pyLucXor: A Python Package for Accurate Phosphorylation Site Localization

Zhongchun Wei, Qixuan Yue, Lei Deng, and Mingze Bai*

Chongqing Key Laboratory of Big Data for Bio Intelligence, Chongqing University of Posts and Telecommunications, Chongqing 400065, China

Abstract. Mass spectrometry-based phosphoproteomics has emerged as an indispensable tool for deciphering protein signaling networks; however, accurate localization of phosphorylation sites remains a persistent analytical challenge. Here, we present pyLucXor, a Python package that implements the LuciPHOr localization algorithm within a framework built upon PyOpenMS. A central methodological contribution of this work is the introduction of an alanine-decoy-based target-decoy strategy, which enables direct, site-level estimation of the false localization rate (FLR). Benchmarking on the publicly available synthetic phosphopeptide dataset PXD000138 demonstrates that, under stringent 1% site-level FLR control, pyLucXor confidently localized 51,468 phosphorylation sites with a localization accuracy of 98.55%, compared with 48,186 sites (98.84% accuracy) identified by the original LuciPHOr. These results indicate that pyLucXor recovers a substantially greater number of phosphorylation sites under equivalent FLR constraints while maintaining comparable localization accuracy. In summary, the alanine-decoy-based global FLR estimation framework provides a statistically principled and computationally scalable quality-control solution for large-scale phosphoproteomic studies.

1 Introduction

Protein phosphorylation is among the most prevalent and biologically consequential post-translational modifications (PTMs), modulating protein activity, localization, and interactions in a reversible manner [1, 2]. Advances in high-resolution tandem mass spectrometry (MS/MS) and enrichment chemistries have enabled phosphoproteomics to measure phosphorylation events at proteome scale, providing a quantitative window into signaling networks and disease mechanisms [3, 4]. Despite these advances, reliable interpretation hinges on not only identifying a phosphopeptide sequence but also localizing the phosphate group to the correct residue(s) within that peptide.

The localization problem is non-trivial for several reasons. First, many peptides contain multiple potential acceptor residues (serine, threonine, tyrosine; S/T/Y), leading to multiple site isomers that can explain the same precursor mass [5]. Second, fragmentation spectra may lack sufficient site-determining ions, and phosphorylation is prone to neutral loss, both of which dilute diagnostic evidence [6]. Third, large-scale studies generate millions of peptide–spectrum matches (PSMs), making computational efficiency and statistically consistent quality control essential [7]. As a result, different tools can yield diverging localization confidence at the same nominal score threshold, and thresholds may not transfer across datasets [8, 9]. This gap motivates methods that are (i) scalable, (ii) interoperable with modern pipelines, and (iii) capable of reporting FLR with transparent statistical meaning.

Several algorithms have become standard for site localization. Ascore scores site isomers based on the presence of site-determining ions using a binomial model [10]. PhosphoRS performs probabilistic inference and reports site-level probabilities that capture partial localization [11]. LuciPHOr extends target–decoy concepts to localization and estimates FLR using modified decoy construction and density modeling [12, 13]. These tools are often integrated into broader proteomics pipelines such as the Trans-Proteomic Pipeline [14]. Despite their effectiveness, existing methods often suffer from heterogeneous interfaces, limited interoperability, and the lack of a unified framework for consistent FLR estimation across large-scale phosphoproteomic datasets.

Here we present pyLucXor, a Python re-implementation of the LuciPHOr localization logic built on PyOpenMS [15]. Beyond re-implementation, we propose an alanine-decoy-based target-decoy approach (TDA) for direct FLR estimation that is computationally light and biologically justified, and we implement a cross-file global FLR module to support study-level quality control.

* Corresponding author: Mingze Bai, baimz@cqupt.edu.cn

2 Methods

2.1 Overview of the pyLucXor workflow and supported formats

pyLucXor is designed to fit into established proteomics pipelines by adopting standard open formats. Raw vendor files are typically converted to mzML for spectra storage, while identification results are exchanged via idXML in the OpenMS ecosystem [7, 16]. pyLucXor consumes mzML (MS/MS spectra) and idXML (PSMs and candidate modified peptide annotations), performs localization analysis, and writes back scores, probabilities, and FLR estimates into idXML user parameters. For modified peptide representation, pyLucXor follows standardized encoding practices to preserve reproducibility and interoperability [17, 18]. Figure 1 outlines the brief analysis workflow of pyLucXor.

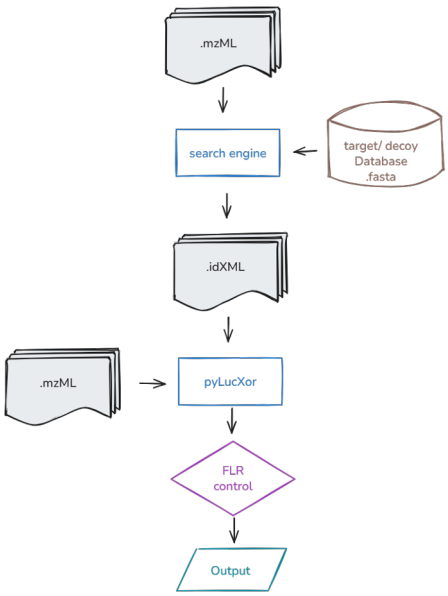

Fig. 1. Brief analysis workflow of pyLucXor.

2.2 Alanine-decoy-based target–decoy approach for direct FLR estimation

The core idea is to create an external set of decoy sites that mimic the scoring behavior of true phosphorylation sites while being biologically implausible as true sites. Alanine is well-suited for this purpose because it lacks the hydroxyl group required for phosphorylation under physiological conditions. In the alanine-decoy strategy, alanine residues within candidate peptides are treated as additional decoy acceptor positions. We introduce phosphorylation mass shifts at alanine positions (annotated as PhosphoDecoy) alongside true phosphorylation annotations (Phospho) during localization scoring. Any localized modification assigned to alanine is therefore counted as a false localization event.

We estimate FLR directly from counts as:

$$FLR = N_{(PhosphoDecoy)} / (N_{(Phospho)} + N_{(PhosphoDecoy)}) \quad (1)$$

where $N_{(PhosphoDecoy)}$ is the number of localizations assigned to alanine-decoy positions and $N_{(Phospho)}$ is the number assigned to S/T/Y positions. This formulation supports both per-file and cross-file aggregation, enabling study-level quality control. To avoid bias, peptides without localization ambiguity (all candidate acceptor sites are modified) can optionally be excluded from FLR estimation.

2.3 Integrated localization algorithms and unified interface

pyLucXor implements the LuciPHOr localization logic: it estimates FLR using a modified target-decoy design and models score distributions using kernel density estimation (KDE), yielding global and local FLR assignments [12, 13]. The original LuciPHOr is retained as an external reference method for benchmarking, enabling direct performance comparison and ensuring that any observed differences can be attributed to implementation or modeling variations rather than methodological inconsistencies. pyLucXor stores all results with consistent naming conventions and supports downstream consensus strategies within broader phosphoproteomic workflows.

2.4 Software architecture and scalability considerations

pyLucXor is implemented in Python with performance-critical components optimized via vectorized numerical computation. The architecture includes: (i) spectrum indexing for rapid MS/MS retrieval, (ii) peptide parsing and site-isomer generation with safeguards against combinatorial explosion, (iii) fragment ion generation with support for neutral loss behavior, (iv) a KDE-based scoring and FLR modeling module, (v) parallel PSM processing, and (vi) a global FLR aggregation module that can pool evidence across multiple runs. This design targets large-scale datasets where both throughput and reproducibility are critical [7, 14].

2.5 Benchmark dataset and evaluation protocol

We benchmark pyLucXor on PXD000138, a synthetic phosphopeptide dataset with known phosphorylation sites, widely used for objective localization evaluation [19, 20]. We report (i) the number of localized sites under a stringent site-level FLR threshold (1%), (ii) site-level localization accuracy against ground truth, (iii) overlap between pyLucXor and LuciPHOr outputs, and (iv) computational performance (processing speed and total runtime) under a standardized environment.

3 Results and Discussion

3.1 Phosphorylation site identification

Both pyLucXor and LuciPHOr processed 111,588 PSMs. The marginal difference arises from the minimum

requirement of LuciPHOr of 50 PSMs per file for model construction, which pyLucXor inherits. As presented in Table 1, under 1% site-level FLR control on PXD000138, pyLucXor identified 51,468 well-resolved phosphorylation sites, compared to 48,186 identified by the original LuciPHOr (local FLR < 0.01). pyLucXor thus recovers approximately 6.8% more well-resolved sites than LuciPHOr under equivalent FLR constraints, reflecting improvements in the re-implementation and the alanine-decoy-based FLR modeling. Other phosphorylation sites, including those of intermediate quality, were not included in the statistical analysis.

Table 1. Comparison of phosphorylation site counts across localization tools on PXD000138

Tool	Total PSMs	Total Phospho Sites	Well Resolved Phospho Sites	Uncertain Phospho Sites
LuciPHOr	111588	118625	48186	37337
pyLucXor	111588	117341	51468	38970

The distribution in Figure 2 is consistent with expected phosphoproteomics characteristics: singly phosphorylated peptides dominate, while multi-phosphorylated peptides are substantially rarer and more difficult to localize. As the number of modified sites increases, evidence becomes diluted and the number of competing site isomers increases, emphasizing the need for rigorous localization QC in multi-site peptides.

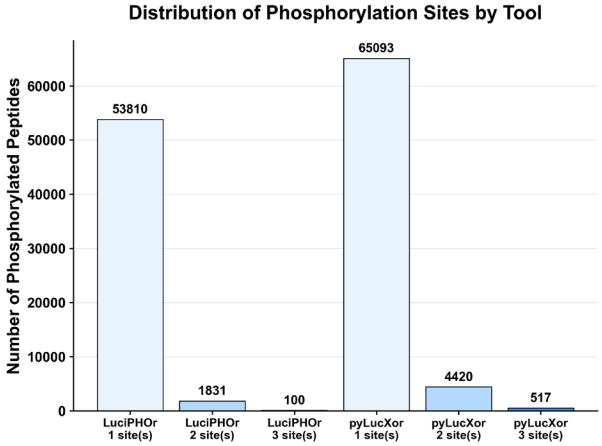

Fig. 2. Distribution of identified peptides by the number of localized phosphorylation sites (1 site, 2 sites, 3 sites) across tools.

3.2 Overlap analysis across tools

To assess consistency, we compared the overlap of localized phosphopeptides between pyLucXor and LuciPHOr. As shown in Figure 3, a substantial common core of identified phosphopeptides indicates that high-confidence signals are consistently recovered by both tools, while tool-specific identifications reflect differences in implementation and FLR modeling assumptions.

The overlap structure demonstrates that pyLucXor successfully reproduces the core localization behavior of LuciPHOr while additionally recovering a unique set of high-confidence sites enabled by the improved alanine-decoy-based FLR estimation.

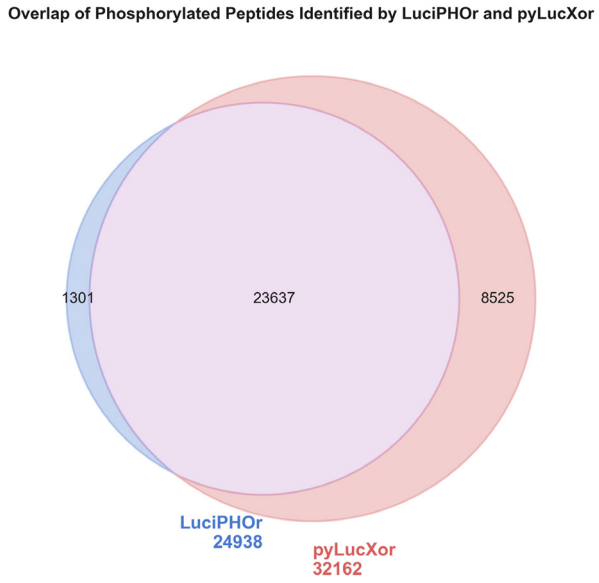

Fig. 3. Venn diagram of phosphopeptides identified by different localization tools under matched quality constraints.

3.3 Site-level localization accuracy

Using the synthetic ground truth with known phosphorylation site annotations, pyLucXor achieved a localization accuracy of 98.55%, closely matching LuciPHOr (98.84%). As presented in Table 2, both FLR-aware methods attain high accuracy, supporting the utility of decoy-based approaches for consistent quality control. The minor accuracy difference between pyLucXor and LuciPHOr (0.29 percentage points) is attributable to differences in implementation details and the alanine-decoy FLR modeling, while pyLucXor identifies a larger number of well-resolved sites overall. This slight decrease in accuracy suggests a trade-off between sensitivity and precision, where pyLucXor recovers additional sites at the cost of a marginal increase in localization errors.

Table 2. Localization accuracy of pyLucXor and LuciPHOr on PXD000138 using ground-truth site annotations

Tool	Total Sites	Correct Sites	Incorrect Sites	Site Accuracy
LuciPHOr	37354	36921	433	0.9884
pyLucXor	40073	39490	583	0.9855

3.4 The global FLR estimation

Single-run FLR calibration can be unstable for sparse strata and for large cohorts where evidence is distributed over many files. The alanine-decoy-based global FLR estimation in pyLucXor aggregates evidence across files and yields a study-level QC curve suitable for large-scale reanalysis projects. As shown in Figure 4, the overall morphology of the FLR curves exhibits characteristic cumulative distribution profiles: a steep ascent is observed in the low FLR region (0-1%), with subsequent

gradual plateauing. This configuration is consistent with the expected distribution of site localization quality, whereby most high-confidence site localizations are enriched in the low FLR region, and the incremental gain of additional sites diminishes as the FLR threshold increases.

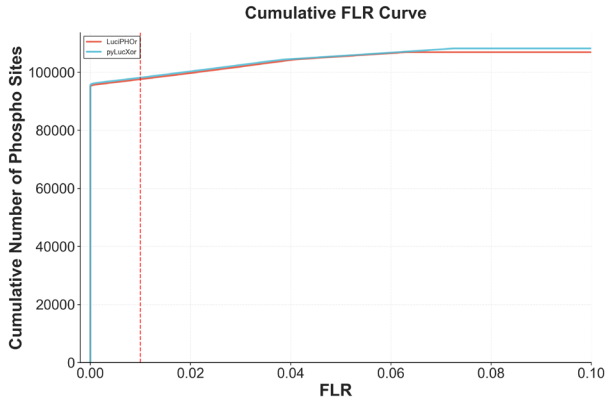

Fig. 4. Cumulative global FLR curves comparing tools.

3.5 Computational performance

The computational performance evaluation was conducted under a standardized environment equipped with a 2-core CPU and 7 GB of RAM, using the first RAW file (1.raw) from the PXD000138 dataset. pyLucXor processed 3,697 peptides (1,948 phosphopeptides) in 26.01 seconds, yielding a throughput of 142.12 IDs/s. The two-stage procedure and density-based FLR modeling incur additional computation, but offer clearer statistical interpretability for quality assurance in large datasets. Compared to the Java version of the original LuciPHOr, pyLucXor benefits from a native Python implementation that integrates directly into modern proteomics pipelines without requiring a separate Java runtime.

4 Conclusion

pyLucXor provides a scalable and statistically grounded framework for phosphorylation site localization in large-scale proteomics. By re-implementing the LuciPHOr algorithm within a PyOpenMS-based pipeline, pyLucXor achieves native compatibility with modern Python-driven proteomics workflows without requiring a separate Java runtime, thereby addressing a key interoperability limitation of the original tool.

A central methodological contribution of this work is the introduction of an alanine-decoy-based target–decoy approach for direct FLR estimation. Alanine residues, which lack the hydroxyl group required for phosphorylation under physiological conditions, serve as biologically plausible decoy acceptor sites. By quantifying localizations assigned to alanine positions relative to true S/T/Y residues, pyLucXor estimates FLR in a manner that is both computationally efficient and statistically transparent. This strategy can be naturally extended to cross-file aggregation, enabling study-level quality control in large cohorts where per-file FLR calibration may be unstable due to limited PSM counts or sparse data strata.

Benchmarking on the synthetic phosphopeptide dataset PXD000138 shows that pyLucXor identifies a greater number of well-resolved phosphorylation sites than the original LuciPHOr under an equivalent 1% FLR threshold, while maintaining comparable localization accuracy. Overlap analysis indicates that pyLucXor largely reproduces the core localization behavior of LuciPHOr, while recovering additional high-confidence sites through improved FLR modeling. The cumulative global FLR curves further support the statistical consistency of the alanine-decoy framework across the full score range.

Overall, pyLucXor offers a practical solution for phosphorylation site localization in large-scale datasets, with an emphasis on transparent FLR estimation and compatibility with existing analysis workflows.

References

1. S.J. Humphrey, D.E. James, M. Mann, Protein phosphorylation: a major switch mechanism for metabolic regulation. Trends Endocrinol. Metab. **26**, 676-687 (2015). https://doi.org/10.1016/j.tem.2015.09.013
2. J.V. Olsen, B. Blagoev, F. Gnad, B. Macek, C. Kumar, P. Mortensen, M. Mann, Global, in vivo, and site-specific phosphorylation dynamics in signaling networks. Cell **127**, 635-648 (2006). https://doi.org/10.1016/j.cell.2006.09.026
3. A. Lundby, A. Secher, K. Lage, N.B. Nordsborg, A. Dmytriyev, C. Lundby, J.V. Olsen, Quantitative maps of protein phosphorylation sites across 14 different rat organs and tissues. Nat. Commun. **3**, 876 (2012). https://doi.org/10.1038/ncomms1871
4. K. Engholm-Keller, M.R. Larsen, Technologies and challenges in large-scale phosphoproteomics. Proteomics **13**, 910-931 (2013). https://doi.org/10.1002/pmic.201200484
5. P.J. Boersema, S. Mohammed, A.J.R. Heck, Phosphopeptide fragmentation and analysis by mass spectrometry. J. Mass Spectrom. **44**, 861-878 (2009). https://doi.org/10.1002/jms.1599
6. K.A. Ramsbottom, A. Prakash, Y.P. Riverol, O.M. Camacho, M.-J. Martin, J.A. Vizcaino, E.W. Deutsch, A.R. Jones, Method for independent estimation of the false localization rate for phosphoproteomics. J. Proteome Res. **21**, 1603 (2022). https://doi.org/10.1021/acs.jproteome.1c00827
7. C. Dai, J. Pfeuffer, H. Wang, P. Zheng, L. Kall, T. Sachsenberg, V. Demichev, M. Bai, O. Kohlbacher, Y. Perez-Riverol, quantms: a cloud-based pipeline for quantitative proteomics enables the reanalysis of public proteomics data. Nat. Methods **21**, 1-5 (2024). https://doi.org/10.1038/s41592-024-02343-1
8. D.D. Shteynberg, E.W. Deutsch, D.S. Campbell, M.R. Hoopmann, U. Kusebauch, D. Lee, L.

Mendoza, M.K. Midha, Z. Sun, A.D. Whetton, R.L. Moritz, PTMProphet: fast and accurate mass modification localization for the Trans-Proteomic Pipeline. J. Proteome Res. **18**, 4262 (2019). https://doi.org/10.1021/acs.jproteome.9b00205

9. M. Locard-Paulet, D. Bouyssie, C. Froment, O. Burlet-Schiltz, L.J. Jensen, Comparing 22 popular phosphoproteomics pipelines for peptide identification and site localization. J. Proteome Res. **19**, 1338-1345 (2020). https://doi.org/10.1021/acs.jproteome.9b00679
10. S.A. Beausoleil, J. Villen, S.A. Gerber, J. Rush, S.P. Gygi, A probability-based approach for high-throughput protein phosphorylation analysis and site localization. Nat. Biotechnol. **24**, 1285-1292 (2006). https://doi.org/10.1038/nbt1240
11. T. Taus, T. Kocher, P. Pichler, C. Paschke, A. Schmidt, C. Henrich, K. Mechtler, Universal and confident phosphorylation site localization using phosphoRS. J. Proteome Res. **10**, 5354-5362 (2011). https://doi.org/10.1021/pr200611n
12. D. Fermin, S.J. Walmsley, A.-C. Gingras, H. Choi, A.I. Nesvizhskii, LuciPHOr: algorithm for phosphorylation site localization with false localization rate estimation using modified target-decoy approach. Mol. Cell. Proteomics **12**, 3409 (2013). https://doi.org/10.1074/mcp.M113.028928
13. D. Fermin, D. Avtonomov, H. Choi, A.I. Nesvizhskii, LuciPHOr2: site localization of generic post-translational modifications from tandem mass spectrometry data. Bioinformatics **31**, 1141 (2014). https://doi.org/10.1093/bioinformatics/btu788
14. E.W. Deutsch, L. Mendoza, D.D. Shteynberg, M.R. Hoopmann, Z. Sun, J.K. Eng, R.L. Moritz, Trans-Proteomic Pipeline: robust mass spectrometry-based proteomics data analysis suite. J. Proteome Res. **22**, 615-624 (2023). https://doi.org/10.1021/acs.jproteome.2c00624
15. H.L. Rost, U. Schmitt, R. Aebersold, L. Malmstrom, pyOpenMS: a Python-based interface to the OpenMS mass-spectrometry algorithm library. Proteomics **14**, 74-77 (2014). https://doi.org/10.1002/pmic.201300246
16. J. Pfeuffer, T. Sachsenberg, O. Alka, M. Walzer, A. Fillbrunn, L. Nilse, O. Schilling, K. Reinert, O. Kohlbacher, OpenMS - a platform for reproducible analysis of mass spectrometry data. J. Biotechnol. **261**, 142-148 (2017). https://doi.org/10.1016/j.jbiotec.2017.05.016
17. R.D. LeDuc, V. Schwammle, M.R. Shortreed, A.J. Cesnik, S.K. Solntsev, J.B. Shaw, M.J. Martin, J.A. Vizcaino, E. Alpi, P. Danis, N.L. Kelleher, L.M. Smith, Y. Ge, J.N. Agar, J. Chamot-Rooke, J.A. Loo, L. Pasa-Tolic, Y.O. Tsybin, ProForma: a standard proteoform notation. J. Proteome Res. **17**, 1321-1325 (2018). https://doi.org/10.1021/acs.jproteome.7b00851
18. R.D. LeDuc, E.W. Deutsch, P.-A. Binz, R.T. Fellers, A.J. Cesnik, J.A. Klein, T. Van Den Bossche, R. Gabriels, A. Yalavarthi, Y. Perez-Riverol, J. Carver, W. Bittremieux, S. Kawano, B. Pullman, N. Bandeira, N.L. Kelleher, P.M. Thomas, J.A. Vizcaino, Proteomics Standards Initiative's ProForma 2.0: unifying the encoding of proteoforms and peptidoforms. J. Proteome Res. **21**, 1189-1195 (2022). https://doi.org/10.1021/acs.jproteome.1c00771
19. Y. Perez-Riverol, C. Bandla, D.J. Kundu, S. Kamatchinathan, J. Bai, S. Hewapathirana, N.S. John, A. Prakash, M. Walzer, S. Wang, J.A. Vizcaino, The PRIDE database at 20 years: 2025 update. Nucleic Acids Res. **53**, D543-D553 (2025). https://doi.org/10.1093/nar/gkae1011
20. H. Marx, S. Lemeer, J.E. Schliep, L. Matheron, S. Mohammed, J. Cox, M. Mann, A.J.R. Heck, B. Kuster, A large synthetic peptide and phosphopeptide reference library for mass spectrometry-based proteomics. Nat. Biotechnol. **31**, 557-564 (2013). https://doi.org/10.1038/nbt.2585