STANDARDS, TOOLING AND BENCHMARKS TO PROBE REPRESENTATION LEARNING ON PROTEINS

Joaquin Gomez Sanchez
Technical University of Munich
joaquin.gomez@tum.de

Sebastian Franz
Technical University of Munich
sebastian.franz@tum.de

Michael Heinzinger
Technical University of Munich
mheinzinger@rostlab.org

Burkhard Rost
Technical University of Munich
rost@in.tum.de

Christian Dallago
Technical University of Munich
christian.dallago@tum.de

Summary

With the advent of novel foundational approaches to represent proteins, a race to evaluate and assess their effectiveness to embed biological data for a variety of downstream tasks, from structure prediction to protein engineering [1], has gained tremendous traction. While tasks like protein 3D structure prediction from sequence have well characterized datasets and methodological approaches [2], many others, for instance probing the ability to encode protein function from sequence, lack standardization. This becomes particularly relevant when employing experimental biological datasets for machine learning, as curating biologically meaningful data splits requires biological intuition, whilst engineering appropriate machine learning models requires data science expertise. Gold standard experimental datasets annotated with machine learning relevant metadata are thus scarce and often scattered in different file formats in the literature, using a variety of metrics to measure success, hindering rapid evaluation of new foundational representation techniques or machine learning models built on top of them. To address these challenges, we propose a suite of solutions including a) standards for sequence datasets and embedding interfaces, b) curated and machine learning metadata annotated protein sequence datasets, c) machine learning architectures and training scripts, and d) an extensible, automatic evaluation pipeline connecting all these components. In practice, we described new, broad data standards for machine learning protein sequence datasets, including definitions for predictions of a categorical attribute for a residue in a sequence (e.g., secondary structure), or predicting a single value for the entire sequence (e.g., protein fitness). We expanded a previous collection of datasets for protein engineering (FLIP [1]) by adding five traditional tasks from the literature, like residue secondary structure [3, 4], residue conservation [5], and protein subcellular location prediction [6, 7, 8]. We created a novel software solution (biotrainer) that collects machine learning architectures used for protein predictions and exposes a reproducible training pipeline that can consume any dataset adhering to the newly proposed data standards. Lastly, we connected all components in a new software solution (autoeval), which collects definitions for embedding methods, datasets and downstream machine learning models to automatically evaluate them. With these solutions, biological experimentalists can contribute new datasets and even train standard models using popular embedding methods, while machine learning researchers can easily plug in new foundational models or architectures in a common interface and test them on a variety of tasks against other solutions. In turn, the combination of solutions presented here unlocks the ability of interest groups to create challenges around new biological datasets, new machine learning architectures, new foundational models, or a combination thereof.

Availability: Biotrainer, the software solution containing machine learning architectures, reproducible training runs and deployable models, is available open source at https://github.com/sacdallago/biotrainer. FLIP v2, the collection of curated protein sequence sets annotated with machine learning-relevant attributes such as train, test and validation assignations, is available at https://github.com/J-SNACKKB/FLIP/. Autoeval, assessing the accuracy of various foundational protein representation models on protein tasks from FLIP v2 using architectures from biotrainer is available open source at https://github.com/J-SNACKKB/autoeval/.

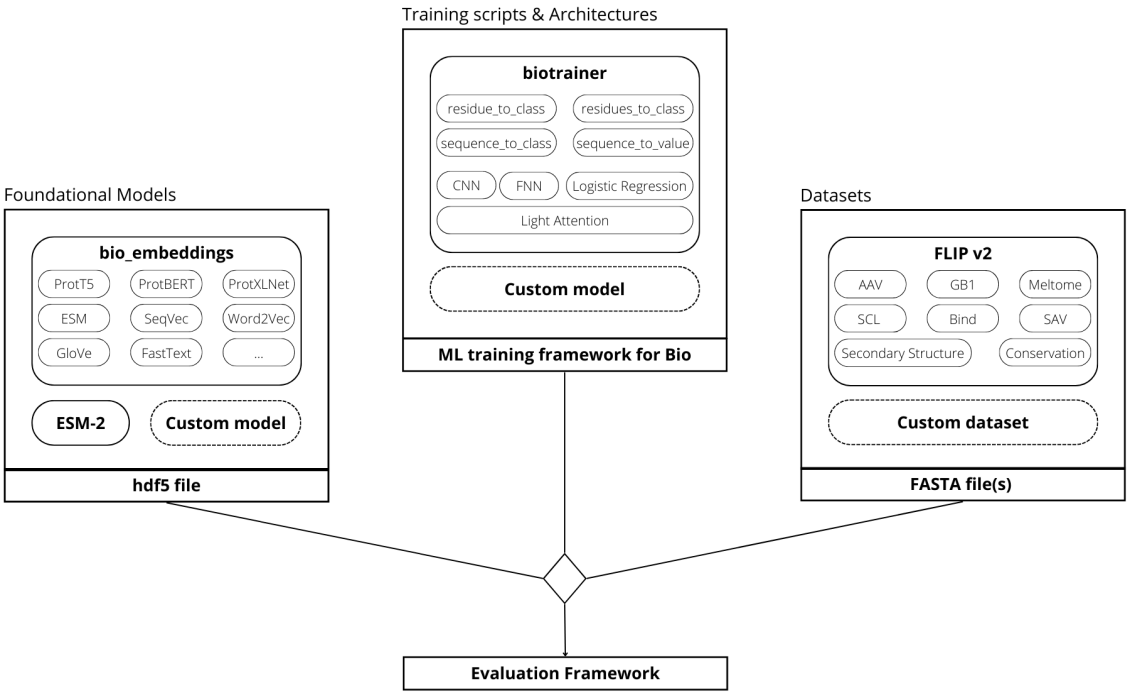

Figure 1: **From foundational models to datasets, training frameworks and automatic evaluations unlocking rapid advancement in machine learning for protein sequences.** *Foundational Models* included in frameworks like bio_embeddings [9] or custom scripts for ESM-2 [10] allow to embed protein sequence sets in computable formats storing the results in standardized hdf5 files. These files include sequences embedded either as matrices for residue prediction tasks (s.a. secondary structure), or as vectors for protein prediction tasks (s.a. subcellular localization). *Training scripts and Architectures* allow to train machine learning models to predict attributes of proteins using annotated FASTA files and hdf5 embedding files. In particular, our novel solution biotrainer collects several architectures from the literature (s.a. Light Attention [8]) intro reproducible workflows enabling the most common protein prediction tasks, for instance predicting categorical attributes of residues (residue_to_class) or sequences (sequence_to_class). *Datasets* stored as FASTA files with novel standards for curated sequence datasets with annotations relevant for machine learning pipelines. In particular, we expand FLIP [1] with several protein sequence machine learning datasets (subcellular localization prediction (SCL) [8], secondary structure prediction [11], single amino acid variant (SAV) effect prediction [5], and binding ability to metal, nucleic acids and small molecules (Bind) [12]), annotated with split information (train, test and validation), and easy to parse labels. *Evaluation Framework* uses stored embeddings of annotated FASTA files are linked to pre-defined machine learning training scripts to automatically evaluate the ability of foundational models and machine learning architectures to predict a plethora of protein properties.

References

[1] Christian Dallago, Jody Mou, Kadina E Johnston, Bruce J Wittmann, Nicholas Bhattacharya, Samuel Goldman, Ali Madani, and Kevin K Yang. Flip: Benchmark tasks in fitness landscape inference for proteins. *bioRxiv*, 2021.

[2] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021.

[3] Michael Schantz Klausen, Martin Closter Jespersen, Henrik Nielsen, Kamilla Kjaergaard Jensen, Vanessa Isabell Jurtz, Casper Kaae Soenderby, Morten Otto Alexander Sommer, Ole Winther, Morten Nielsen, Bent Petersen, et al. Netsurfp-2.0: Improved prediction of protein structural features by integrated deep learning. *Proteins: Structure, Function, and Bioinformatics*, 87(6):520–527, 2019.

[4] John Moult, Krzysztof Fidelis, Andriy Kryshtafovych, Torsten Schwede, and Anna Tramontano. Critical assessment of methods of protein structure prediction (casp)—round xii. *Proteins: Structure, Function, and Bioinformatics*, 86:7–15, 2018.

[5] Céline Marquet, Michael Heinzinger, Tobias Olenyi, Christian Dallago, Kyra Erckert, Michael Bernhofer, Dmitrii Nechaev, and Burkhard Rost. Embeddings from protein language models predict conservation and variant effects. *Human genetics*, pages 1–19, 2021.

[6] José Juan Almagro Armenteros, Casper Kaae Sønderby, Søren Kaae Sønderby, Henrik Nielsen, and Ole Winther. Deeploc: prediction of protein subcellular localization using deep learning. *Bioinformatics*, 33(21):3387–3395, 2017.

[7] Vineet Thumuluri, José Juan Almagro Armenteros, Alexander Rosenberg Johansen, Henrik Nielsen, and Ole Winther. Deeploc 2.0: multi-label subcellular localization prediction using protein language models. *Nucleic Acids Research*, 2022.

[8] Hannes Stärk, Christian Dallago, Michael Heinzinger, and Burkhard Rost. Light attention predicts protein location from the language of life. *Bioinformatics Advances*, 1(1):vbab035, 2021.

[9] Christian Dallago, Konstantin Schütze, Michael Heinzinger, Tobias Olenyi, Maria Littmann, Amy X Lu, Kevin K Yang, Seonwoo Min, Sungroh Yoon, James T Morton, et al. Learned embeddings from deep learning to visualize and predict protein sets. *Current Protocols*, 1(5):e113, 2021.

[10] Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Sal Candido, and Alexander Rives. Language models of protein sequences at the scale of evolution enable accurate structure prediction. *bioRxiv*, 2022.

[11] Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, Ghalia Rehawi, Yu Wang, Llion Jones, Tom Gibbs, Tamas Feher, Christoph Angerer, Martin Steinegger, Debsindhu Bhowmik, and Burkhard Rost. Prottrans: Toward understanding the language of life through self-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):7112–7127, 2022.

[12] Maria Littmann, Michael Heinzinger, Christian Dallago, Konstantin Weissenow, and Burkhard Rost. Protein embeddings and deep learning predict binding residues for various ligand classes. *Scientific Reports*, 11(1):1–15, 2021.