functional evidence from biological data, wherein functional impact is broadly
215 defined and highly variable. In contrast, recommendation-based models, in which groups of
216 objects that fit some criteria without asserting any specific entity within that group is a better fit for
217 that criteria than any other, more accurately reflects the intended goal of PTM prioritization –
218 namely the recommendation of likely-functional PTMs for investigative protein research. To build
219 a model capable of recommendation-based classification, we introduced a new summation
220 metric, *Fzero,* that emphasizes not only ROC metrics, but also captures the recall of true positive
221 data (see Methods). Due to the intended goal of SAPH-ire to identify "potential positives", a high
222 precision (true positives captured / total positives) was not only unwanted but indicative of a poor
223 model and therefore not included as a summary metric.

224 For model training, we used a SAPH-ire training dataset generated in 2018, consisting of
225 435,750 PTMs coalesced by multiple sequence alignment into 272,968 MAPs (**Figure 1A**; see
226 Methods). From each of more than a dozen architecture and training permutations, 200 models
227 were stochastically trained and evaluated by Fzero and Recall and then filtered to identify the
228 most consistently high performing architectures (Batch 1). The agreement between most of the
229 models made collective intelligence approaches redundant. Therefore, an additional series of
230 models trained with an added feature (family-weighted observation source count) was generated
231 (Batch 2). In general, both batches varied only slightly in terms of ROC AUC (0.68 – 0.73), but
232 varied dramatically in recall (0.5 – 0.8), suggesting that significant gains in model performance

were achieved by considering recall in addition to ROC AUC (**Figure 1B**). The top performing models from these two independent evaluations were averaged together to represent the final SAPH-ire score that outperformed either top model alone (**Figure 1B** (inset)). This final score resulted in excellent predictive (AUC = 0.792) and recall (AUC = 0.694) performance (**Figure 1C**).

Figure 1. Numerical summary of SAPH-ire TFx development, training, and performance. (A) Diagram of the training dataset with positive and negative classes indicated. Note that Modified Alignment Positions (MAPs) with less than 2 literature sources of support were included in the negative class as they were not used at any point for training the model (see methods). (B) Plot of true positive recall versus ROC AUC versus Fzero (color) for the top two of 200 models generated using ten (Batch 1) or eleven (Batch 2) features (MAP-level analysis). Dashed box (on same scale) indicates performance of a combined model (SAPH-ire TFx) that is generated by taking the average score of Batches 1 and 2 on the expanded dataset (shown further below). (C) ROC and recall threshold curves for the training dataset are shown (based on MAP-level analysis).

Evaluating SAPH-ire TFx on an expanded PTM dataset

Next, we evaluated the performance of SAPH-ire TFx on an expanded PTM dataset in which an additional 102,475 PTMs were modeled – 3,233 of which were known functional but to which the algorithm was blind during its initial training (**Figure 2A**). With this expanded dataset, a completely new set of protein family alignments were also generated and model input features extracted. Thus, the expanded dataset represents a completely new dataset never before seen by the model. Overall ROC AUC and recall for the outcoming data performed better than on the original training dataset in terms of ROC and recall-threshold metrics (AUC_{ROC} 0.794 vs. 0.792, and AUC_{Recall} 0.764 versus 0.694) (**Figure 2B**). We further evaluated the model by comparing the distribution of SAPH-ire TFx scores for unknown and known functional MAPs wherein each MAP was placed into a bin described by the known function source count (KFSC) – a count of the unique literature sources containing evidence of functional impact for a PTM within the MAP, which serves as a proxy for confidence in the assignment of known function status (**Figure 2C**). This analysis revealed that SAPH-ire scored PTMs with greater KFSC values significantly higher ($p<0.0001$) than those with lower KFSC (3+ref/2ref; 2ref/1ref, and 1ref/0refs). This trend was also

evident for the 3,233 “New” PTMs of known functional status in the expanded dataset to which algorithm was completely blind during initial training (**Figure 2C**, dashed lines). Taken together, these data show that SAPH-ire TFx is an effective recommendation model for functional PTMs and is robust in its performance across independent datasets.

Figure 2. Application of SAPH-re TFx to an expanded dataset. The SAPH-ire TFx model summarized in figure 1 was applied to a completely new expanded dataset derived from new protein sequence alignments and containing additional PTMs curated in 2019. (A) Venn diagram showing the relationship between the training and expanded datasets. Due to stringent self-imposed UniProt sequence filters, a small fraction of PTMs in the training dataset were lost, and the new expanded dataset contained 102,475 newly curated PTMs. (B) ROC and recall curves for SAPH-ire TFx results from the expanded dataset. (C) Frequency distribution of SAPH-ire TFx scores relative to true positive status in terms of known function source count (1, 2, or 3+ references). Area contained by solid lines corresponds to total expanded PTM dataset. Area contained by dashed lines corresponds to expanded PTMs not contained in the original training dataset from figure 1. All statistical data shown is at the MAP level.

Biological Significance

We hypothesized that MAPs containing sites of PTM that coincide with a known pathogenic mutation should have higher SAPH-ire scores than those that are benign when silenced, if the model can recommend functionally significant PTMs. To test this hypothesis, we compared the SAPH-ire TFx scores of PTMs whose sites correspond to mis-sense mutations curated in the ClinVar disease linkage database (34). We observed significantly higher scores ($p=0.0003$) for PTMs coincident with pathogenic mutations (median score = 0.475) compared to benign mutations (median score = 0.033) (**Figure 3**). As expected from the sigmoid response curve of the SAPH-ire model, the score distributions exhibit two regions of enrichment. The first region encompasses PTMs with a score less than .05, in which there is a slightly greater enrichment (37%) for benign versus pathogenic mutations. The second region represents the top 10% of PTMs by SAPH-ire TFx score (score above .95), and above this threshold there is significant overrepresentation (156%) of pathogenic compared to benign mutations. The enrichment of high scoring pathogenic-SNP-coincident PTMs follows with the corollary that if the

295 model is capable of distinguishing between PTMs with and without functional impact, then MAPs
296 that coincide with disease-causing SNPs should be overrepresented compared to benign SNPs
297 as the SAPH-ire score increases. However, it's also important to note that there are also several
298 SNP/PTM sites for which the model does not support this corollary. This seeming incongruity, can
299 be readily explained as not every SNP labeled as pathogenic is necessarily dependent on the
300 effect of lost PTM. Alternatively, PTMs coincident with SNP mutations that are labelled benign
301 could still have functional significance but not be pathogenic when silenced.

302

303 **Figure 3. SAPH-ire TFx selective enrichment of PTMs coincident with causal disease-linked**
304 **mutations.** PTMs contained in the expanded dataset that are coincident with disease-linked
305 mutations curated in the ClinVar database as benign or pathogenic were compared by frequency
306 distribution of SAPH-ire TFx scores. The median score for each class is indicated as a red bar in
307 the box plot. Significance testing (p-value) was determined using a T-test.

308

309 **Recommendation thresholding is essential for model interpretation**

310 Thresholding is an essential component of a recommendation model that provides a useful
311 filter for large datasets. Prior models for the functional prioritization of PTMs, while providing
312 quantitative distributions of potential function, have largely struggled to provide well-defined
313 thresholds that support confident recommendations for further investigation. To address this
314 problem, we modeled the tradeoff between true and false positive rates using ROC curves. Our
315 goal was to set a minimum threshold score over which MAPs could be considered having a high
316 chance of functionality. Due to our desire for SAPH-ire scores to be agnostic across different PTM
317 types, we first considered that the selected thresholds must not create a bias in distribution of
318 PTMs occurring above that threshold. To evaluate this, we plotted the percent representation of
319 each of the most common PTMs in the dataset relative to SAPH-ire TFx score (**Figure 4A**). The
320 relative representation of each PTM type deflected significantly above a score of 0.9897 but was
321 stable below this point and above 0.945, the range between which we defined as ideal for
322 thresholding.

Next, we evaluated the ROC curves for the highest confidence true positive MAPs (KFSC >1, >2) (**Figure S1**), and unfurled each curve to reveal the independent rates for true and false positives with respect to the SAPH-ire TFx score. From these curves, three thresholds were chosen within the ideal range (0.95, 0.9719, and 0.987) that strike a balance between true positive hits and false positive recommendations (**Figure 4B**). These recommendation thresholds provide useful landmarks to interpret SAPH-ire TFx scores for a protein or family of interest, as shown here for family IPR000043 (**Figure 4C**). They also serve as benchmarks for evaluating other models and future versions of SAPH-ire.

Figure 4. Recommendation thresholding for SAPH-ire TFx. (A) Plot of the percent representation for different PTM types relative to SAPH-ire TFx score, revealing an ideal threshold range inside which no one PTM becomes over or underrepresented. (B) Unfurled pareto fronts showing true positive (TP) and false positive (FP) rates above given SAPHire score. Rates shown are KFSC > 1 (lower confidence) or KFSC > 2 (higher confidence). Dashed vertical lines represent chosen thresholds where TP and FP rates are as follows (KFSC >2): 0.95 – TP=0.82, FP=0.11; 0.972 – TP=0.7, FP=0.07; 0.987 – TP=0.53, FP=0.04. (C) Representative rank-ordered SAPH-ire TFx plot for family IPR000043 (Adenosylhomocysteinase-like family) with indicated thresholds shown for reference. Shown on an exponential scale to emphasize rank distribution.

Comparison to other models

SAPH-ire TFx is a unique model that is one of a small number of published algorithms aimed at functional prioritization of PTMs, and also the first recommendation model for functional PTMs. We therefore sought to draw comparisons with these models to gauge overall performance. Two predominant models currently exist in the public domain: SAPH-ire FPx (23) – an 8-feature neural network PTM ranking model; and a Phosphosite Functional Score (PFS) model (36) – a 59-feature gradient boosting machine learning model trained to identify functional phosphosites.

To evaluate the three models equivalently, we compared model scores for phosphosites represented in all three datasets. PFS was built using a dataset containing 116,268 phosphosites that resulted from selective re-analysis of raw mass spectrometry data files collected from a broad range of eukaryotic organisms (36). Comparing our source database to the PFS dataset revealed

354 71% overlap (82,279 phosphosites), however, this number dropped in response to strict protein
355 family membership criteria (see Methods). Of PTMs that fall within InterPro whole sequence
356 families, 236,982 represent unique phosphosites that were analyzed by SAPH-ire TFx, and
357 49,935 of these overlap with ~43% of the PFS dataset (**Figure 5A**). Inclusion of SAPH-ire FPx
358 data, which was based on PTMs curated in early 2017, resulted in a final comparable dataset of
359 24,695 phosphosites, which were compared in terms of score distribution as well as ROC and
360 recall metrics.

361

362 **Figure 5. Benchmarking SAPH-ire TFx against prior PTM functional prioritization models.**
363 SAPH-ire TFx was compared head-to-head with the two prior machine learning models for
364 functional prioritization: SAPH-ire FPx (23) and Phosphosite Functional Score (PFS) (36). (A)
365 Venn diagram describing the overlap between the expanded dataset (reported here) and
366 phosphosite datasets for the other two models. Three-way model comparisons were conducted
367 with 24,695 phosphosites. Pairwise model comparisons (PFS vs. SAPH-ire TFx) were also
368 conducted with 49,935 overlapping phosphosites (Figure S2). (B) Comparison of the score
369 distributions for PTMs binned by category of unknown function, known function (1, 2, or 3+
370 sources), or known by neighbor (KbN) determined by SAPH-ire TFx protein family alignments.
371 Dashed lines indicate the thresholds quantitatively determined for SAPH-ire TFx or loosely
372 recommended by other models. (C) Inter-quartile range relative to known functional status, based
373 on the distributions shown in B. (D) Comparison of ROC and recall curves for each model. (E) Pie
374 chart representation of the percentage of recalled versus mis-called (Missed) PTMs based on
375 thresholds shown in B [0.95 threshold used for SAPH-ire TFx] (top). Number of recommendations
376 deduced from these percentages applied to the whole dataset for each model (bottom).
377 Recommendations are also shown for each of the thresholds established for SAPH-ire TFx in
378 figure 4.
379

380 In general, SAPH-ire FPx and PFS perform similarly in most respects – in part because
381 they were both built to maximize ROC AUC but not recall. Both models result in broad and
382 overlapping score distributions that are significantly different but modestly distinct between sites
383 of known and unknown function (**Figure 5B**). This results from broad score distributions that
384 change marginally across bins of increasing KFSC. SAPH-ire FPx tends to have lower average
385 scores that are compressed for the unknown function category and don't increase dramatically
386 until reaching KFSC >2 true positive status. PFS exhibits higher overall scores compared to
387 SAPH-ire FPx but shows comparable responsiveness to increasing KFSC. The score distribution
388 of the two models as shown by inter-quartile range also increases by almost 2-fold with increasing

389 KFSC, which is counter to the expectation for increased confidence in classification (**Figure 5C**).
390 Consequently, the recommendation thresholds used for each model must be low to enable either
391 model to capture even a small percentage of true positive phosphosites. A separate analysis
392 comparing only PFS and SAPH-ire TFx, which includes a larger phosphosite overlap (49,935
393 phosphosites), showed similar results (**Figure S3**).

394 In contrast to SAPH-ire FPx and PFS, the score distributions for SAPH-ire TFx become
395 less, rather than more broad with increasing KFSC, concomitant with the expectation for greater
396 confidence with increasing score (**Figure 5B,C**). ROC and recall curves for all three models show
397 that this difference is largely due to improved recall performance of SAPH-ire TFx, while ROC
398 AUC is otherwise similar between the three different models (**Figure 5D**). The practical
399 consequences of the differences between SAPH-ire TFx and other models is evident in terms of
400 number of missed calls based on recommended thresholds, where as many as 32% of highly
401 confident true positive functional PTMs ($KFSC_{MAP} > 2$) are mis-called by previous models – a
402 quantity that is lowered to less than 6% in SAPH-ire TFx (**Figure 5E**). This also results in an
403 increase in the number of phosphosites recommended as functionally impactful at all thresholds
404 (see Conclusions).

405 We next compared all three models to a recently published fourth model that is not based
406 on machine learning, but rather on a derivative of sequence homology modeling (37). In brief, this
407 method calls conserved phosphorylation hotspots that are defined based on the sequence
408 conservation of protein regions in domain families that are densely populated with observed
409 phosphorylation sites. In general, high scores were enriched for conserved phosphosite hotspots
410 regardless of model, with SAPH-re TFx exhibiting the best overall performance in terms of recall
411 and score distribution across KFSC (**Figure S3**). Taken together, these data strongly support the
412 conclusion that SAPH-ire TFx is a robust and effective model for the classification of PTM function,
413 which surpasses the performance of previous models.

414

415 **SAPH-ire web interface**

416 To support the community with easy access to SAPH-ire TFx, we have launched a web
417 resource (https://saphire.biosci.gatech.edu). This resource provides an intuitive and dynamic
418 interface for interacting with SAPH-ire predictions and allows for quick prioritization of PTMs of
419 interests in context of either a protein or family of interest. In the results window for the tool, one
420 can quickly retrieve and analyze the SAPH-ire scores and general statistics for the associated
421 PTMs, as well as quickly identify known functional PTMs in the family (**Figure 6**). Additionally,
422 SAPH-ire predictions are made available via API.

423
424 **Figure 6. SAPH-ire TFx web tool.** Screen shots from https://saphire.biosci.gatech.edu showing
425 architecture of the SAPH-ire TFx web tool. Key features include dynamic customizable graphic
426 and tabular display that can be filter-updated by the user and downloaded to produce figures.
427 Tabular data is searchable and is color coordinated to indicate PTMs aligned with or
428 correspondent with known functional data, including PubMed reference IDs, hyperlinks to protein
429 (UniProt) or protein family (InterPro) information, among other features. Additionally, information
430 is machine accessible through a corresponding API provided at the same domain.
431

432 **DISCUSSION**

433 Using systematic architecture and training optimization procedures we have created a new
434 machine learning model and complimentary web-accessible tool – SAPH-ire TFx – that is capable
435 of confidently recommending PTMs of likely functional significance. To our knowledge, this is the
436 first recommendation tool that does not rely solely on probability-based ranking for functional
437 prioritization. As estimated by ROC and recall analyses, the model is shown to be highly predictive
438 for recall of PTMs of known function, and this property is enhanced at increasing score thresholds
439 provided by the model. After its development, we tested the model with an expanded dataset to
440 which it had never been previously exposed, showing that its performance characteristics are
441 robust. To estimate its performance with physiologically relevant predictions, we demonstrated
442 that the model enriches PTMs that are coincident with pathogenic point mutations significantly
443 over those coinciding with benign mutations – an expected trend for models that predict the
444 functional importance of protein residues. In a series of benchmarking tests, we further showed

445 that SAPH-ire TFx outperforms existing machine learning or conservation-based hotspot models
446 (including one of our previous models) in all respects, including ROC, recall, and prediction
447 confidence. Finally, we provide quantitatively-validated thresholds that maximize confidence,
448 recall, and recommendations of unknown function PTMs (the ultimate goal of the model). All
449 model output is provided in a user-accessible web interface that is intended to maximize
450 knowledge about any given PTM in the context of all PTMs for a protein and/or protein family.

451 An important consideration in the design and interpretation of any model intended to
452 predict whether a PTM is functionally significant is the absence of true negative data – i.e. PTMs
453 for which definitive evidence demonstrates that they have no biological impact. This inevitable
454 characteristic of PTM data prohibits the estimation of accuracy as an evaluation metric and forces
455 reliance on true positive data alone as an evaluation metric. Our current perspective is that models
456 failing to efficiently capture true positive data fail to recognize the combinations of protein features
457 underlying the breadth of PTM functions that are known to exist in the eukaryotic proteome.
458 SAPH-ire TFx is the first model for the prediction of functional PTMs that maximizes the recall of
459 true positive data and so hopefully minimizes this concern as much as possible. As we have
460 shown here, models that do not optimize recall suffer from broad score distributions that force
461 extremely conservative thresholding to capture a reasonable (albeit small) fraction of true positive
462 data (**Figure 5E**). The accumulation of additional experimental data over time will certainly
463 highlight the robustness of these and future models.

464

465 **ACKNOWLEDGEMENTS**

466 We would like to thank Zahra Nassiri-Toosi, Wei Li, and other members of the Torres lab
467 for careful review and beta-testing of the SAPH-ire website. Special thanks also to Jiani Long and
468 Ragy Haddad for pilot work on SAPH-ire and web-tool development. Special thanks to Dr. Peng
469 Qiu of Georgia Institute of Technology for critical review of the manuscript and contributions to
470 model design. This work was funded by National Institutes of Health R01 GM117400 to M.T.

REFERENCES

1. Prabakaran S, Lippens G, Steen H, Gunawardena J. Post-translational modification: Nature's escape from genetic imprisonment and the basis for dynamic information encoding [Internet]. Vol. 4, Wiley Interdisciplinary Reviews: Systems Biology and Medicine. 2012 [cited 2012 Nov 5]. p. 565–83. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22899623

2. Nussinov R, Tsai C-J, Xin F, Radivojac P. Allosteric post-translational modification codes. Trends Biochem Sci [Internet]. 2012 Oct [cited 2012 Oct 9];37(10):447–55. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22884395

3. Lothrop AP, Torres MP, Fuchs SM. Deciphering post-translational modification codes. FEBS Lett [Internet]. 2013 Apr 17 [cited 2014 Feb 22];587(8):1247–57. Available from: http://www.sciencedirect.com/science/article/pii/S0014579313000926

4. Csizmok V, Forman-Kay JD. Complex regulatory mechanisms mediated by the interplay of multiple post-translational modifications. Curr Opin Struct Biol [Internet]. 2018 Feb 1 [cited 2019 Aug 6];48:58–67. Available from: https://www.sciencedirect.com/science/article/abs/pii/S0959440X17301306

5. Gibson DG, Glass JI, Lartigue C, Noskov VN, Chuang R-Y, Algire M a, et al. Creation of a bacterial cell controlled by a chemically synthesized genome. Science [Internet]. 2010 Jul 2;329(5987):52–6. Available from: http://www.ncbi.nlm.nih.gov/pubmed/20488990

6. Reimand J, Bader GD. Systematic analysis of somatic mutations in phosphorylation signaling predicts novel cancer drivers. Mol Syst Biol [Internet]. 2014;9(1):637–637. Available from: http://msb.embopress.org/cgi/doi/10.1038/msb.2012.68

7. Reimand J, Wagih O, Bader GD. Evolutionary constraint and disease associations of post-translational modification sites in human genomes. PLoS Genet [Internet]. 2015;11(1):e1004919. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25611800

8. Walsh CT, Garneau-Tsodikova S, Gatto GJ. Protein posttranslational modifications: The chemistry of proteome diversifications. Vol. 44, Angewandte Chemie - International Edition. 2005. p. 7342–72.

9. Darling AL, Uversky VN. Intrinsic disorder and posttranslational modifications: The darker side of the biological dark matter. Front Genet. 2018;9(MAY):1–18.

10. Chen C, Huang H, Wu CH. Protein Bioinformatics Databases and Resources. Methods Mol Biol [Internet]. 2017;1558(302):3–39. Available from: https://linkinghub.elsevier.com/retrieve/pii/B9780444519801500124

11. Pascovici D, Wu JX, McKay MJ, Joseph C, Noor Z, Kamath K, et al. Clinically Relevant Post-Translational Modification Analyses-Maturing Workflows and Bioinformatics Tools. Int J Mol Sci [Internet]. 2018 Dec 20;20(1). Available from: http://www.ncbi.nlm.nih.gov/pubmed/30577541

12. Doll S, Burlingame AL. Mass spectrometry-based detection and assignment of protein posttranslational modifications. ACS Chem Biol [Internet]. 2015 Jan 16;10(1):63–71. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25541750

13. Vlastaridis P, Kyriakidou P, Chaliotis A, Van de Peer Y, Oliver SG, Amoutzias GD. Estimating the total number of phosphoproteins and phosphorylation sites in eukaryotic

514 proteomes. Gigascience. 2017;6(2):1–11.

515 14. Landry CR, Levy ED, Michnick SW. Weak functional constraints on phosphoproteomes.
516 Trends Genet [Internet]. 2009 May;25(5):193–7. Available from:
517 http://www.ncbi.nlm.nih.gov/pubmed/19349092

518 15. Beltrao P, Albanèse V, Kenner LR, Swaney DL, Burlingame A, Villén J, et al. Systematic
519 functional prioritization of protein posttranslational modifications. Cell [Internet]. 2012 Jul
520 20 [cited 2013 May 21];150(2):413–25. Available from:
521 http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3404735&tool=pmcentrez&ren
522 dertype=abstract

523 16. Strumillo MJ, Oplová M, Viéitez C, Ochoa D, Shahraz M, Busby BP, et al. Conserved
524 phosphorylation hotspots in eukaryotic protein domain families. Nat Commun [Internet].
525 2019;10(1):1977. Available from: http://dx.doi.org/10.1038/s41467-019-09952-x

526 17. Minguez P, Letunic I, Parca L, Bork P. PTMcode: a database of known and predicted
527 functional associations between post-translational modifications in proteins. Nucleic Acids
528 Res [Internet]. 2013 Jan [cited 2014 Jan 13];41(Database issue):D306-11. Available from:
529 http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3531129&tool=pmcentrez&ren
530 dertype=abstract

531 18. Minguez P, Letunic I, Parca L, Garcia-Alonso L, Dopazo J, Huerta-Cepas J, et al.
532 PTMcode v2: a resource for functional associations of post-translational modifications
533 within and between proteins. Nucleic Acids Res [Internet]. 2015 Jan;43(Database
534 issue):D494--502. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25361965

535 19. Minguez P, Parca L, Diella F, Mende DR, Kumar R, Helmer-Citterich M, et al.
536 Deciphering a global network of functionally associated post-translational modifications.
537 Mol Syst Biol [Internet]. 2012 Jan [cited 2012 Oct 25];8(599):599. Available from:
538 http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3421446&tool=pmcentrez&ren
539 dertype=abstract

540 20. Dewhurst HM, Choudhury S, Torres MP. Structural Analysis of PTM Hotspots (SAPH-
541 ire)--A Quantitative Informatics Method Enabling the Discovery of Novel Regulatory
542 Elements in Protein Families. Mol Cell Proteomics [Internet]. 2015 Aug;14(8):2285–97.
543 Available from: http://www.mcponline.org/lookup/doi/10.1074/mcp.M115.051177

544 21. Johnson JR, Santos SD, Johnson T, Pieper U, Strumillo M, Wagih O, et al. Prediction of
545 Functionally Important Phospho-Regulatory Events in Xenopus laevis Oocytes. PLOS
546 Comput Biol [Internet]. 2015;11(8):e1004362. Available from:
547 http://dx.plos.org/10.1371/journal.pcbi.1004362

548 22. Torres MP, Dewhurst H, Sundararaman N. Proteome-wide Structural Analysis of PTM
549 Hotspots Reveals Regulatory Elements Predicted to Impact Biological Function and
550 Disease. Mol Cell Proteomics [Internet]. 2016 Nov 3 [cited 2016 Oct 11];15(11):3513–28.
551 Available from: http://www.ncbi.nlm.nih.gov/pubmed/27697855

552 23. Dewhurst HM, Torres MP. Systematic analysis of non-structural protein features for the
553 prediction of PTM function potential by artificial neural networks. Liu B, editor. PLoS One
554 [Internet]. 2017 Feb 22;12(2):e0172572. Available from:
555 http://dx.plos.org/10.1371/journal.pone.0172572

556 24. Xiao Q, Miao B, Bi J, Wang Z, Li Y. Prioritizing functional phosphorylation sites based on
557 multiple feature integration. Sci Rep [Internet]. 2016;6(April):24735. Available from:

http://www.ncbi.nlm.nih.gov/pubmed/27090940%5Cnhttp://www.ncbi.nlm.nih.gov/pubmed/27090940

25. Hornbeck P V, Zhang B, Murray B, Kornhauser JM, Latham V, Skrzypek E. PhosphoSitePlus, 2014: mutations, PTMs and recalibrations. Nucleic Acids Res [Internet]. 2015;43(Database issue):D512-20. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4383998&tool=pmcentrez&rendertype=abstract

26. Li J, Jia J, Li H, Yu J, Sun H, He Y, et al. SysPTM 2.0: an updated systematic resource for post-translational modification. Database (Oxford) [Internet]. 2014;2014:bau025. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3975108&tool=pmcentrez&rendertype=abstract

27. Huang K-Y, Su M-G, Kao H-J, Hsieh Y-C, Jhong J-H, Cheng K-H, et al. dbPTM 2016: 10-year anniversary of a resource for post-translational modification of proteins. Nucleic Acids Res [Internet]. 2016 Jan;44(D1):D435--D446. Available from: http://www.ncbi.nlm.nih.gov/pubmed/26578568

28. Altschul SFF, Gish W, Miller W, Myers EWW, Lipman DJJ. Basic local alignment search tool. J Mol Biol [Internet]. 1990 Oct 5;215(3):403–10. Available from: http://www.ncbi.nlm.nih.gov/pubmed/2231712

29. Mitchell A, Chang H, Daugherty L, Fraser M, Hunter S, Lopez R, et al. The InterPro protein families database : the classification resource after 15 years. Nucleic Acids Res [Internet]. 2015 Jan 28;43(D1):D213–21. Available from: http://nar.oxfordjournals.org/lookup/doi/10.1093/nar/gku1243

30. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res [Internet]. 2004;32(5):1792–7. Available from: http://www.ncbi.nlm.nih.gov/pubmed/15034147

31. Abadi M, Barham P, Chen J, Chen Z, Davis A, Dean J, et al. TensorFlow: A system for large-scale machine learning [Internet]. 2016 [cited 2018 Sep 29]. Available from: https://ai.google/research/pubs/pub45381

32. McKinney W. Data Structures for Statistical Computing in Python [Internet]. 2010. p. 51–6. Available from: http://conference.scipy.org/proceedings/scipy2010/mckinney.html

33. Dupret G, Koda M. Bootstrap re-sampling for unbalanced data in supervised learning. Eur J Oper Res [Internet]. 2001 Oct;134(1):141–56. Available from: www.elsevier.com/locate/dsw

34. Landrum MJ, Lee JM, Benson M, Brown GR, Chao C, Chitipiralla S, et al. ClinVar: improving access to variant interpretations and supporting evidence. Nucleic Acids Res [Internet]. 2018 Jan;46(D1):D1062--D1067. Available from: http://www.ncbi.nlm.nih.gov/pubmed/29165669

35. R Core Team. R: A language and environment for statistical computing [Internet]. R Foundation for Statistical Computing, Vienna, Austria.; 2013. Available from: http://www.r-project.org/

36. Ochoa D, Jarnuczak AF, Gehre M, Soucheray M, Kleefeldt AA, Vieitez C, et al. The functional landscape of the human phosphoproteome. bioRxiv [Internet]. 2019

601 Feb;541656. Available from: https://www.biorxiv.org/content/10.1101/541656v1

602 37. Strumillo MJ, Oplova M, Vieitez C, Ochoa D, Shahraz M, Busby BP, et al. Conserved
603 phosphorylation hotspots in eukaryotic protein domain families. bioRxiv [Internet].
604 2018;391185. Available from: https://www.biorxiv.org/content/early/2018/08/13/391185

605 **SUPPLEMENTAL TABLES AND FIGURES**
606
607 **Table S1. Frequency of PTM types analyzed by SAPH-ire TFx.**
608
609 **Table S2. List of InterPro families analyzed in SAPH-ire TFx.**
610
611 **Table S3. List of organisms represented by SAPH-ire TFx.**
612
613 **Table S4. Description of features used in the SAPH-ire TFx model.**
614
615 **Figure S1. ROC curves at different KFSC thresholds.** (unfurled in Figure 4).
616
617 **Figure S2. Pairwise comparison between SAPH-ire TFx and the PFS machine learning**
618 **models.** These data include 49,935 phosphosites that overlap between the SAPH-ire TFx and
619 PFS datasets. (A) Score distribution relative to functional status. (B) Inter-quartile ranges from the
620 distributions in A. (C). ROC and recall curves for the score comparison of each model.
621
622 **Figure S3. Comparison of SAPH-ire FPx, PFS, and SAPH-ire TFx models relative to**
623 **phosphosite conservation hotspot analysis.** Phosphosites localized within conserve
624 phosphosite hotspots predicted by Strumillo et al. were used to bin data from the three-model
625 comparison shown in figure 5. (A) Score distribution relative to functional status relative to
626 predicted hotspots. (B) Inter-quartile ranges from the distributions in A.

FIGURE 1

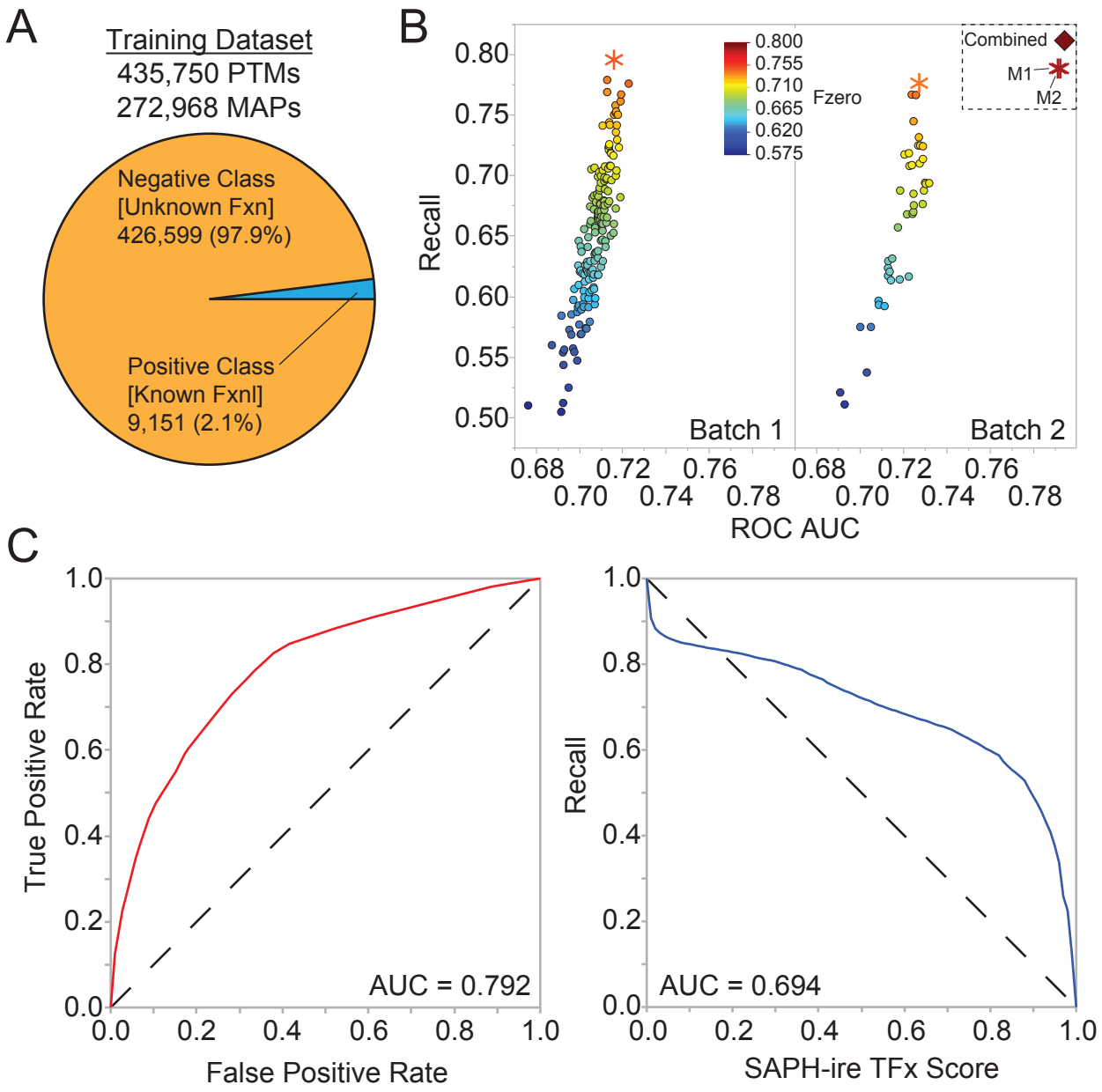

A
Training Dataset
435,750 PTMs
272,968 MAPs
Negative Class
[Unknown Fxn]
426,599 (97.9%)
Positive Class
[Known Fxnl]
9,151 (2.1%)
B
Recall
ROC AUC
Fzero
0.800
0.755
0.710
0.665
0.620
0.575
Combined
M1
M2
Batch 1
Batch 2
C
True Positive Rate
False Positive Rate
AUC = 0.792
Recall
SAPH-ire TFx Score
AUC = 0.694

FIGURE 2

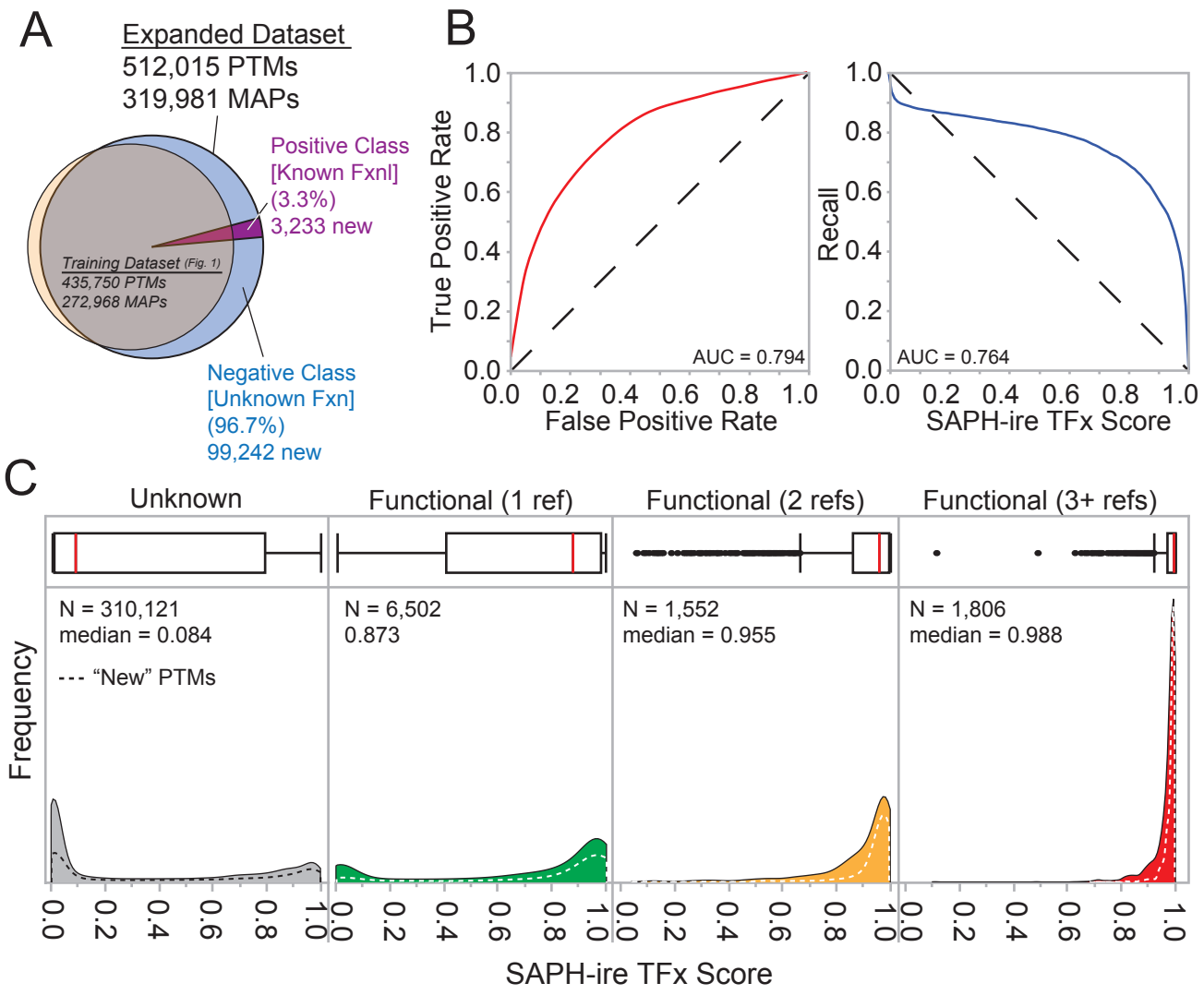
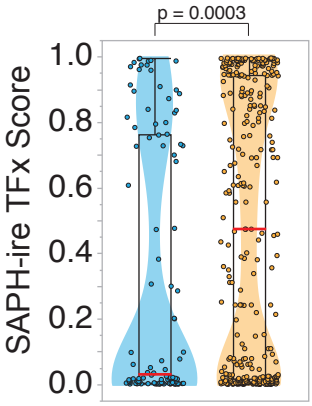

A
Expanded Dataset
512,015 PTMs
319,981 MAPs
Positive Class
[Known Fxnl]
(3.3%)
3,233 new
Training Dataset (Fig. 1)
435,750 PTMs
272,968 MAPs
Negative Class
[Unknown Fxn]
(96.7%)
99,242 new
B
True Positive Rate
False Positive Rate
AUC = 0.794
Recall
SAPH-ire TFx Score
AUC = 0.764
C
Unknown
Functional (1 ref)
Functional (2 refs)
Functional (3+ refs)
N = 310,121
median = 0.084
--- "New" PTMs
N = 6,502
0.873
N = 1,552
median = 0.955
N = 1,806
median = 0.988
Frequency
SAPH-ire TFx Score

FIGURE 3



	Benign	Pathogenic
N	102	322
median	0.033	0.475

FIGURE 4

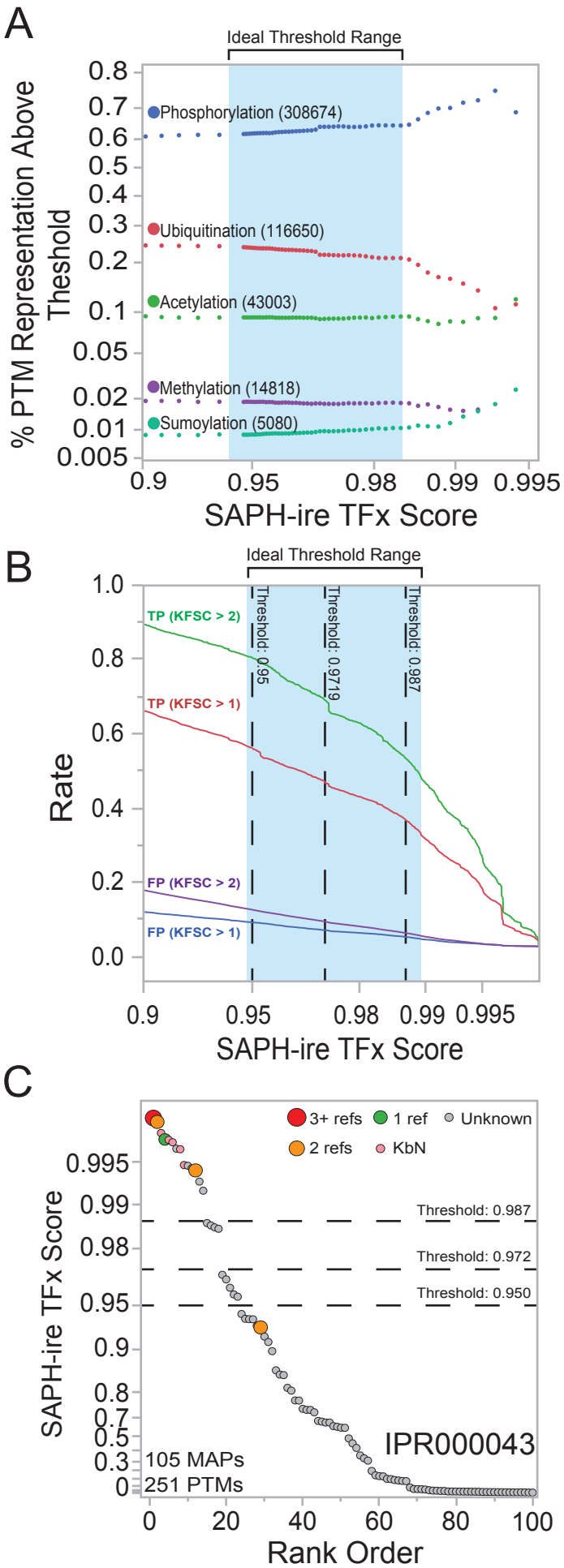
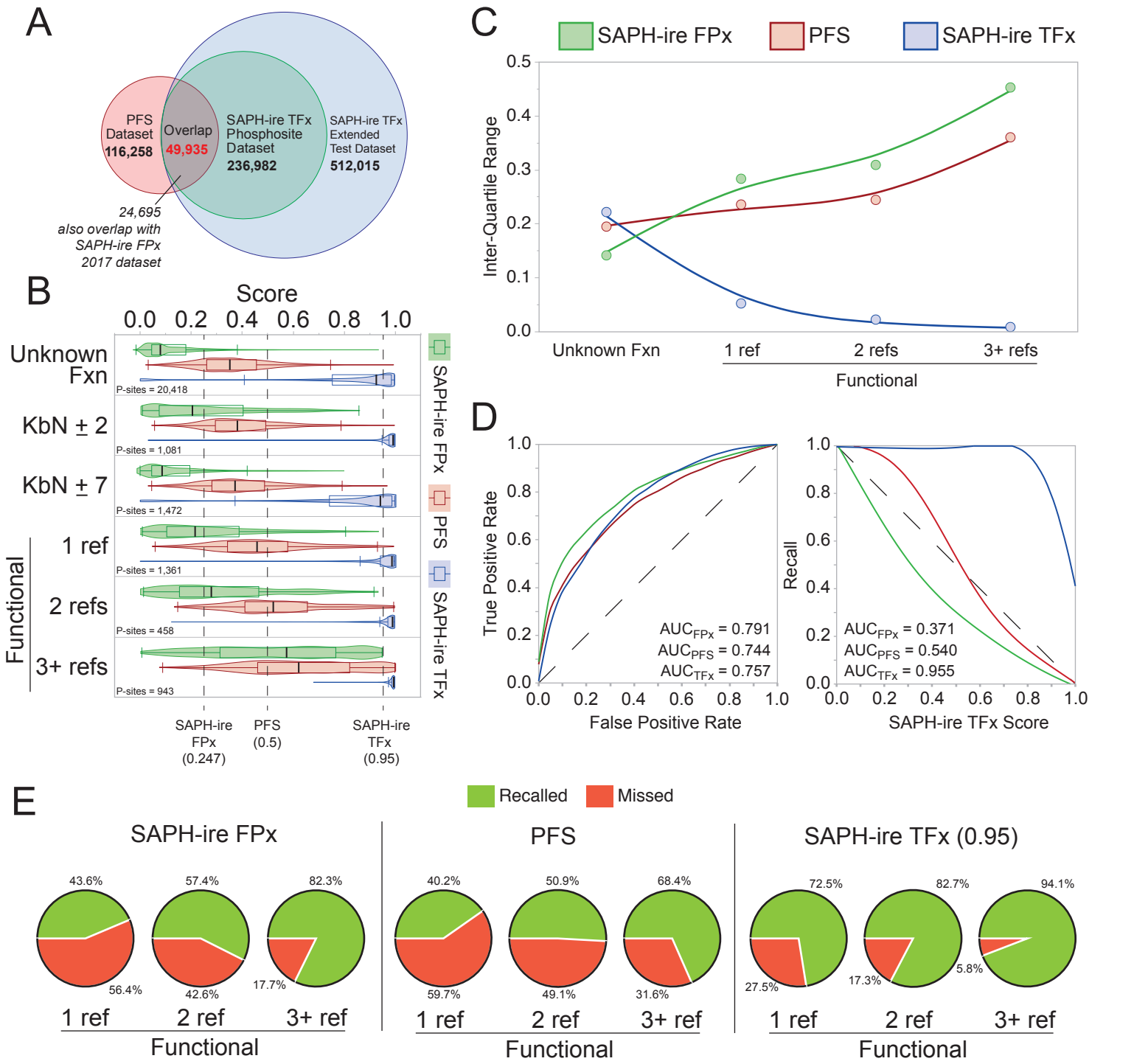

A
Ideal Threshold Range
% PTM Representation Above Theshold
Phosphorylation (308674)
Ubiquitination (116650)
Acetylation (43003)
Methylation (14818)
Sumoylation (5080)
SAPH-ire TFx Score
B
Ideal Threshold Range
Threshold: 0.95
Threshold: 0.9719
Threshold: 0.987
TP (KFSC > 2)
TP (KFSC > 1)
FP (KFSC > 2)
FP (KFSC > 1)
Rate
SAPH-ire TFx Score
C
3+ refs
1 ref
Unknown
2 refs
KbN
Threshold: 0.987
Threshold: 0.972
Threshold: 0.950
IPR000043
105 MAPs
251 PTMs
SAPH-ire TFx Score
Rank Order

FIGURE 5



Recommendations

SAPH-ire FPx	**PFS**	**SAPH-ire TFx**
21,708/151,764 (14.3%)	11,056/116,258 (9.5%)	*Thresh 0.95*: 133,719/512,015 (26.1%)
		Thresh 0.972: 101,632/512,015 (19.8%)
		Thresh 0.987: 63,827/512,015 (12.5%)

FIGURE 6

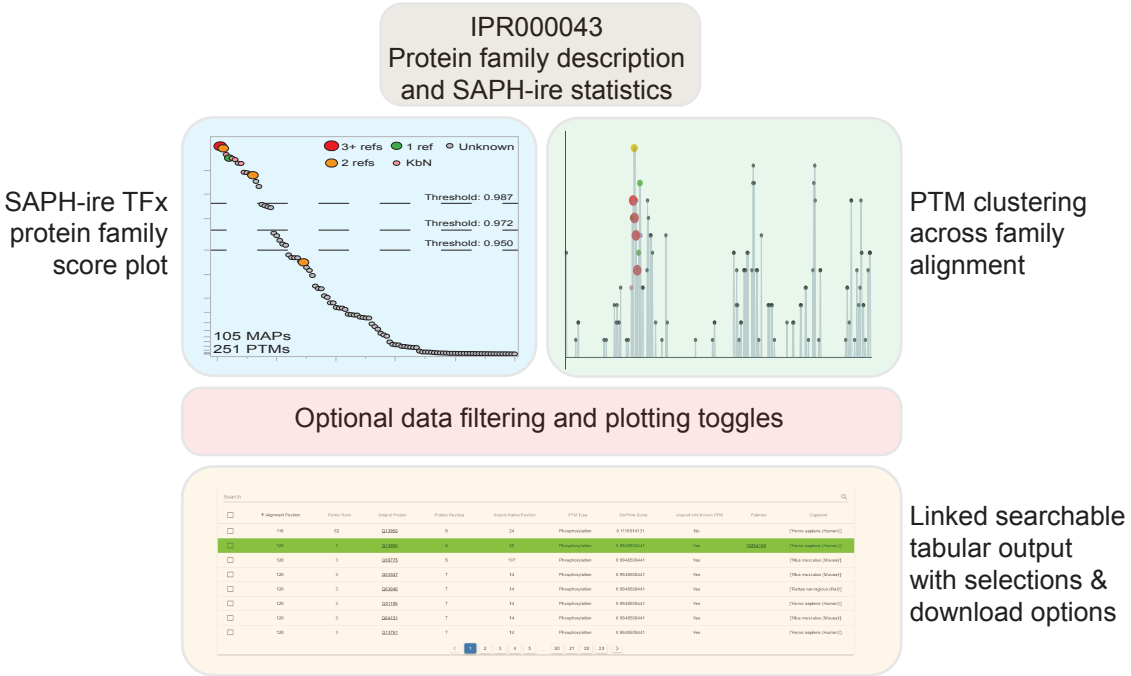

IPR000043
Protein family description
and SAPH-ire statistics
SAPH-ire TFx
protein family
score plot
3+ refs
1 ref
Unknown
2 refs
KbN
Threshold: 0.987
Threshold: 0.972
Threshold: 0.950
105 MAPs
251 PTMs
PTM clustering
across family
alignment
Optional data filtering and plotting toggles
Linked searchable
tabular output
with selections &
download options

FIGURE S1

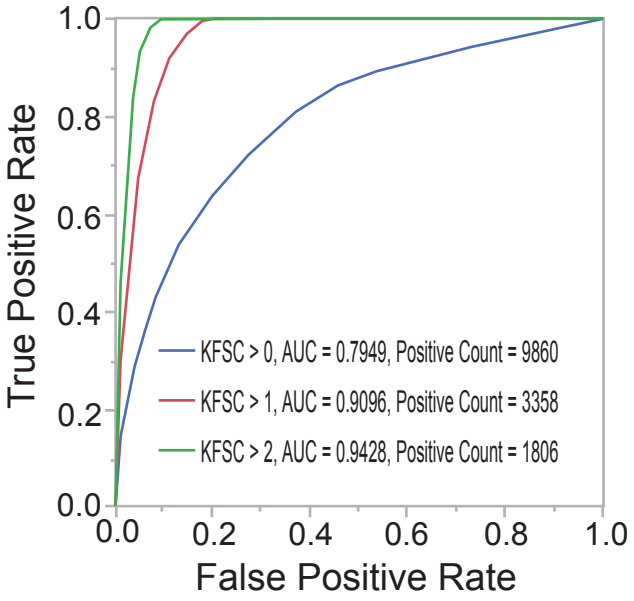

1.0
0.8
0.6
0.4
0.2
0.0
True Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
False Positive Rate
KFSC > 0, AUC = 0.7949, Positive Count = 9860
KFSC > 1, AUC = 0.9096, Positive Count = 3358
KFSC > 2, AUC = 0.9428, Positive Count = 1806

FIGURE S2

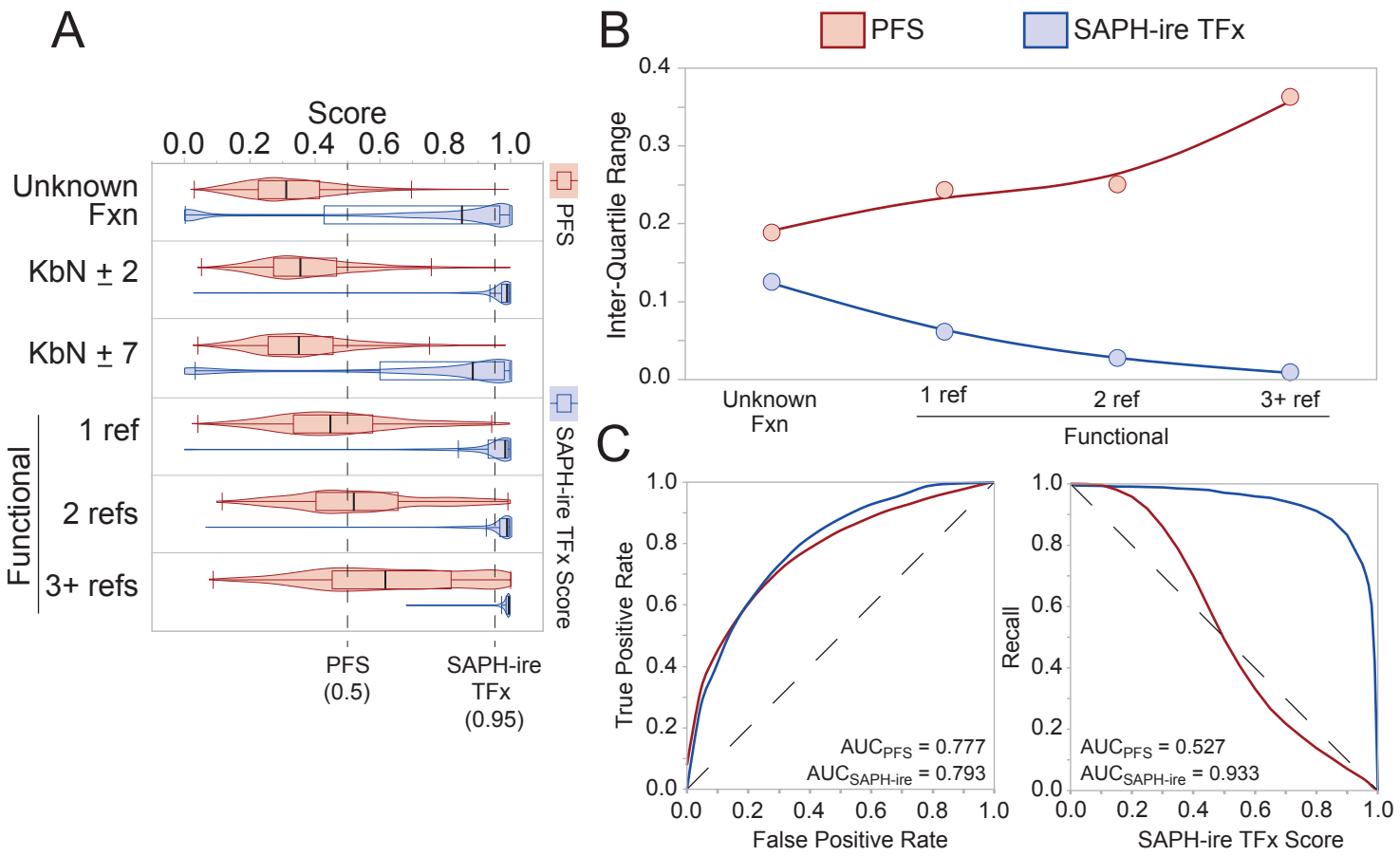
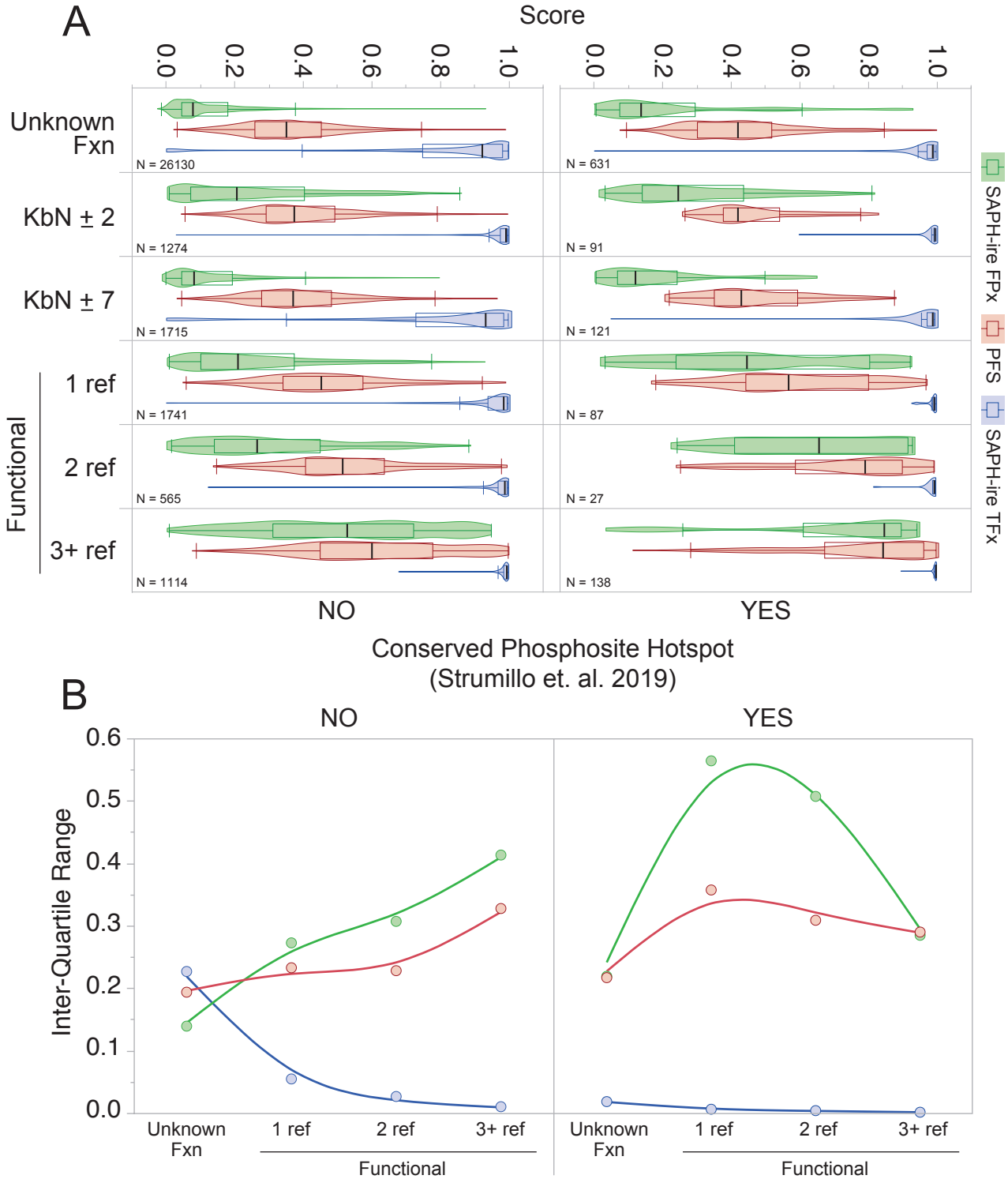

A
Score
0.0 0.2 0.4 0.6 0.8 1.0
Unknown Fxn
KbN ± 2
KbN ± 7
1 ref
2 refs
3+ refs
Functional
PFS
SAPH-ire TFx Score
PFS (0.5)
SAPH-ire TFx (0.95)
B
PFS
SAPH-ire TFx
Inter-Quartile Range
0.0 0.1 0.2 0.3 0.4
Unknown Fxn
1 ref
2 ref
3+ ref
Functional
C
True Positive Rate
False Positive Rate
AUC_PFS = 0.777
AUC_SAPH-ire = 0.793
Recall
SAPH-ire TFx Score
AUC_PFS = 0.527
AUC_SAPH-ire = 0.933

FIGURE S3


A
Score
0.0
0.2
0.4
0.6
0.8
1.0
Unknown Fxn
KbN ± 2
KbN ± 7
Functional
1 ref
2 ref
3+ ref
N = 26130
N = 1274
N = 1715
N = 1741
N = 565
N = 1114
N = 631
N = 91
N = 121
N = 87
N = 27
N = 138
SAPH-ire FPx
PFS
SAPH-ire TFx
NO
YES
Conserved Phosphosite Hotspot
(Strumillo et. al. 2019)
B
NO
YES
Inter-Quartile Range
0.0
0.1
0.2
0.3
0.4
0.5
0.6
Unknown Fxn
1 ref
2 ref
3+ ref
Functional