Biological molecular function: Beyond homology

Jason E. McDermott

Integrated Discovery Sciences Directorate, Pacific Northwest National Laboratory, Richland, Washington, USA
Jason.McDermott@pnnl.gov

Yana Bromberg
Department of Biology, Emory University, Atlanta, Georgia, USA
yana.bromberg@emory.edu

Hannah Carter
Department of Medicine, Division of Medical Genetics, University of California San Diego, La Jolla, California, USA
hkcarter@health.ucsd.edu

Travis Wheeler
Pharmacy Practice & Science Department, University of Arizona, Tucson, Arizona, USA
twheeler@arizona.edu

Computational annotation of molecular function has long relied on a powerful but fundamentally retrospective strategy: identify a sequence or structure similar to a characterized molecule and transfer what is known. Improved alignment algorithms, structure prediction, and biological language models (LMs) have extended the reach of this paradigm, yet large numbers of genes and proteins, particularly those from viruses, microbiomes, and non-model organisms, remain without useful functional descriptions. Moreover, similarity does not by itself resolve how molecular activity changes with context, i.e. cellular state, genomic neighborhood, interaction network, etc. This session explores complementary routes to function annotation beyond exclusive reliance on homology. The papers in the session use genomic neighborhoods, genetic fitness measurements, knowledge graphs, biological network topology, metabolic feasibility, learned sequence representations, open-ended reaction generation, and context-dependent descriptions of retained protein elements. Together, they suggest that "beyond homology" is not a single replacement technology but rather a shift toward integrating evidence across scales, representing function in context, evaluating methods under genuinely difficult generalization settings, and reporting uncertainty or failure when available data do not support a functional claim. The session papers also expose unresolved challenges: context can improve system-level coherence without identifying the responsible protein; generative models can produce plausible chemistry without assigning it correctly; and calibrated abstention may be more informative than a forced prediction. This session frames these advances and limitations as a research agenda for moving from annotation transfer toward mechanistic and testable hypotheses of biological function.

Keywords: molecular function; protein function prediction; biological dark matter; genomic context; knowledge graphs; metabolic modeling; network biology; protein language models; uncertainty calibration.

© 2026 The Authors. Open Access chapter published by World Scientific Publishing Company and distributed under the terms of the Creative Commons Attribution Non-Commercial (CC BY-NC) 4.0 License.

1. Introduction

Advances in sequencing technology have vastly outpaced experimental characterization of protein function. It is estimated that in well-characterized organisms 30-60% of proteins can not be assigned specific functional roles. The situation is even worse in metagenomes and viromes, where diversity and evolutionary distance from the known collection of molecules is a drastically limiting factor (Hatfull, 2015; Lobb, et al., 2020; Schloss and Handelsman, 2005). The vast majority of functional annotations are assigned on the basis of gene/protein homology, captured by sequence similarity. Sequence alignment, profile Hidden Markov Models, and increasingly sensitive database-search methods are used across the sequence landscape (Altschul, et al., 1990; Buchfink, et al., 2021; Eddy, 2011; Roddy, et al., 2024; Steinegger and Soding, 2017). Predicted three-dimensional structures now permit searches across structural space, recovering relationships that may not be apparent from primary sequence (van Kempen, et al., 2023) and emerging protein LMs aim to identify distributed sequence patterns to produce representations useful for many downstream predictive tasks (Bileschi, et al., 2022; Rives, et al., 2021). These advances have greatly expanded the detectable neighborhood around known biology.

Yet expanding that neighborhood is not the same as understanding function. Most similarity-based approaches remain anchored to an existing annotation: a new molecule is interpreted through resemblance to something previously studied. Their success therefore depends on the coverage, accuracy, and granularity of reference knowledge. Existing labels often describe only one layer of function, for example catalytic activity, while omitting the conditions, partners, location, regulation, or phenotypic consequences, that determine what the molecule does in a living system.

The phrase “molecular function” itself hides several distinct questions. What chemical transformation can a protein catalyze? What substrate does it act upon? In which compartment, condition, or organism does that activity matter? What role does it play in a pathway or interaction network? How do regulatory elements or domain combinations alter the effect of an otherwise familiar component? A protein can retain a recognizable domain yet have a wholly different biological role when found in a different context (Jeffery, 1999; Piatigorsky and Wistow, 1991) or included within a different domain architecture (Bashton and Chothia, 2007; Todd, et al., 2001). Conversely, proteins without detectable homology can converge on related system-level roles (McDermott, et al., 2019). Function is therefore not only a property of an isolated sequence; it is an interaction among molecular capabilities, environment, and the level at which the question is asked.

This session, “Biological molecular function: Beyond homology,” builds on the PSB 2026 session on methods and benchmarks for finding function in biological dark matter (McDermott, et al., 2026). Discussion at that session identified the dependence of most annotation strategies on sequence or structural similarity as both an enduring strength and an important frontier. The challenge is not to discard homology, as evolutionary relationships remain among the strongest sources of functional evidence and establish a demanding baseline (Hamp, et al., 2013). Rather, the goal is to determine what additional evidence becomes informative where direct annotation transfer becomes unreliable, incomplete, or conceptually insufficient.

The accepted papers in this session span microbes, bacteriophages, metagenomes, metabolism, fusion proteins, and cancer signaling. Their methods differ substantially, but they collectively move away from treating a protein as an isolated sequence requiring a single inherited label. They use

neighboring genes, condition-specific mutant fitness, heterogeneous knowledge graphs, network topology, organism-level biochemical constraints, and retained regulatory elements. Other contributions ask whether learned molecular representations can support direct enzyme discovery or generation of biochemical reactions, RNA–protein interactions, pharmacogenomics, and structure-based cryptic-pocket detection, extending context from molecular partners and substrates to explicit local geometry. Importantly, these papers do not present a simple story of homology being replaced. Several report that performance declines under stricter generalization tests, that improvements at the system level do not translate to protein-level labels, or that generated chemistry is not correctly assigned. These outcomes help improve problem definition and point toward evaluation, uncertainty, and abstention as essential components of functional inference.

2. Session Contributions

The session papers can be organized around four complementary ideas: (1) context can supply functional evidence that is absent from an isolated sequence, (2) models can infer biochemical capability directly from molecule representations, while system-level constraints can filter molecular predictions for biological plausibility, (3) progress depends on representations that capture context-dependent effects and on uncertainty estimates to distinguish supported hypotheses from extrapolation, and (4) molecular partners, substrates, and local structural environments define function in ways that require interaction-aware, condition-specific, or auditable representations.

2.1. Function from genomic, phenotypic, and network context

Genomic organization is one of the most direct forms of non-homology evidence available at scale. In prokaryotes and viruses, neighboring genes often participate in related processes or functional modules. Hoang et al. introduce Synnotate, which masks a target gene and annotates it using the standardized annotations and arrangement of genes in a 21-gene neighborhood (Hoang, et al., 2027). The model uses a sparse-attention transformer and provides confidence estimates, attribution to individual neighbors, and retrieved reference neighborhoods. Under genus-disjoint cross-validation, it reaches high accuracy in annotating prokaryotic and bacteriophage genes. Combining model confidence with a synteny score further defines trusted regions. Synnotate is able to confidently annotate nearly 400K prokaryotic dark genes and 42K of unannotated phage proteins.

The method identifies candidates missed by direct sequence and structural searches while retaining interpretable evidence for each prediction demonstrating a practical model for inference beyond homology: use an independent contextual signal, test it across taxonomic boundaries, expose the evidence supporting a prediction, and provide boundaries on the calibrated region. Its application to bacteriophages is especially relevant because viral genomes contain extensive uncharacterized sequence space and often lack tractable experimental systems.

Experimental phenotypes provide another layer of context. Anthony et al. describe GIMME, a heterogeneous biological knowledge graph (KG) integrating nearly eight million random-barcode transposon sequencing fitness measurements with genes, experiments, strains, experimental metadata, metabolic knowledge, text embeddings, and protein language-model embeddings (Anthony, et al., 2027). Their graph-attention model, GFM1, predicts condition-specific gene knockout-determined fitness across ten *Pseudomonas* strains. Relative to a matched multilayer

perceptron, message passing significantly improves predictions for held-out experiments and held-out genes. The largest benefits are observed for highly connected, conditionally essential genes, recovering many new true-essential gene–condition observations.

These results illustrate the value of combining molecular representations with experimental phenotype under many conditions, but also point out limitations. The graph model's advantage disappears under a more stringent cluster-aware split, indicating that successful message passing still depends on informative connections to related training genes. Thus, context can extend functional prediction beyond isolated sequence similarity without eliminating dependence on the structure and coverage of known biology. The study reveals this limitation by using increasingly demanding splits, an evaluation practice that should become standard for claims of generalization beyond homology.

At a different scale, Jacobs et al. ask whether non-homologous cancer driver genes can be grouped by convergent roles in signaling networks (Jacobs, et al., 2027). INSPIRE performs within-species alignment of signaling subnetworks associated with colorectal cancer driver genes, comparing interaction patterns and topology rather than relying on gene-level resemblance alone. Consensus clusters show stronger within-group CRISPR co-essentiality than between-group pairs and identify exploratory epistatic relationships. Aggregating genes into the inferred modules also reveals significant mutual exclusivity that is not apparent in single-gene analysis. Although molecular annotations such as Pfam and Gene Ontology contribute to alignment of interaction endpoints, the relationships among drivers emerge from network organization and shared system-level roles. INSPIRE predicts the impact that distinct genes have on a common network program, broadening the notion of function beyond transfer of a biochemical label.

Together, Synnotate, GIMME, and INSPIRE move across local genome organization, condition-specific phenotype, and cellular signaling topology. They show that context can identify functional patterns at multiple scales, but also that the strength of functional inference depends on graph connectivity, available measurements, and the biological level at which function is defined.

2.2. From molecular representations to biochemical capability

An ambitious route beyond homology is to predict general biochemical activity directly from learned representations, common to specific functions, rather than retrieving a label from a similar annotated protein. REFinder, presented by Prabakaran and Bromberg, targets enzyme discovery in metagenomic sequence data (Prabakaran and Bromberg, 2027). Its first stage uses REBEAN, an alignment-free DNA language model, to classify reads into broad Enzyme Commission (EC) classes. Reads predicted to encode enzymes are then routed into focused assembly, reducing the need to assemble an entire metagenome and functionally labeling previously unseen, "dark matter" protein genes/sequences. Optional sequence-, domain-, structure-, metal-binding-, and representation-reliability analyses provide additional evidence for the resulting candidates.

Across human saliva and murine gut metagenomes, REFinder recovers approximately 2-3 times more enzymes than sequence similarity alone producing many annotated sequences without close Swiss-Prot matches. Note that in one sample, thousands of these sequence-divergent candidate oxidoreductases had strong structural matches to known enzyme folds. Class-targeted assembly can cost up to 6.4-fold less than whole-metagenome assembly in terms of computation and can recover poorly covered (lower-depth sequencing) genes. The framework therefore links functional inference

to the design of the computational discovery process itself: predicted activity determines which reads are assembled.

The interpretation of these candidates requires care. Structural correspondence supports a fold but does not establish catalytic activity, broad EC classes provide limited specificity, and the present evaluation covers two comparatively well-studied microbiome types. Nonetheless, REFinder demonstrates how alignment-free learned signal can prioritize previously inaccessible sequence space and generate testable enzyme candidates at metagenomic scale.

Balamukundhan et al. push direct inference further with RxnPred, an encoder–decoder transformer architecture that generates reaction SMILES from protein sequence and structure embeddings (Balamukundhan, et al., 2027). The model combines ESM-2 sequence representations, ESM-IF1 structural representations, and ChemBERTa metabolite representations. Unlike classification into an EC category or retrieval from a supplied reaction pool, RxnPred attempts open-ended generation of the biochemical transformation itself. This formulation is appealing because the reaction is closer to mechanistic function than a broad class label and because generation could, in principle, describe chemistry absent from the training set.

RxnPred is competitive for reactions observed during training, trading precision for increased recall. It can also generate previously unseen reaction strings, including some valid reactions found in the Rhea reaction knowledgebase (Bansal, et al., 2022). However, these reactions are generally assigned to the wrong enzyme, and the model does not recover the correct unseen reactions in evaluations where retrieval methods can rank them because those reactions are supplied at inference. Performance also decreases as test proteins become less similar to training proteins. These findings distinguish two capabilities that are easily conflated: generating chemically plausible reactions and grounding a particular reaction in the molecular determinants of a particular protein. The first is emerging; the second remains a central challenge.

METEOR, developed by Niu and Hoehndorf, addresses biochemical function from the opposite direction (Niu and Hoehndorf, 2027). Instead of asking only which reaction a protein resembles, it asks which collection of predicted reactions can coexist in a mass-balanced metabolic network capable of supporting growth. METEOR aggregates protein-level EC predictions into reaction confidence and uses confidence-weighted mixed-integer optimization under mass-balance and biomass-feasibility constraints to select an organism-level reaction repertoire. Across enzyme predictors and bacterial genomes, the resulting networks contain significantly fewer dead-end metabolites than raw thresholded predictions. All repaired METEOR networks grow under the defined medium, whereas none of the raw thresholded networks do. The method also recovers more curated sub-threshold EC assignments vs. conventional thresholding followed by repair.

The constraints improve global biochemical coherence but do not consistently improve per-protein annotation accuracy or pathway inference against a size-matched baseline, and selected reactions are not assigned to specific proteins or paralogs. This distinction is consequential. Organism-level feasibility can reveal that a capability is probably present without identifying the molecule that carries it out. METEOR therefore suggests that function prediction should sometimes be decomposed into at least two tasks: determining the biological capabilities required by the system and attributing those capabilities to molecular actors.

Taken together, REFinder, RxnPred, and METEOR span molecular function discovery, generation, and constraint. Learned representations expand the space of candidate activities; generative models test whether chemistry can be produced without a closed label set; and metabolic models enforce physical and physiological consistency. Their limitations suggest that robust enzyme-function inference will likely require these forms of evidence to interact rather than compete – for example, using generative proposals as hypotheses, genomic or structural evidence for molecular grounding, and system constraints to evaluate organism-level plausibility.

2.3. Representing context-dependent effects and calibrated uncertainty

Moving beyond homology also requires better representations of what a functional annotation means. Murakami examines fusion proteins, for which identifying retained domains or protein segments may be straightforward, while predicting the biological consequence of their combination is not (Murakami, 2027). The proposed signed element-contribution framework represents fusion function as a sum of context-dependent contributions from retained elements. Elements may act as effectors, autoinhibitory regions, oligomerization modules, or neutral components, and their default positive or negative contributions can reverse under specified conditions.

This representation addresses a gap in common resources. A fusion can preserve an enzymatic or interaction domain while losing an autoinhibitory segment, converting a familiar molecular component into a constitutively active or pathogenic one. Here, homology may correctly identify every retained part yet fail to describe the sign of its contribution in the new architecture. The study finds that retention of effector elements and loss of autoinhibitory elements depart from random breakpoint expectations, supporting the relevance of signed roles. It also highlights that autoinhibitory regions remain poorly represented in machine-readable annotations.

The work uses conformal prediction to retrieve precedents and does not provide a prediction when no known mechanism is sufficiently similar (~15% of the queries). At the same time, mechanism-class identification does not outperform baseline because the current role vectors mostly encode how much of a protein is retained rather than which specific mechanism remains. This negative result is informative: no uncertainty procedure can recover distinctions that the representation and source annotations do not contain. Progress therefore depends not only on larger models but also on richer, explicitly mechanistic data resources.

Calibration and limitation connect this work to the other contributions. Synnotate limits high-confidence annotation to a trusted region. GFM1's split-dependent performance identifies settings where graph context does not support generalization. RxnPred exposes the difference between plausible output and correct molecular assignment. METEOR separates system coherence from protein-level certainty. Across these papers, a responsible beyond-homology method is not one that always emits a label. It is one that communicates which evidence supports a claim, the biological scale at which the claim is valid, and when the available evidence is insufficient.

2.4. Molecular interactions, substrate-dependent effects, and auditable structural inference

Islam et al. investigate RNA–protein interaction prediction with Qfusion, which combines fixed ESM-2 protein and RNA-FM RNA embeddings through an eight-qubit, eight-layer circuit with cross-register correlators (Islam, et al., 2027). In family-and-clan-disjoint evaluation, Qfusion attains

a mean AUROC somewhat higher than a parameter-matched classical control. The AUPRC difference from the matched control is significant across evaluation folds, whereas the AUROC difference is not. Although the circuit itself has no trainable parameters, the complete model includes trainable projection and readout components. Because evaluation uses a classical statevector simulator, the study claims neither runtime nor sample-complexity quantum advantage. Its contribution is a competitive cross-modal coupling mechanism whose ablations identify cross-register correlators as important to the observed gain.

Xu et al. examine variant-effect prediction for six cytochrome P450 pharmacogenes, with activity modeling focused on CYP2C9 (Xu, et al., 2027). Ambiguous AlphaMissense calls are enriched across these proteins. Within the ambiguous CYP2C9 class, AlphaMissense scores correlate weakly with measured activity, while ESM-2 retains some signal. A CYP2C9-specific nearest-neighbor model combining ESM-2 embeddings with AlphaMissense and ESM-2 scores improves overall activity correlation and ambiguous-class correlation. These biochemical improvements do not reliably transfer to sparse clinical labels. Moreover, more than a tenth of the variants annotated for multiple drugs have distinct effects across substrates. This finding motivates substrate-conditioned prediction as a future direction rather than establishing a validated substrate-aware model, and underscores that biochemical activity prediction and clinical interpretation are distinct targets.

Yu and Wu present an algebraic counting field for detecting cryptic pocket-lining residues from apo protein structures. At inference, the method uses no sequence search, structural-template transfer, or protein language model. It expands explicit geometric, physicochemical, and topological quantities into discretized neighborhood features, with scores reconstructable from lookup-table counts, integer multiplicities, and spatial aggregation. This provides exact numerical traceability, although not by itself a causal or mechanistic explanation.

On a locked external set of apo–holo pairs, the method performance was somewhat better than that of P2Rank. A second predeclared external evaluation did not confirm improvement from an expanded feature set. The principal result is therefore competitive prediction with an auditable scoring path and explicit external validation, not universal state-of-the-art performance. The output identifies benchmark cryptic residues, not ligand identity, affinity, druggability, or therapeutic efficacy.

3. Conclusions

The papers in this session present “beyond homology” not as a rejection of evolutionary inference but as an expansion of the evidence and representations used to understand molecular function. Sequence and structural similarity remain indispensable. However, they are incomplete when no characterized relative is detectable, when an annotation lacks mechanistic specificity, or when biological consequence depends on context that an isolated sequence cannot encode.

Several common principles emerge. First, context is measurable and computationally useful. Gene neighborhoods, perturbation phenotypes, interaction networks, and organism-level metabolic constraints each add information that cannot be reduced to pairwise sequence similarity. Second, function exists at multiple scales. A model may correctly infer a system capability without assigning it to a protein, or may identify a network role without proposing a biochemical mechanism. Claims

and benchmarks should state clearly which level is being predicted. Third, direct and generative models require grounding. Producing a plausible enzyme class or reaction is not equivalent to establishing that a particular molecule performs it. Independent genomic, structural, chemical, phenotypic, and physiological evidence will be needed to convert predictions into compelling hypotheses. Molecular partners, substrate identity, and local structural geometry further show that the relevant context may reside at the interface between molecules or in a specific physical environment.

Finally, rigorous negative results, calibrated confidence, and clearly accounting for limitations are central to progress. The hardest test of a beyond-homology method is not performance on a random split populated by close relatives. It is whether useful information survives lineage separation, sequence-cluster separation, unseen chemistry, sparse context, or a missing mechanistic precedent. The studies represented here make those boundaries visible. Collectively, they point toward a future in which functional annotation is less often a single inherited label and more often an evidence-weighted, context-specific, and testable account of what a molecule can do, where it does it, and how confidently we know.

Acknowledgments

Funding to support this work was provided to J.E.M. by Pacific Northwest National Laboratory's Predictive Phenomics Initiative conducted under the Laboratory Directed Research and Development Program. PNNL is a multiprogram national laboratory operated by Battelle for the U.S. Department of Energy under Contract No. DE-AC05-76RL01830. Funding to support J.E.M. and T.J.W. was provided under NIH/NIDCR U01DE034176-01. Y.B. was supported by was supported by National Science Foundation award number 2310114.

References

Altschul, S.F.*, et al.* Basic local alignment search tool. *J Mol Biol* 1990;215(3):403-410.

Anthony, W.*, et al.* Predicting microbial gene fitness with a biological knowledge graph. *Pacific Symposium on Biocomputing* 2027.

Balamukundhan, S., Jagarlapudi, S.N. and Raman, K. RxnPred: A generative framework for enzymatic reaction prediction. *Pacific Symposium on Biocomputing* 2027.

Bansal, P.*, et al.* Rhea, the reaction knowledgebase in 2022. *Nucleic Acids Res* 2022;50(D1):D693-D700.

Bashton, M. and Chothia, C. The generation of new protein functions by the combination of domains. *Structure* 2007;15(1):85-99.

Bileschi, M.L.*, et al.* Using deep learning to annotate the protein universe. *Nat Biotechnol* 2022;40(6):932-937.

Buchfink, B., Reuter, K. and Drost, H.G. Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nat Methods* 2021;18(4):366-368.

Eddy, S.R. Accelerated Profile HMM Searches. *PLoS Comput Biol* 2011;7(10):e1002195.

Hamp, T.*, et al.* Homology-based inference sets the bar high for protein function prediction. *BMC Bioinformatics* 2013;14 Suppl 3(Suppl 3):S7.

Hatfull, G.F. Dark Matter of the Biosphere: the Amazing World of Bacteriophage Diversity. *J Virol* 2015;89(16):8107-8110.

Hoang, K.*, et al.* Synnotate: Genomic-context based annotation of the prokaryotic and viral dark proteome. *Pacific Symposium on Biocomputing* 2027.

Islam, A., Subramanian, G. and Singh, R. RNA-Protein Interaction Prediction Using Hybrid Quantum-Classical Fusion. *Pacific Symposium on Biocomputing* 2027.
Jacobs, C.*, et al.* INSPIRE: A network-based method for quantifying functional convergence in cancer driver genes. *Pacific Symposium on Biocomputing* 2027.
Jeffery, C.J. Moonlighting proteins. *Trends Biochem Sci* 1999;24(1):8-11.
Lobb, B.*, et al.* An assessment of genome annotation coverage across the bacterial tree of life. *Microb Genom* 2020;6(3).
McDermott, J.*, et al.* Prediction of bacterial E3 ubiquitin ligase effectors using reduced amino acid peptide fingerprinting. *PeerJ* 2019;7(e7055).
McDermott, J.E.*, et al.* Session Introduction: Biological molecular function: methods and benchmarks for finding function in biological dark matter. *Pac Symp Biocomput* 2026;31:417-424.
Murakami, K. A signed element-contribution framework for describing the function of fusion proteins. *Pacific Symposium on Biocomputing* 2027.
Niu, K. and Hoehndorf, R. METEOR: Evidence-weighted reaction selection reconciles enzyme prediction with growth feasibility. *Pacific Symposium on Biocomputing* 2027.
Piatigorsky, J. and Wistow, G. The recruitment of crystallins: new functions precede gene duplication. *Science* 1991;252(5009):1078-1079.
Prabakaran, R. and Bromberg, Y. REFinder: Mining new enzymes from metagenomes. *Pacific Symposium on Biocomputing* 2027.
Rives, A.*, et al.* Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc Natl Acad Sci U S A* 2021;118(15).
Roddy, J.W., Rich, D.H. and Wheeler, T.J. nail: software for high-speed, high-sensitivity protein sequence annotation. *bioRxiv* 2024.
Schloss, P.D. and Handelsman, J. Metagenomics for studying unculturable microorganisms: cutting the Gordian knot. *Genome Biol* 2005;6(8):229.
Steinegger, M. and Soding, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat Biotechnol* 2017;35(11):1026-1028.
Todd, A.E., Orengo, C.A. and Thornton, J.M. Evolution of function in protein superfamilies, from a structural perspective. *J Mol Biol* 2001;307(4):1113-1143.
van Kempen, M.*, et al.* Fast and accurate protein structure search with Foldseek. *Nat Biotechnol* 2023.
Xu, H.*, et al.* Homology-Based Variant-Effect Predictors Break Down on Cytochrome P450 Pharmacogenes. *Pacific Symposium on Biocomputing* 2027.