Abstract
Preliminary: This final report corresponds to a grant (DE-SC0021216) that was awarded to the University of Montana. Mid-way through the grant period, I relocated from the University of Montana to the University of Arizona. The grant was ended at University of Montana in late 2022, with all efforts concluding on 08/26/22; the remaining funds supporting the project were relinquished by University of Montana, and were later awarded to University of Arizona under a new grant, with start date 04/01/23. This report focuses on results of research efforts at UMontana through 08/26/22.
Results: We made progress in each of the three aims of the proposal. We released software that identifies and fills gaps in the annotation of metabolic proteins within bacterial genomes. We made substantial progress in developing software for alignment-based annotation of protein coding DNA, allowing for coding frameshifts caused by sequencing error. Finally, we made notable progress in developing AI methods (specifically: a neural embedding model) for identifying similarities between protein sequences based on amino-wise latent vectors. These efforts were supplemented by development of methods for protein modeling in support of predicting protein-drug binding activity, and by my leadership of a team in the NIH/DOE 2021 Petabyte-Scale Sequence Search code-a-thon.

Research findings
Aim 1 (A Bayesian-inference framework for integrating multi-omics data): Individual proteins need not be examined and labeled in isolation – their involvement in metabolic, regulatory, and signaling networks can provide additional support for annotation of their presence and function. Consider a gene network found in an individual: one of more present genes may escape annotation due to some combination of sequence divergence, sequencing error, or true novelty. We developed an ensembled deep neural network to predict the presence of a (KEGG) metabolic module even when annotation support for that module is incomplete, for example due to incomplete sequencing/annotation of the constituent proteins. The tool, MetaPathPredict, was released shortly after the end of the funding period, and makes robust predictions of KEGG module presence within highly incomplete genomes; it also produces confidence estimates that correlate with accuracy. A pre-print is available at https://doi.org/10.1101/2022.12.21.521254, and the manuscript is in its second round of review at eLife. The software, MetaPathPredict, can be downloaded from https://github.com/d-mcgrath/MetaPathPredict.

During the course of our work on metagenomes, we became increasingly interested in the presence of phage in these genomes. Shortly after the funding period, we released a tool, VIBES, that annotates full or partial integrations of phage sequence detected in bacterial genomes. VIBES can be downloaded from https://github.com/TravisWheelerLab/VIBES. A pre-print of the manuscript (now in press at Nucleic Acids Research) can be found at https://doi.org/10.1101/2023.10.17.562434.

Aim 2 (Accounting for sequencing error in alignment-based annotation): During the project period, we made progress toward development of BATH, a tool for the annotation of protein-

coding DNA via translated alignment. Built on the HMMER3 code base, BATH annotates DNA sequences based on direct alignment to protein sequences or profile hidden Markov models (pHMMs). BATH also employs novel frameshift-aware algorithms to detect frameshift-inducing indels and produce accurate annotations of frameshifted DNA. On benchmarks involving protein-coding sequence with or without frameshift-inducing sequencing error, BATH demonstrates substantially better sensitivity than other translated search tools (tblastn, LAST, MMseqs2, DIAMOND), and in particular produces greater annotation coverage and accurate E-values. BATH was released shortly after the funding period, and is available for download at https://github.com/TravisWheelerLab/BATH; the descriptive manuscript is in review at Bioinformatics Advances (preprint: https://doi.org/10.1101/2023.12.31.573773).

Aim 3 (Deep learning approaches for alignment-free annotation): During the funding period, we made progress towards development of a novel neural embedding model designed to compute amino-wise latent vectors for protein sequences. The model, which we call NEAR, was trained on sequence alignments generated by the widely used phmmer tool (in HMMER), and produces per-amino-acid embeddings that support rapid identification of similar proteins. NEAR was released after the funding period; benchmarking experiments reveal improved performance and enhanced computational efficiency of NEAR compared to state-of-the-art neural embedding models specifically developed for protein sequences. On commodity GPU hardware ($300 card available for installation on a mid-level laptop or desktop), our neural network model finds nearly all matches identified by HMMER with ~100x speedup. The software implementation is called NEAR, and is available in a github repository (https://github.com/TravisWheelerLab/NEAR), with preprint available at https://doi.org/10.1101/2024.01.25.577287 (in review at Bioinformatics Advances).

Additional: Early in the SARS-Cov2 pandemic, we were approached by a researcher at the NIH Rocky Mountain Labs to collaborate on development of a drug discovery pipeline. Motivated by the situational urgency of the pandemic and the conceptual/methodological overlap between representation of sequences at the sequence and structural levels, we joined the search for novel drugs. Among other advances, our neural network architecture demonstrated an AUC of ~0.85 for classification of active vs. decoy compounds in the PDBbind database, a substantial increase over the state of the art. We used this model as part of our new pipeline (https://drugsniffer.org/, https://doi.org/10.3389/fphar.2022.874746) for scanning through billions of candidate ligands to find several thousand with high predicted binding affinity to target proteins, as part of a large community effort to improve prediction of novel SARS-CoV2 drugs (https://doi.org/10.1002/minf.202300262).

I led one of four groups within a 2021 NIH/DOE-organized code-a-thon on benchmark development for peta-scale sequence search, emphasizing sensitive search at peta-scale. Among other things, this lead to development of a new analysis of the risks of using reversed sequences as a source of benchmark decoy sequences (in press at Bioinformatics Advances, preprint https://doi.org/10.1101/2023.06.19.545636) and a piloted benchmark harness ()https://github.com/TravisWheelerLab/bagel).