ISSN: 2694-1708
DOI: 10.33552/ICBC.2026.03.000566
Insights in Chemistry and Biochemistry
Iris Publishers

Opinion Article

Mapping The Dark Interactome: Overcoming Founder Effects to Decode Enzyme-Substrate Networks

Wang, R, and Biggar KK*

Institute of Biochemistry, Carleton University, Ottawa ON, Canada K1S 5B6

Corresponding authors: Biggar KK, Institute of Biochemistry, Carleton University, Ottawa ON, Canada K1S 5B6.

Received Date: March 03, 2026

Published Date: March 11, 2026

Opinion Article

The true functional complexity of the human proteome extends far beyond the simple linear sequences that are encoded within our genome. It is governed by a highly dynamic, combinatorial landscape derived from a large number of Possible Post-Translational Modifications (PTMs) that collectively influence protein function. With hundreds of distinct chemical modifications now recognized, and likely many more to be discovered, PTMs dictate protein function, stability, and interactions dynamically and in real-time. This intricate PTM landscape is managed by families of specific enzymes that act as 'writers' to add these chemical tags or 'erasers' to remove them. Unraveling these enzyme-substrate networks is central to understanding control over cellular signaling, as well as the dysregulatory events that drive complex pathologies from inflammatory diseases to cancer.

Despite the need to accurately and reliably map these networks, a vast proportion of the proteome remains functionally uncharacterized, what could be referred to as a "dark interactome" of yet-to-be identified enzyme and substrate pairings. As the field rapidly adopts Machine Learning (ML) to accelerate discovery, we are confronting a critical methodological crossroad. To truly expand and validate these regulatory networks, the computational biochemistry community must move beyond generalized prediction architectures and actively address the systemic biases embedded in our foundational data. Given the unique functions and diversity across PTM-modifying enzymes, the future of substrate discovery likely relies on reduced-bias curation and ML training strategies, likely executed on an enzyme-by-enzyme basis.

- **The challenge of founder bias:** Historically, the expansion of known substrate networks has been dictated by the limitations of the biochemical tools available at the time of study. In studies of post-translational phosphorylation, the availability of pan-phospho antibodies and robust enrichment strategies have historically allowed for the relatively rapid cataloging of kinase networks [1]. This early success facilitated pioneering computational approaches, such as NetworKIN (https://networkin.info/), which integrated consensus motifs with contextual interaction data to refine kinase-substrate predictions. However, for many other critical PTMs, the discovery pipeline has been severely bottlenecked.

The over-reliance on limited experimental data directly feeds into a significant *founder effect* or *founder bias* within modern prediction strategies. When an enzyme is initially characterized, the first few substrates discovered (typically highly abundant proteins or technically accessible modification sites residing in proteotryptic peptides optimized for mass spectrometry identification) may act to disproportionately define our understanding of that enzyme's recognition motif (e.g., we are more likely to discover new substrates using sequence features from that of known substrates). Consequently, public databases become heavily enriched with targets that closely mirror these original substrates, while simultaneously suffering from a near-total absence of experimentally validated negative examples. The issue here is that when modern computational approaches to prediction, such as deep learning models and neural networks, are trained on these skewed datasets, they inevitably echo the founder effect. They learn an artificially narrow, rigid view of the enzyme's substrate permissiveness, failing to capture the true biophysical and structural features that dictate molecular recognition. In the end, we are at risk building sophisticated algorithms that are

exceptionally good at predicting what we already know to be true (e.g., overfit), while remaining entirely blind to the broader "dark interactome".

- **Lysine methylation: an example study in specificity:** The limitations of generalized models and biased training data become starkly apparent when examining complex, chemically nuanced protein modifications like lysine methylation. Unlike phosphorylation, protein methylation lacks robust, modification-specific pan-affinity reagents, making traditional enrichment uniquely challenging [2]. For years, the accepted non-histone substrates for key regulatory enzymes were heavily restricted to a handful of intensely studied targets. Consider SET8, a lysine methyltransferase overexpressed in numerous malignancies; its known substrate pool was historically dominated by targets like p53 (at lysine 382) and histone H4 (at lysine 20) [3,4]. When standard permutation array-based predictions are built solely around these founder motifs, they perform poorly when applied to the wider proteome for validation studies [6]. The assumption that all substrates will rigidly adhere to a simple, sequential consensus motif is a flawed oversimplification of the enzyme's true active site chemistry.
- **A shift toward reduced-bias, ML-Hybrid curation:** To overcome these historical bottlenecks around substrate identification for PTM-modifying enzymes, the field is now witnessing a necessary shift away from reliance on such pre-existing, historically biased, databases. Methodological advancements now beginning to emphasize ML-based approaches that generate comprehensive, enzyme-specific training data from the ground up. However, this strategy hinges on high-throughput experimental generation of training data prior to computational modeling. By chemically synthesizing representative PTM proteomes (e.g., peptide representations of all known modification sites) and subjecting them directly to *in vitro* enzymatic activity, the true permissiveness of an enzyme can be characterized without preconceived biases. However, this approach should still be framed within the limitations of an *in vitro* training data study. These yields relevant, balanced datasets comprising of both positive and, critically, validated negative examples for one specific enzyme that can be used to generate informative ML models for the prediction of an enzyme's substrate network. When applied to enzymes like SET8 (as highlighted previously) or the sirtuin family of NAD+-dependent deacetylases, this reduced-bias methodology drastically outperforms traditional models [6]. Recent ML-based implementations have been able to achieve *in vitro* precision rates exceeding 35-40%, successfully validating hundreds of novel modification sites and improving traditional substrate identification from single digit success metrics. More importantly, these approaches redefine an enzyme's known permissiveness, proving that many active sites are capable of accept a much broader array of flanking amino acids than classical founder motifs ever suggested.
- **Broader implications and next-generation computational strategies:** The implications of mapping these enzyme-specific interactomes extend far beyond basic biochemistry. High-accuracy ML models are becoming essential tools for precision medicine, allowing researchers to rapidly profile how specific genomic missense mutations cause a gain or loss of PTM sites. This capacity to predict "neo-substrates" provides a possibility for entirely new functional narratives for tumourigenesis and disease progression, offering more-informative avenues for therapeutic development.

As the scope of this challenge broadens, the computational strategies underpinning these models are rapidly evolving [7,8]. The integration of structural predictors, such as AlphaFold, with state-of-the-art numerical encoding is enabling researchers to capture deep biophysical components from sequence data alone. Furthermore, the advent of massive Protein Language Models (PLMs), such as ESM-2, driven by transformer architectures, provides an unprecedented ability to learn complex, evolutionary-scale representations of proteins. However, the true utility of these advanced PLMs in mapping the "dark interactome" will only be realized when they are fine-tuned with high-quality, reduced-bias data. Ultimately, the most sophisticated language model cannot correct for a fundamental lack of biological ground truth in its training data. Only by training our most advanced computational tools in unbiased, experimentally validated data can we hope to accurately navigate the full functional landscape of the human proteome.

References

1. Linding R, Jensen LJ, Ostheimer GJ, van Vugt MA, Jørgensen C, et al. (2007) Systematic discovery of *in vivo* phosphorylation networks. Cell 129(7): 1415-1426.
2. Carlson SM, Moore KE, Green AM, Martin GM, Gozani O (2014) Proteome-wide enrichment of proteins modified by lysine methylation. Nature Communications 5(1): 3150.
3. Shi, X, Kachirskaia I, Yamaguchi H, West LE, Wen H, et al. (2007) Modulation of p53 function by SET8-mediated methylation at lysine 382. Molecular cell 27(4): 636-646.
4. Li Z, Nie F, Wang S, Li L (2011) Histone H4 Lys 20 monomethylation by histone methylase SET8 mediates Wnt target gene activation. Proceedings of the National Academy of Sciences. 108(8): 3116-3123.
5. Kudithipudi S, Dhayalan A, Kebede AF, Jeltsch A (2012) The SET8 H4K20 protein lysine methyltransferase has a long recognition sequence covering seven amino acid residues. Biochimie 94(11): 2212-2218.
6. Ridgeway NH, Chopra A, Lukinović V, Feldman M, Charih F, et al. (2025) Machine learning-driven prediction of substrates for enzymes introducing or removing protein post-translational modifications. Communications Chemistry 8(1): 340.
7. Lin Z, Akin H, Rao R, Hie B, Zhu Z, et al. (2023) Evolutionary-scale prediction of atomic-level protein structure with a language model. Science 379(6637): 1123-1130.