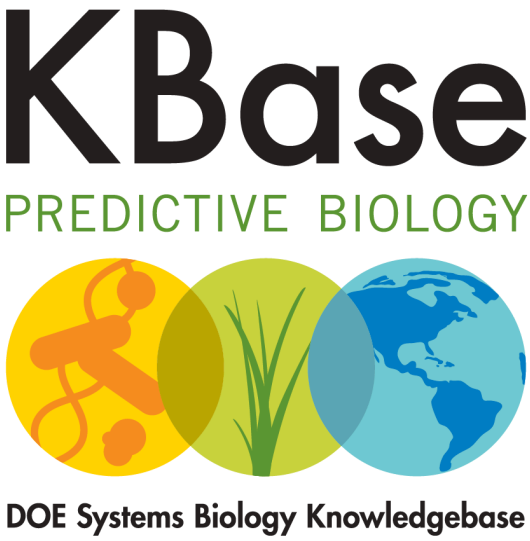

DOE BSSD Performance Management Metrics Report FY 2025 Q4: Incorporating AI/ML for Next-Generation Capabilities

Authors: Paramvir Dehal[1] (psdehal@lbl.gov), Chris Henry[2] (chenry@mcs.anl.gov),Gazi Mahmud[1] (GaziMahmud@lbl.gov), Elisha-Wood Charlson[1] (elishawc@lbl.gov), Roy Kamimura[1] (royk@lbl.gov), Adam Arkin[1] (aparkin@lbl.gov)

[1]Lawrence Berkeley National Laboratory, Berkeley, CA 94720 and [2]Argonne National Laboratory, Argonne IL 60439

KBase's Evolution Toward Predictive Biology

Since its inception, KBase has been committed to enabling data interoperability (Q1), advancing collaboration in team-oriented science (Q2), and expanding partnerships with external data resources and modeling communities (Q3). These foundations now converge in FY25 Q4, as KBase begins embedding artificial intelligence (AI) and machine learning (ML) directly into its workflows.

This marks a turning point where KBase moves from a platform of integrated data and tools to a next-generation knowledge environment for predictive biology. The incorporation of AI/ML enables automation of repetitive tasks, deeper reasoning over complex datasets, and more interactive guidance for users. These capabilities will accelerate the DOE BER mission of connecting genomes to phenotypes to ecosystems.

Key Emerging AI/ML Capabilities in KBase

The KBase Narrative Agent

The KBase Narrative Agent is a prototype AI-driven system designed to enhance the usability and efficiency of the KBase platform. Its goal is to lower the barrier to computational biology, accelerate research, and make advanced bioinformatics tools more accessible to the BER research community.

The agent leverages recent advances in large language models (LLMs) for scientific reasoning, task automation, and knowledge retrieval. By integrating these capabilities into the KBase Narrative environment, the system acts as an interactive co-pilot that:

- Guides users in refining their analysis goals and clarifying their scientific questions.
- Designs customized workflows by retrieving relevant information from KBase documentation and tutorials using retrieval-augmented generation[1] (RAG).
- Automates workflow execution through a modular multi-agent framework.
- Interprets and summarizes results, generating markdown narratives to support scientific reporting and publication.

The Narrative Agent is built using LangGraph[2], a stateful agent framework built on LangChain[3]. In this architecture, each workflow is expressed as a directed graph of modular nodes that encapsulate specific tasks such as planning, execution, or summarization. This structure enables:

- Iterative planning and re-execution, allowing the agent to refine analysis as new results emerge.
- Parallel task execution, supporting modular and scalable workflows such as genome annotation or metagenome assembly.
- Human-in-the-loop (HITL) checkpoints, ensuring domain experts can intervene at critical decision points without disrupting the overall programmatic structure.

Together, these capabilities provide a flexible and controllable agentic workflow system that reduces the steep learning curve for new users while allowing advanced researchers to scale complex analyses. In FY25, the initial prototypes have focused on workflow guidance and result interpretation, with the long-term vision of transforming the Narrative Agent into a more fully functional KBase Research Assistant capable of interfacing with the Narrative and the KBase BER Data Lakehouse. This Research Assistant would then be embedded across KBase to support interactive, AI-augmented science at scale.

The KBase Gene Function Agent

The KBase Gene Function Agent addresses one of the most difficult challenges in biology: assigning reliable functions to genes. This capability is foundational to nearly all other types of biological analysis in KBase from specific pathway engineering for biomanufacturing, genotype-to-phenotype prediction for microbes and phage, and understanding microbial community function and activity. Current annotation pipelines often provide conflicting or incomplete results, particularly for less-studied microbial genomes. To overcome this, the Gene Function Agent integrates multiple lines of evidence and applies AI reasoning to generate improved gene function predictions.

At its core, the agent draws on curated experimental resources, including the gene re-annotations from RbTnSeq fitness assays published by Price et al[4]. and made available through the Fitness Browser[5]. These data provide experimentally supported insights into gene essentiality and function across a wide range of microbial organisms. The Gene Function Agent complements this with literature evidence retrieved through PaperBLAST[6], enabling it to incorporate relevant publications into the annotation process.

The agent operates in two stages:

1. Evidence synthesis: it reviews the RbTnSeq fitness data, relevant publications, and any overlapping annotations, then outputs an updated gene function prediction along with a narrative explanation of its reasoning.
2. Evaluation: a separate evaluation LLM compares the agent's predictions against the gold-standard human annotations curated in the Fitness Browser. This dual-agent design provides both performance benchmarking and transparency, ensuring that improvements over existing annotations can be quantified and audited.

This approach transforms gene annotation from a static process into an evidence-driven, iterative workflow, where machine reasoning and human curation can be tightly integrated. In FY25, early results demonstrate the feasibility of AI-assisted annotation pipelines that reason across experimental datasets and literature simultaneously, laying the groundwork for scalable curation of millions of genes in FY26 and beyond. To accomplish this, we will utilize the KBase BER Data Lakehouse to organize, link, and calculate all lines of evidence- beyond the two sources above- and allow the gene function agent to iteratively update and harmonize gene function annotations across the data store.

A Generative AI Foundation Model for Genome Composition

Understanding the gene content of microbial genomes is essential for connecting sequence data to phenotype and ecosystem function. However, draft microbial genomes and metagenome-assembled genomes (MAGs) are often incomplete or fragmented, making it difficult to recover accurate functional profiles. To address this, KBase is developing a generative AI foundation model for microbial genome composition, trained directly on species-level pangenomes spanning thousands of microbial lineages from the GTDB[7] resource. The initial goal of the generative framework is to probabilistically fill in holes to predict full genetic content of partial MAGs. In the future, this model will be extended to aid in the design of augmented genomes for biotechnological applications in biomanufacturing and agricultural support.

This work builds on earlier research applying generative adversarial networks (GANs) to *Pseudomonas* genomes[8], where gene presence/absence matrices could be realistically generated, and missing content in incomplete genomes identified. Building from that proof of concept, KBase is now developing a large language model (LLM)-based generative framework, designed specifically for microbial data.

The approach involves several innovations:

- Custom tokenizer: encoding pangenome gene family IDs, gene functions (EggNOG, UniRef), and contig breaks so the model can “read” genomes at the gene level.
- Pre-training on ~500,000 annotated microbial genomes, treating each genome as a structured document containing organism metadata, genome quality, and gene-level features.
- Training objectives adapted to biological data.
- Contig shuffling: making the model robust to unordered draft assemblies.
- Missing gene prediction: allowing the model to recover absent genes given the surrounding genomic context.
- Fine-tuning tasks: genome contamination detection, chimera checking, contig ordering, and genome completion.

Early results show that the model achieves 80–90% accuracy in predicting both core and accessory genes missing from draft genomes, a major advance toward scalable genome recovery. Looking forward, the vision is to enable natural language queries, where a user could specify: “Build a root-associated pseudomonad that tolerates drought stress and produces a desired metabolite.” The model would then generate a base genome plus recommended gene additions or modifications to meet those criteria.

This foundation model thus serves a dual purpose: (1) improving genome quality and completeness for BER datasets (especially metagenome-assembled genomes), and (2) laying the groundwork for synthetic biology applications, where users could explore hypothetical microbial genomes designed for specific traits or environments.

Protein Language Model for Sequence Search

The determination of protein similarity is a key component of numerous genome annotation, comparative genomics, and omics data analysis pipelines, yet this computation has become increasingly expensive to run as the volume of available reference genomic data has grown geometrically over the years. With the advent of protein language models (PLMs) from recent advances in deep learning for proteins, we finally have a means of making protein similarity search vastly more scalable and more sophisticated with benefits for accuracy. Previously, the protein similarity computation revolved around the alignment of a query protein sequence against a reference set of sequences to determine the similarity of the query to all available reference sequences. PLMs offer a means of effectively translating a protein sequence into a high-dimensional vector, translating the computationally expensive string comparison operation in the protein similarity workflow into a much more performant and scalable vector comparison operation (e.g. cosine similarity). To leverage these advantages, we applied the ESM2 PLM to precompute mean-averaged embeddings for every protein in UniProt, storing the resulting vectors in a vector database to support rapid distance query. To query this database, we then compute an embedding for the query sequence and then query the vector DB with the resulting embedding.

While this approach already greatly improved performance, we encountered challenges that needed to be addressed to make this database operate efficiently. First, the storage of ESM2 embeddings for more than 100M proteins in UniProt was costly in memory requirements and computation to execute queries. To address this challenge, we quantized the database, converting floating-point embeddings into binary embeddings. This greatly reduced memory requirements and query compute costs for our PLM similarity service, but most surprisingly, for ESM2 models over a specific size, this improved accuracy as well (Figure 1). In this case, accuracy was determined based on the ability of the PLM similarity metric to correctly map query proteins to the same PFAM and UniRef90 protein cluster as classic string-comparison-based approaches.

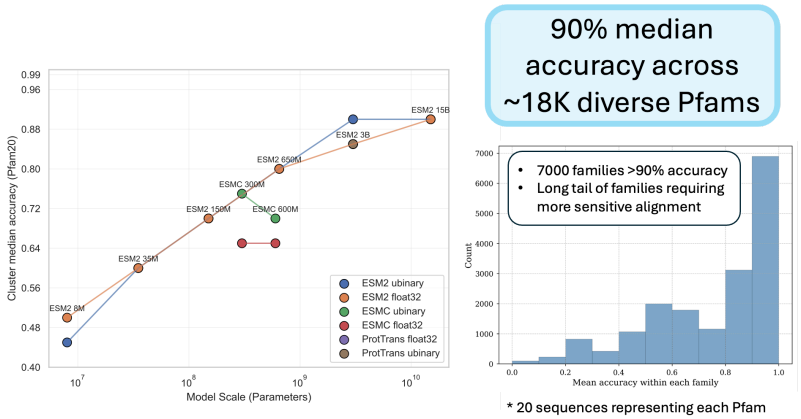

Figure 1. Accuracy of mapping query proteins into proper UniRef and PFAM protein clusters based on embedding similarity using EMS2 embeddings from a variety of model sizes.

Finally, we made one other enhancement to improve the performance of our similarity service by implementing a hierarchical search strategy in which the query sequence is first compared against a smaller number of representative sequences, before descending the search tree to obtain hits with increasing resolution. With all of our enhancements, we now have a PLM-based protein query service that can query all proteins in a single genome against all of UniProt in less than 1 minute. The accuracy of the top hit for mapping to proper UniRef protein families is over 90% with our quantized models, but additionally, with this service, it is possible to obtain the top 100 hits and still perform classical sequence alignments to further boost clustering accuracy. This PLM service now represents a foundational asset within the KBase Lakehouse, capable of supporting both large-scale curation tasks and user-level interactive searches.

Machine Learning Classifiers for Microbial Phenotype Prediction

A major challenge in microbial systems biology is predicting nutrient utilization phenotypes—whether a microbe can grow on a particular carbon source—directly from its genome. Traditional experimental approaches (e.g., Biolog plates, culture-based assays) are resource-intensive and biased toward culturable taxa, leaving many microbial capabilities uncharacterized. This activity is a critical step in predicting environmental distributions of microbes and the optimality of a host to serve a metabolic engineering need.

To address this, KBase has developed machine learning classifiers that predict microbial phenotypes from genomic features. These models are trained on an extensive dataset of 819 microbial genomes tested across 242 carbon sources, drawing on multiple experimental resources (ATLeaf, Biolog, Marine, and PMI datasets).

The classifiers integrate diverse genomic annotations, including RAST, KOFAM, UniProt, eggNOG, and UniRef, converting gene-level information into feature matrices for machine learning. Models were trained using Random Forest and CatBoost[9] algorithms within a standardized evaluation framework, with performance assessed by cross-validation, cross-dataset testing, and phylogeny-informed train/test splits.

Key outcomes:

- High intra-dataset accuracy: Classifiers achieved strong performance when trained and tested within the same dataset (balanced accuracy typically >0.8).
- Cross-dataset challenges: Generalization across datasets was more difficult due to phylogenetic and experimental biases, though biologically informed feature selection improved robustness.
- Explainable predictions: Importantly, the models identify key genomic features (e.g., KEGG pathway genes, transporters, regulators) driving phenotype predictions, providing mechanistic insights into nutrient utilization pathways.
- Accuracy: Depending on the phenotype, classifiers achieved 80–95% accuracy, making them among the most reliable genome-to-phenotype models to date.

Together, these phenotype classifiers advance KBase's mission to move from genome sequence to functional prediction at scale. Beyond practical utility (e.g., predicting traits of MAGs or isolates), they offer scientific discovery potential by revealing which genetic features are most predictive of microbial traits across environments.

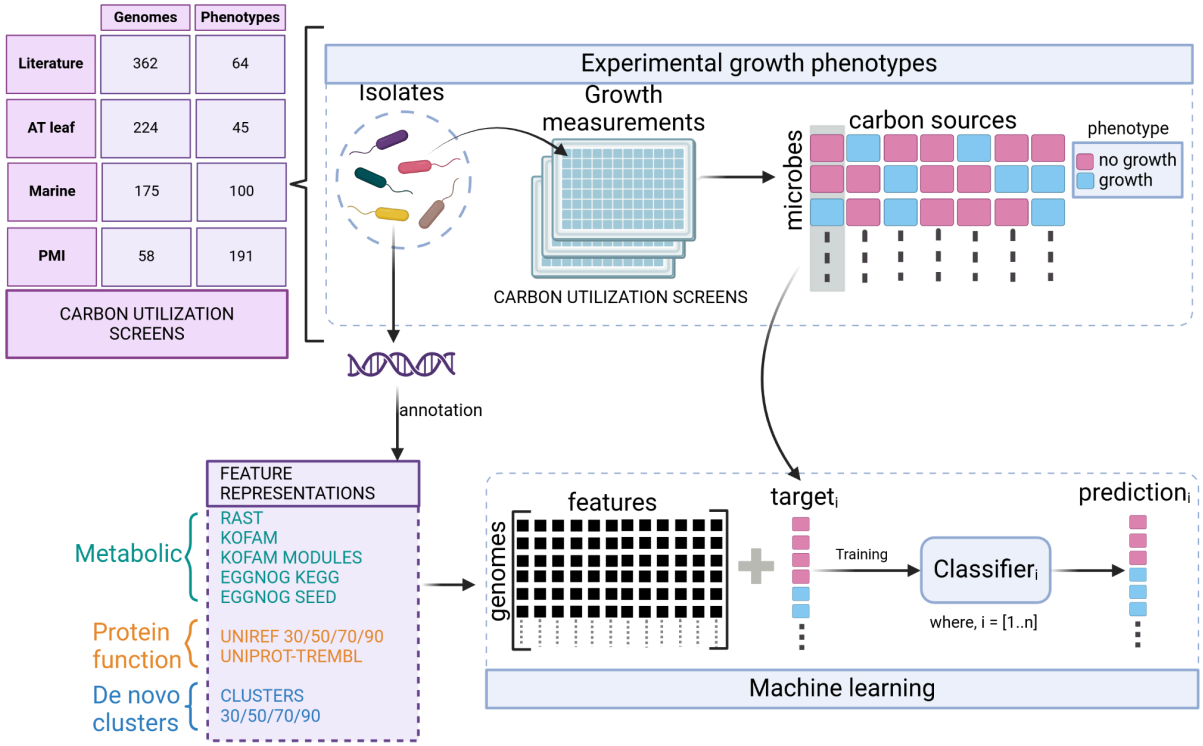

Figure 2: The microbial carbon utilization prediction workflow. (A) Overview of the data, characteristics, and machine learning workflow for carbon utilization prediction. The experimental data were collected by measuring the growth phenotypes (growth vs no growth) of a group of microbial isolates in 242 unique carbon sources. The resulting growth/no-growth outcomes are tabulated and utilized as the training data for the machine learning workflow. Our study utilizes four datasets (Literature, ATLeaf, Marine, and PMI). The analysis workflow first includes annotating the genome sequences of the microbial isolates using various annotation tools. The count matrices generated are then used as features for the machine learning workflow to predict the outcomes of the carbon utilization screens.

Building Toward FY26 Priorities

The AI/ML capabilities developed in FY25 represent first-of-kind prototypes that directly embed machine reasoning into KBase. Each effort lays the groundwork for future scaling:

- Narrative Agent (Research Assistant): expanding from guided workflows to fully interactive, human-in-the-loop research companions that integrate documentation, analysis, and interpretation.

- Gene Function Agent: scaling from curated subsets to millions of genes, combining experimental data (e.g., RbTnSeq) with literature evidence for high-confidence annotations.
- Phenotype Classifiers: extending predictive models to multi-omics data integration, enabling more accurate and interpretable genotype-to-phenotype predictions across diverse ecosystems.
- Protein Language Model: becoming the foundation indexing layer for rapid, scalable search and clustering across DOE's expanding protein databases.
- Generative Genome Model: advancing from missing gene recovery to complete genome design and evaluation, integrating with KBase's Lakehouse for contamination detection, genome finishing, and ultimately synthetic genome prototyping.

Together, these efforts move KBase from being a platform for data interoperability and analysis to becoming a knowledge environment powered by agentic AI. FY26 will focus on transitioning these prototypes into production-ready, integrated capabilities that span data ingestion, annotation, model-building, and interactive exploration.

Conclusion

FY25 Q4 marks a strategic inflection point for KBase. After three quarters of building infrastructure through interoperability (Q1), collaboration (Q2), and partnerships (Q3), KBase has now taken the first steps in embedding AI/ML capabilities at the core of the user experience and scientific workflows.

With agents that guide analyses, models that predict gene functions and phenotypes, PLMs that transform sequence search, and generative models that can reconstruct or even design microbial genomes, KBase is beginning to redefine how scientists interact with data. These advances shift KBase from a toolbox into a co-scientist platform that can reason, explain, and generate.

Most importantly, this work is being built on top of the KBase Data Lakehouse, which enables AI-ready data integration and serves as a model for a larger cross-BER integrated data lakehouse. By augmenting the lakehouse with an agentic AI infrastructure, the KBase AI/ML team will help define a reference implementation of a scalable, extensible, explainable AI architecture to accelerate BER science.

Looking ahead, KBase will continue to pioneer AI-driven predictive biology in support of DOE BER's mission — enabling researchers not just to analyze data, but to predict, design, and engineer biological systems that address the nation's energy and environmental challenges. This includes working closely with projects, such as OPAL, for demonstrating an agentic orchestration layer across BER automated labs and with BERIL for developing agent-ready APIs to accelerate data acquisition, harmonization and analysis to enable development of scientist agents. This positions KBase as a key platform and testbed for cross-BER AI initiatives and ensures data, models, and laboratory experiments can be integrated in a common infrastructure.

References

1. Lewis, P. *et al.* Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. (2020).
2. GitHub - langchain-ai/langgraph: Build resilient language agents as graphs. *GitHub* https://github.com/langchain-ai/langgraph.
3. LangChain. https://www.langchain.com/.
4. Price, M. N. *et al.* Mutant phenotypes for thousands of bacterial genes of unknown function. *Nature* **557**, 503–509 (2018).
5. Website. https://fit.genomics.lbl.gov/cgi-bin/myFrontPage.cgi.
6. Price, M. N. & Arkin, A. P. PaperBLAST: Text Mining Papers for Information about Homologs. *mSystems* **2**, (2017).
7. Chaumeil, P.-A., Mussig, A. J., Hugenholtz, P. & Parks, D. H. GTDB-Tk v2: memory friendly classification with the genome taxonomy database. *Bioinformatics* **38**, 5315–5316 (2022).
8. Kesapragada, M. *et al.* Generative model for Pseudomonad genomes. in *NeurIPS 2022 Workshop on Learning Meaningful Representations of Life* (2022).
9. CatBoost - state-of-the-art open-source gradient boosting library with categorical features support. https://catboost.ai/.