GenPTM: A Generalizable Framework for Protein Post-Translational Modification Information Extraction from the Scientific Literature

Shovan Bhowmik[1*], Karen Ross[2], Chuming Chen[1,3], Cathy Wu[1,3], K. Vijay-Shanker[1]

[1*]Department of Computer and Information Sciences, University of Delaware, Newark, Delaware, United States.
[2]Department of Biochemistry and Molecular & Cellular Biology, Georgetown University Medical Center, Washington, D.C., United States.
[3]Center for Bioinformatics and Computational Biology, University of Delaware, Newark, Delaware, United States.

*Corresponding author(s). E-mail(s): shovan@udel.edu;
Contributing authors: ker25@georgetown.edu; chenc@udel.edu; wuc@udel.edu; vijay@udel.edu;

Supplementary Information

S1 Comprehensive PTM Mapping Table

To construct the PTM mapping table, we selected a diverse list of post-translational modifications and their common lexical patterns in the literature. This enabled us to incorporate a rich set of mapping items for both modification terms and associated chemical group terms across different PTM types.

Table S1: Comprehensive PTM mapping table

PTM Class	Modification Mentions	Chemical or Group Mentions	Residues Modified
Phosphorylation	phosphorylated, phosphorylation, phospho modification	phosphate, PO4, pSer, pThr, pTyr	Ser, Thr, Tyr
Serine phosphorylation	serine phosphorylation, pSer	phosphate	Ser
Threonine phosphorylation	threonine phosphorylation, pThr	phosphate	Thr
Tyrosine phosphorylation	tyrosine phosphorylation, pTyr	phosphate	Tyr
Histidine phosphorylation	histidine phosphorylation, pHis	phosphate	His
Acetylation	acetylated, acetylation	acetyl group, acetyl CoA	Lys, N terminus
Lysine acetylation	lysine acetylation, acetyl Lys	acetyl group	Lys
N terminal acetylation	N terminally acetylated	acetyl group	N terminal residues
Methylation	methylated, methylation	methyl group	Lys, Arg
Lysine methylation	Kme, mono methyl Lys	methyl group	Lys
Arginine methylation	Rme, arginine methylation	methyl group	Arg
Ubiquitination	ubiquitinated, ubiquitination	ubiquitin, Ub	Lys
Mono ubiquitination	mono ubiquitinated	ubiquitin	Lys
Poly ubiquitination	polyubiquitinated	ubiquitin chain	Lys
Sumoylation	sumoylated, SUMOylated	SUMO1, SUMO2, SUMO3	Lys
NEDDylation	NEDDylated	NEDD8	Lys
ISGylation	ISGylated	ISG15	Lys
UFMylation	UFMylated, UFM1 conjugated	UFM1 group, UFM1 modifier	Lys
N linked glycosylation	N glycosylated	GlcNAc, N glycan	Asn
O linked glycosylation	O glycosylated	GalNAc, O glycan, T antigen	Ser, Thr
O GlcNAcylation	O GlcNAcylated	O GlcNAc	Ser, Thr
C mannosylation	C mannosylated	mannose, Man	Trp
Glycation	glycated	reducing sugar, AGE	Lys, Arg
Palmitoylation	palmitoylated	palmitate, palmitoyl group	Cys
Myristoylation	myristoylated	myristate	N terminal Gly
Prenylation	prenylated	prenyl group	Cys
Farnesylation	farnesylated	farnesyl group	Cys
Geranylgeranylation	geranylgeranylated	geranylgeranyl group	Cys
S nitrosylation	S nitrosylated	S nitroso, SNO	Cys
S glutathionylation	S glutathionylated	GSH, glutathione	Cys

PTM Class	Modification Mentions	Chemical or Group Mentions	Residues Modified
Disulfide formation	disulfide linked	disulfide bond	Cys
Sulfation	sulfated	sulfate, SO3	Tyr
Nitration	nitrated	nitro group, NO2	Tyr
Hydroxylation	hydroxylated	hydroxyl group	Pro, Lys, Asn
Proline hydroxylation	hydroxyproline	hydroxyl group	Pro
Lysine hydroxylation	hydroxylysine	hydroxyl group	Lys
ADP ribosylation	ADP ribosylated	ADP ribose	Ser, Glu, Asp, Lys, Arg
Mono ADP ribosylation	MARylated	MAR group	Lys, Arg
Poly ADP ribosylation	PARylated	PAR chain	Ser
Succinylation	succinylated	succinyl group	Lys
Malonylation	malonylated	malonyl group	Lys
Glutarylation	glutarylated	glutaryl group	Lys
Crotonylation	crotonylated	crotonyl group	Lys
Propionylation	propionylated	propionyl group	Lys
Butyrylation	butyrylated	butyryl group	Lys
Lactylation	lactylated	lactyl group	Lys
Beta hydroxybutyrylation	beta hydroxybutyrylated	beta hydroxybutyryl group	Lys
Citrullination	citrullinated, deiminated	citrulline group	Arg
Deamidation	deamidated	deamidated Asn or Gln	Asn, Gln
Carbonylation	carbonylated	carbonyl group	Lys, Arg, Pro, Thr
Oxidation	oxidized Met	Met sulfoxide, Met sulfone	Met
Pyroglutamylation	pyroglutamylated	pyroglutamate	N terminal Gln or Glu
GPI anchoring	GPI anchored	GPI anchor	C terminus
Biotinylation	biotinylated	biotin	Lys
Lipoylation	lipoylated	lipoic acid	Lys
Polyglutamylation	polyglutamylated	poly Glu chain	Glu
Polyglycylation	polyglycylated	poly Gly chain	Gly
Pupylation	pupylated	Pup protein	Lys
AMPylation	AMPylated	AMP group	Tyr, Ser, Thr
GMPylation	GMPylated	GMP group	Tyr
UMPylation	UMPylated	UMP group	Tyr
Phosphocholination	phosphocholinated	phosphocholine	Ser, Thr
Phosphoglycerylation	phosphoglycerylated	3 phosphoglycerate	Lys
Transglutamination	TGase crosslinked	Gln Lys isopeptide	Gln, Lys
Carboxylation	carboxylated	gamma carboxyl group	Glu

S2 Site-to-Protein Association Rules

We designed several grammar-based rules to associate the protein with a modified site. Additionally, we also developed a few heuristics when grammar rules are not possible to capture the association relation.

A common active sentence structure can be represented as: "... modifies P at S". Here, P denotes the substrate protein and S the modified site. The dependency pattern typically follows:

Pattern 1:
VERB >dobj P
VERB >prep PREP (at|on|in)
PREP >pobj S

Here, "VERB" indicates the action word describing the modification (e.g., phosphorylates, acetylates), while "PREP" corresponds to the prepositional term (e.g., at, on) that introduces the site mention. The dependency labels "dobj", "prep", and "pobj" denote the grammatical relations in the dependency parse, where the substrate protein P functions as the direct object (dobj) of the action verb, and the site S is linked through a prepositional phrase (pobj) governed by "PREP". Together, these dependencies capture the syntactic pathway from the modifying action to the protein and its associated site. This pattern is captured by parsing the sentence structure mentioned above from SpaCy's dependency parse tree and illustrated in Fig. 1.

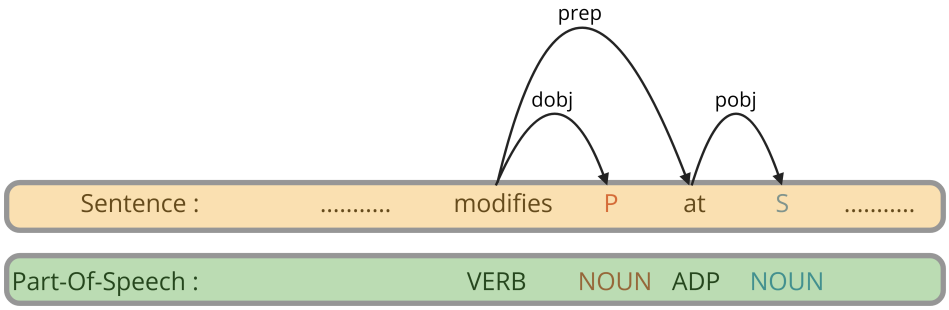

Fig. 1 Dependency Parse Tree of an Active Sentence Pattern.

The sentence fragment "... modifies P at S" illustrates a typical dependency configuration used to identify protein–site associations. In this structure, the main verb (modifies) establishes a direct object (dobj) dependency with the substrate protein (P) and a prepositional (prep) dependency with the preposition (at). The preposition subsequently governs the site mention (S) through a prepositional object (pobj) relation. The corresponding part-of-speech tags are displayed beneath the sentence to provide additional syntactic context.

We also encounter reverse patterns such as: "... modifies S of P". For this type, we design a separate rule depending on the dependency structure.

Similarly, for passive structures "P is modified at S" or "P that is modified at S", we design rules to associate site to their corresponding protein mention. Additionally, nominalized form of verbs are also frequently found in the biomedical text such as "... promotes modification of P at S". For such patterns, we use the following dependency representation:
Pattern 2:
VERB >dobj NOUN
NOUN >prep PREP(of)
PREP >pobj P
VERB >prep PREP(at|on|in)
PREP >pobj S

When we observe Pattern 2, we make an action: S >> site_of >> P.
Example 2: Mechanistically, we show that CPT1A promotes succinylation of MFF at lysine 302 (K302), which protects against Parkin-mediated ubiquitin-proteasomal degradation of MFF. [PMID: 37291333] [**?**]

In Example 2, "lysine 302" is recognized as the modification site of "MFF".

For the remaining grammar-based rules, the corresponding structural templates and grammar patterns are provided below.
Pattern 3:
Sentence Structure: P is modified at S
VERB >nsubjpass P
VERB >prep PREP(at|on|in|within)
PREP >pobj S
Action: S >> site_of >> P

Pattern 4:
Sentence Structure: P which is modified at S
P >(relcl|acl) VERB
VERB >prep PREP(at|on|in|within)
PREP >pobj S
Action: S >> site_of >> P

Pattern 5:
Sentence Structure: P is modified at S (For different parsing)
P >nsubj AUX
AUX >acomp ADJ
ADJ >prep PREP(at|in|on|within)
PREP >pobj S
Action: S >> site_of >> P

Pattern 6:
Sentence Structure: P undergoes modification at S
VERB >compound P
NOUN >compound VERB

NOUN >prep PREP(at|on|in|within)
PREP >pobj S
Action: S >> site_of >> P

Pattern 7:
Sentence Structure: identified S in P VERB >dobj S
VERB >prep PREP(at|on|in)
PREP >pobj P
Action: S >> site_of >> P

Pattern 8:
Sentence Structure: conjugated to P via S
VERB >prep PREP(to)
PREP >pobj P
VERB >prep ADP
ADP >pobj S
Action: S >> site_of >> P

Pattern 9:
Sentence Structure: Detection of P modified on S
NOUN >acl VERB
NOUN >prep PREP(of)
PREP >pobj P
VERB >prep PREP(at|in|on)
PREP >pobj S
Action: S >> site_of >> P

Pattern 10:
Sentence Structure: Modification of P occurs at S
VERB >nsubj NOUN
NOUN(*ends with 'tion' or 'ing'*) >prep PREP(of)
PREP >pobj P
VERB >prep PREP(at|in|on)
PREP >pobj S
Action: S >> site_of >> P

Pattern 11:
Sentence Structure: Modification of P at S
NOUN >prep PREP(of)
PREP >pobj P
NOUN >prep PREP(at|on|in)
PREP >pobj S
Action: S >> site_of >> P

Pattern 12:
Sentence Structure: Modification of S on P

NOUN >prep PREP(of)
PREP >pobj S
pobj >prep PREP(of|on|in|within)
PREP >pobj P
Action: S >> site_of >> P

Pattern 13:
Sentence Structure: S modification of P
NOUN(*ends with 'tion' or 'ing'*) >compound S
NOUN >prep PREP(of|in|on)
PREP >pobj P
Action: S >> site_of >> P

Pattern 14:
Sentence Structure: P modification at S
NOUN(*ends with 'tion' or 'ing'*) >compound P
NOUN >prep PREP(at|in|on)
PREP >pobj S
Action: S >> site_of >> P

Pattern 15:
Sentence Structure: S of P
S >prep PREP(of|on|in|within)
PREP >pobj P
Action: S >> site_of >> P

Pattern 16:
Sentence Structure: P at S
P >prep PREP(at|on)
PREP >pobj S
Action: S >> site_of >> P

Pattern 17:
Sentence Structure: P modified at S
VERB(*VBN*) >nsubj P
VERB >prep PREP(at|in|on|within)
PREP >pobj S
Action: S >> site_of >> P

Pattern 18:
Sentence Structure: inhibits P through something of S
VERB >dobj NOUN
NOUN >compound P
VERB >prep PREP(of) PREP >pobj S
Action: S >> site_of >> P

Pattern 19:
Sentence Structure: P has S
VERB (*VBZ*) >nsubj P
VERB >dobj S
Action: S >> site_of >> P

Pattern 20:
Sentence Structure: P S
P >compound S
Action: S >> site_of >> P

Pattern 21:
Sentence Structure: P (S)
P >appos S
Action: S >> site_of >> P

Pattern 22:
Sentence Structure: Action on S affects P activities
With this rule, we captured several actions such as mutation, modification, on a site mention that triggers protein activities. We investigated examples to find trigger words such as diminish, enhance, regulate, increase, decrease, and activity words such as activation, function with protein mention. If such a scenario appears, we associate the sites with the proteins.

Example 3: Reversible acetylation of Lys49 could thus regulate the repair activity of NEIL2 in vivo. [PMID: 15175427] [**?**]

In Example 3, the action word is "acetylation", the trigger is "regulate", and the activity word is "activity" itself. Therefore, our rule works for this sentence, and we assign NEIL2 (colored in red) as the associated substrate of "Lys49" (colored in blue).

Pattern 23:
Sentence Structure: Mapping of P reveals S
With this rule, we captured words such as mapping or analysis of a protein mention in a sentence. Our investigation on such cases revealed that if there is any mapping of a protein mention in the sentence, it corresponds to any site mentions present in the sentence. Therefore, we associate the mapped protein mention with the site for such cases.

Generic Site Mentions and "is_a" Propagation
In some cases, generic terms such as "site" or "residue(s)" are mentioned which follow the "site_of" pattern with the substrate protein, while the actual site mention appears elsewhere in the sentence. To capture such cases, we implement an "is_a" propagation rule that links the explicit site mention to the generic site term and the link between the generic site mention and the substrate protein becomes "part_of" relation. We note that part_of relations are conceptually similar to site_of relations, differing only in the intermediate association: when a site is directly connected to a protein, it forms

a site_of relation; when mediated through an intermediate structure (e.g., domain or region), it forms a part_of relation with the intermediate term.

One of the "is_a" pattern is "S is a major site..." where S is the explicit site mention and "site" is the generic term. The pattern for this representation in the dependency tree looks like the following:

Pattern:
AUX >nsubj *site*
AUX >attr S
Action: S >> is_a >> *site*

Example 4: From these studies, it was concluded that Ser-247 is the major site of phosphorylation in the IL-2 receptor and that Thr-250 is a minor site. [PMID: 2941417] [**?**]

Here, we first recognize "site of phosphorylation in IL-2" as a part_of relation between the generic term "site" and "IL-2". Through is_a propagation, we then link "Ser-247" to "site", which ultimately establishes "Ser-247" >> site_of >> IL-2.

The is_a propagation also covers appositive or compound dependencies (e.g., K302 >> is_a >> lysine 302 in Example 2). We design several grammar-based is_a rules by analyzing examples from the literature. We also want to mention here that the is_a propagation is applicable to both sites and proteins.

Conjunction Propagation

In many sentences, multiple sites are listed together using commas and conjunctions such as "and" or "or." In such cases, our grammar-based rules typically establish the site–protein association only for the first site in the sequence. However, because these sites are semantically linked through conjunctions, they all refer to the same substrate. Therefore, once an association is established for the first site, we propagate that association to the remaining conjunctively connected sites. This propagation applies not only to site mentions but also to substrate mentions when they appear in coordinated structures.

Example 5: BAP1 undergoes covalent modification by ubiquitin-fold modifier 1 (UFM1) at Lys51, Lys61, Lys187, and Lys205, enhancing its interaction with pVHL and promoting pVHL stabilization. [PMID: 40644555] [**?**]

In Example 5, our dependency-based rules correctly identify "Lys51" as a modification site associated with the substrate "BAP1." Because "Lys61," "Lys187," and "Lys205" are connected to "Lys51" through conjunctions, the conjunction-propagation module extends the association to these additional sites, ensuring that all listed residues are correctly linked to "BAP1".

Two part_of Relations

Additionally, we identify two part_of relations in sentences where the site is localized to a domain or region of the protein, expressed through phrases such as "in the domain of P," "within the region of P". In these cases, we first create a part_of link between the site and the intermediate region (e.g., domain, region), and another part_of link between that region and the protein P.

Example 6: We have identified S150 in the InsP(3)R2 core suppressor domain (amino acids 1-225) as the specific residue that is phosphorylated by CaMKII. [PMID: 23019322] [**?**]

Here, the system establish two relations, S150 >> part_of >> domain and domain >> part_of >> InsP(3)R2, thus concluding S150 >> site_of >> InsP(3)R2.

Cross-Sentence and Contextual Association
All dependency-based rules described above are implemented using "SpaCy". However, when the site and its substrate protein do not co-occur in the same sentence, we extend our analysis to broader textual contexts. First, we search other sentences within the same abstract to identify if the same site has an established substrate association. If found, we propagate that association. If not, we examine the title or the first two sentences of the abstract, which often contain introductory phrases such as "modification of P" to infer the substrate context, as our investigation reveals that when there is a modification of a protein in the title and modification of a site in the same abstract with no explicit evidence of the substrate of the site, then there is high possibility the the modified protein is the associated substrate for the mentioned site.

S3 Testing Dataset Statistics

Table S2. Testing dataset statistics for the GenPTM tool.

PTM Category	PTM Type	No. of Abstracts	No. of Modified Protein Mentions	No. of Modified Site Mentions
WS-PTMs	Ubiquitination	50	49	8
	Ubiq-Like	25	49	9
	Acetylation	55	41	24
	Phosphorylation	75	68	48
OF-PTMs	Palmitoylation	40	43	10
	ADP-Ribosylation	47	22	
	Succinylation	45	33	23
	Hydroxylation	43	31	10
	UFMylation	40	43	21
LF-PTMs	Lipoylation Citrullination AMPylation	68	38	30
Total		**488**	**417**	**183**

Modified Proteins Mentions: The total number of unique modified protein mentions in each PTM corpus for the given number of abstracts. **Modified Site Mentions:** The total number of unique modified site mentions in each PTM corpus for the given number of abstracts. Unmodified protein and site mentions in the abstracts are not part of the numbers reported here.

S4 Performance of the GenPTM tool (Abstract-level)

Table S3. PTM-wise performance evaluation at the abstract-level across PTM categories under different evaluation criteria.

PTM Category	PTM Type	⟨protein, site⟩ pair			protein-only			site-only		
		P	R	F1	P	R	F1	P	R	F1
WS-PTMs	Ubiquitination	100.00	87.50	93.33	94.23	100.00	97.03	88.89	100.00	94.12
	Ubiq-Like	100.00	88.89	94.12	84.91	91.84	88.24	100.00	88.89	94.12
	Acetylation	95.24	90.91	93.02	91.11	100.00	95.35	95.65	91.67	93.62
	Phosphorylation	100.00	89.47	94.44	95.65	97.06	96.35	97.92	97.92	97.92
OF-PTMs	Palmitoylation	80.00	100.00	88.89	93.48	100.00	96.63	100.00	100.00	100.00
	ADP-Ribosylation	–	–	–	79.17	86.36	82.61	–	–	–
	Succinylation	100.00	100.00	100.00	97.06	100.00	98.51	100.00	100.00	100.00
	Hydroxylation	100.00	100.00	100.00	92.86	85.00	88.76	100.00	100.00	100.00
	UFMylation	100.00	90.48	95.00	77.78	97.67	86.60	100.00	90.48	95.00
LF-PTMs	Lipoylation Citrullination AMPylation	70.37	82.61	76.00	87.50	92.11	89.74	81.08	100.00	89.55

S5 Performance of the GenPTM tool (Instance-level)

Table S4. Instance-level performance of GenPTM across different PTM categories.

	<protein, site> pair			protein-only			site-only		
PTM Category	P	R	F1	P	R	F1	P	R	F1
WS-PTMs	99.17%	93.75%	96.39%	91.18%	92.86%	92.01%	94.56%	96.53%	95.53%
OF-PTMs	96.59%	94.44%	95.51%	88.68%	92.31%	90.46%	100.00%	92.39%	96.05%
LF-PTMs	74.19%	82.14%	77.97%	90.65%	92.38%	91.51%	80.95%	91.89%	86.08%
All PTMs	95.00%	92.68%	93.83%	90.02%	92.57%	91.28%	94.16%	94.51%	94.33%

Table S5. PTM-wise performance evaluation at the instance-level across PTM categories under different evaluation criteria.

PTM Category	PTM Type	⟨protein, site⟩ pair			protein-only			site-only		
		P	R	F1	P	R	F1	P	R	F1
WS-PTMs	Ubiquitination	100.00	88.89	94.12	92.73	100.00	96.23	90.00	100.00	94.74
	Ubiq-Like	100.00	88.89	94.12	83.78	91.18	87.32	100.00	88.89	94.12
	Acetylation	97.73	95.56	96.63	91.60	93.75	92.66	94.87	94.87	94.87
	Phosphorylation	100.00	93.85	96.83	95.86	90.85	93.29	94.44	97.70	96.05
OF-PTMs	Palmitoylation	81.25	100.00	89.66	92.16	95.92	94.00	100.00	85.71	92.31
	ADP-Ribosylation				74.19	74.19	74.19			
	Succinylation	100.00	96.77	98.36	94.05	95.18	94.61	100.00	96.77	98.36
	Hydroxylation	100.00	100.00	100.00	90.57	88.89	89.72	100.00	100.00	100.00
	UFMylation	100.00	84.62	91.67	85.89	93.33	89.46	100.00	84.62	91.67
LF-PTMs	Lipoylation Citrullination AMPylation	74.19	82.14	77.97	90.65	92.38	91.51	80.95	91.89	86.08

The instance-level precision is very similar to the abstract-level precision, and in some cases even slightly higher. This happens because the same entity can appear multiple times within an abstract, which increases both true positives and false positives. This indicates that the system is consistent in identifying repeated mentions of the same entity.

However, the micro-average recall drops by up to 5% across different PTM categories, which leads to a decrease in the overall F1-score for both protein-only and site-only evaluations. This trend is not the same for all PTMs. For example, in ADP-ribosylation and UFMylation, the drop is more noticeable because these categories have fewer instances. As a result, even a small number of additional false negatives can significantly affect performance.

Even with this drop, the overall F1-score at the instance level remains close to 90% or higher for most PTMs, with only a few exceptions (for example, ADP-ribosylation at 74.19%). Altogether, the performance is still consistently better than existing systems developed for predicting specific PTMs, as discussed earlier.