Streamlining value of protein embeddings through bio_embeddings

Christian Dallago[1,2,#,*], Konstantin Schütze[1,#], Michael Heinzinger[1,2,#], Tobias Olenyi[1], Maria Littmann[1,2], Amy X. Lu[3], Kevin K. Yang[4], Seonwoo Min[5], Sungroh Yoon[5,6], Burkhard Rost[1,7]

[1] TUM (Technical University of Munich) Department of Informatics, Bioinformatics & Computational Biology - i12, Boltzmannstr. 3, 85748 Garching/Munich, Germany

[2] TUM Graduate School, Center of Doctoral Studies in Informatics and its Applications (CeDoSIA), Boltzmannstr. 11, 85748 Garching, Germany

[3] University of Toronto, Department of Computer Science & Vector Institute

[4] Microsoft Research New England, Cambridge, MA, 02142

[5] Department of Electrical and Computer engineering, Seoul National University, Seoul 08826, South Korea

[6] Interdisciplinary Program in Bioinformatics, Seoul National University, Seoul 08826, South Korea

[7] Institute for Advanced Study (TUM-IAS), Lichtenbergstr. 2a, 85748 Garching/Munich, Germany & TUM School of Life Sciences Weihenstephan (WZW), Alte Akademie 8, Freising, Germany & Columbia University, Department of Biochemistry and Molecular Biophysics & New York Consortium on Membrane Protein Structure (NYCOMPS), 701 West, 168th Street, New York, NY 10032, USA

# these authors contributed equally to this work
* Corresponding author: christian.dallago@tum.de, http://www.rostlab.org
For correspondence to all authors: support@bioembeddings.com
Tel: +49-289-17-811 (email Rost: assistant@rostlab.org)

Keywords: protein representations; hypothesis testing; feature extraction; annotation visualization; language models; protein annotation pipeline

Summary

Protein-based machine learning models increasingly contribute toward the guidance of experimental decision making. Models capable of quickly classifying entire biomes could help focus experiments on promising novelty. Recently, Language Models (LMs) have been adapted from Natural Language Processing (NLP) to decoding the language of life found in protein sequences. Such protein LMs show enormous potential in generating descriptive representations for proteins from just their sequences at a fraction of the time compared to previous approaches. LMs convert amino acid sequences into embeddings (vector representations) useful for several downstream tasks, including analysis and the prediction of aspects of protein function and structure. A buzzing variety of protein LMs is being generated worldwide that, in its diversity, is likely to shine light on different angles of the protein language. Unfortunately, these resources are scattered over the web. The *bio_embeddings* pipeline offers a unified interface to protein LMs to simply and quickly *embed* large protein sets, to *project* the embeddings in lower dimensional spaces, to *visualize* proteins on interactive scatter plots, and to *extract* annotations through supervised or unsupervised techniques. This enables quick hypothesis generation and testing. The pipeline is accompanied by a web server that offers to *embed*, *project*, *visualize,* and *extract* annotations for small protein datasets directly online, without the need to install software.

Availability

- Installation instructions, examples, notebooks and source code are available at: https://github.com/sacdallago/bio_embeddings
- The bio embeddings API is available at: https://api.bioembeddings.com

EXTENDED ABSTRACT

Language Models (LMs) like ELMo (Peters et al., 2018), BERT (Devlin et al., 2019) and GPT-3 (Brown et al., 2020) improve over previous "generations" by training on increasingly larger natural language corpora. They begin to suggest large models from artificial intelligence (AI) or machine learning (ML) to compete with human experts, at least for some tasks (Manning, 2011). These representations effectively decode the meaning of natural language from large amounts of unlabelled data. This is exactly what makes these approaches so promising for proteins for which the amount of unlabelled data (protein sequences) is outgrowing the amount of labelled data (known protein function and structure) by many orders of magnitude. In fact, the amount of data available for protein sequences even outgrows the largest NLP data sets such as Google's Billion Word data (Chelba et al., 2014) about 500-fold (Elnaggar et al., 2020).

Protein sequences follow similar principles to natural languages: they are formed by tokens (amino acids) that have individual and context dependent meaning, through long- and short-range dependencies such as residue-residue interactions. Thus, similarly to natural language, LMs trained on protein sequences (Alley et al., 2019; AlQuraishi, 2019; Armenteros et al., 2020; Elnaggar et al., 2020; Heinzinger et al., 2019; Lu et al., 2020; Madani et al., 2020; Min et al., 2020; Rao et al., 2019; Rives et al., 2019) capture important *meaning* of the protein sequence language. Proof for such understanding of meaning is that the models can be used for the acid test of predicting aspects of protein structure and function. For instance, SeqVec (Heinzinger et al., 2019) training ELMo (Peters et al., 2018) on UniRef50 (The UniProt Consortium, 2019), showed that the LM's representations clustered protein sequences by function (Heinzinger et al., 2019). In another analogy to NLP, protein LMs may be fine-tuned on specialized sequence sets (analogy to natural language: legal text vs. wikipedia articles) to encode for different protein properties (Armenteros et al., 2020).

The *bio_embeddings* pipeline is targeted to computational biologists and aims to abstract, via a uniform and standardized interface, the use of protein LMs to create protein representations (embeddings). These representations can be used to train machine learning algorithms using "transfer learning" (Raina et al., 2007), or for analytical purposes. Features of the pipeline allow to perform visual analysis of custom sequence sets by drawing protein spaces spawned by their embeddings. The pipeline also incorporates supervised and unsupervised (Littmann et al., 2020; Villegas-Morcillo et al., 2020) approaches on protein embeddings to further enhance analytical potential out of the box. Users of the pipeline can choose to create representations using many protein LMs, including SeqVec (Heinzinger et al., 2019), UniRep (Alley et al., 2019), ESM (Rives et al., 2019), ProtBert-BFD, ProtAlbert, ProtXLNet (Elnaggar et al., 2020). Pipeline runs are reproducible (as configurations are defined in files), and outputs are stored in popular formats, e.g. CSVs, FASTA and HDF5 (The HDF Group, 2000). For researchers contributing new protein LMs, *bio_embeddings* can become a central repository to distribute their work to the community, requiring minimal changes for pipeline consumers to make use of new protein LMs.

To demonstrate the use of the *bio_embeddings* pipeline, we created several example runs. Reported here are outputs for two of them (Figure 1). We ran these examples on a machine equipped with an Nvidia GTX1080 by using SeqVec (Heinzinger et al., 2019) and UMAP (McInnes et al., 2018). Inputs, outputs, and execution steps are outlined in the repository linked in **AVAILABILITY**. For the "disprot" example (1161 protein sequences, 623902 amino acids, DisProt (Hatos et al., 2020)), it took 10 minutes to go from FASTA sequences to interactive plot. The ~7.8x bigger "cath" set (31288 protein sequences, 4872078 amino acids, CATH (Dawson et al., 2017)) also took ~10 minutes to go from FASTA file to annotated plot.

Figure 1 shows that raw SeqVec representations have the potential to distinguish highly disordered proteins (>80% disorder content) from less disordered proteins (<20% disorder content). In an experimental setting where low disorder proteins are the desiderata, large protein sequences libraries can be screened and assessed virtually by plotting embeddings of library sequences against the annotated proteins in DisProt (Hatos et al., 2020). For this purpose, users may also use the unsupervised *extract* stage, which leverages embedding similarity to transfer annotations (Littmann et al., 2020). Filtering by a desiderata in-silico may narrow the amount of in-vitro or in-vivo experiments, potentially accelerating development while decreasing cost.

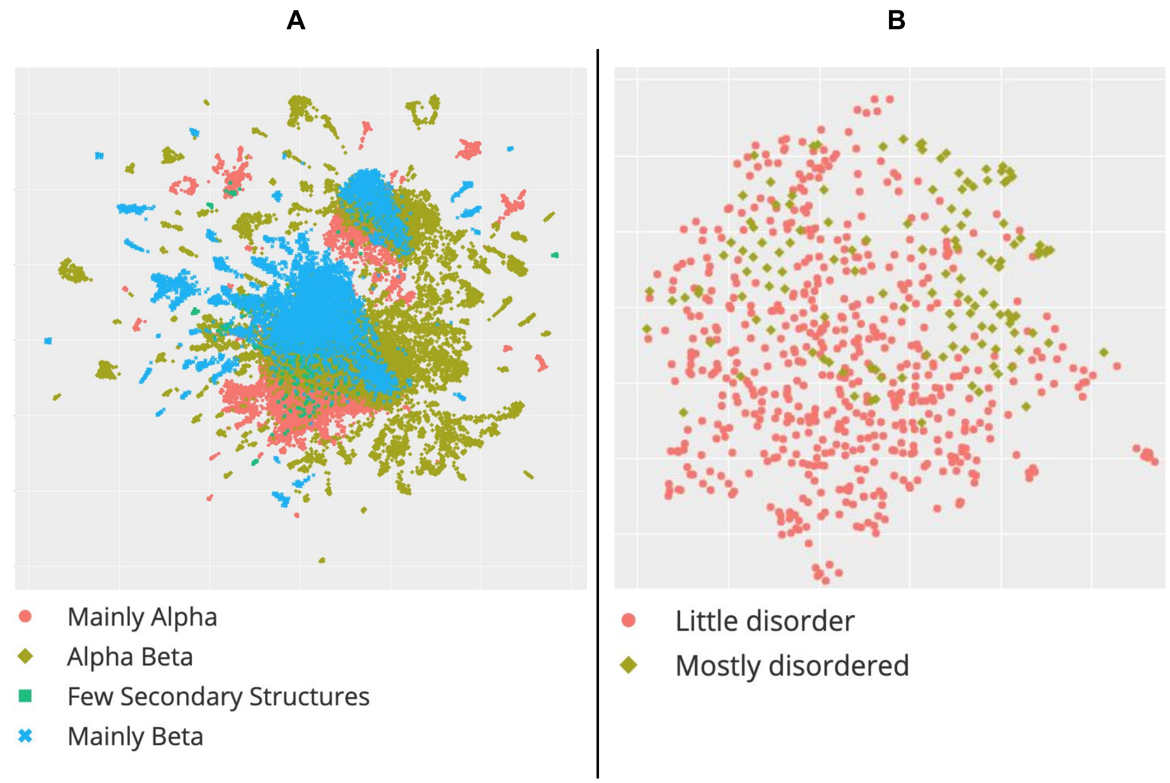

Figure 1 – Projections of per protein embeddings (via SeqVec) using UMAP are able to separate proteins into clusters

Screenshots of interactive plotly visualizations of UMAP (McInnes et al., 2018) reduced per protein embeddings using SeqVec (Heinzinger et al., 2019) for proteins in CATH (**A** - (Dawson et al., 2017)) and DisProt (**B** - (Hatos et al., 2020))). The plots were generated by providing (i) a FASTA file (ii) an annotation file in CSV format, and (iii) a configuration file. Protein annotations for CATH ("Mainly Alpha", "Mainly Beta", "Alpha Beta", "Few Secondary Structures") were parsed from the database. In the case of DisProt, proteins were classified as "Little disorder" if the amount of amino acids annotated as disorder was inferior to 20% of the total amino acidic content for a protein, and as "Mostly disordered" if the amount of disordered amino acids was above 80%. These files are provided for reproducibility in the repository (see **AVAILABILITY**). Interactive visualizations can be accessed here: CATH http://data.bioembeddings.com/public/embeddings/examples/cath.html, DisProt http://data.bioembeddings.com/public/embeddings/examples/disprot.html

ACKNOWLEDGEMENT

Thanks to Tim Karl (TUM) for help with hardware and software; to Inga Weise (TUM) for support with many other aspects of this work. Thanks to Jamie Morton, Tom Sercu, Ali Madani, Daniel Barenberg, Alex Rives, Vladimir Gligorijevic and Josh Meier for constructive discussions around protein language models and their use. Thanks to Roshan Rao, Niel Thomas and Nicholas Bhattacharya for creating and maintaining TAPE. Thanks to all those who deposit their experimental data in public databases, and to those who maintain these databases. In particular, thanks to the Ioanis Xenarios (SIB, Univ. Lausanne), Matthias Uhlen (Univ. Upssala) and their teams at Swiss-Prot and HPA.

FUNDING

This work was supported by the Deutsche Forschungsgemeinschaft (DFG) – project number RO1320/4-1, by the Bundesministerium für Bildung und Forschung (BMBF) – project number 031L0168, and by the BMBF through the programme "Software Campus 2.0 (TU München)" – project number 01IS17049.

CONFLICT OF INTEREST

None declared.

REFERENCES

Alley, E. C., Khimulya, G., Biswas, S., AlQuraishi, M., & Church, G. M. (2019). Unified rational protein engineering with sequence-based deep representation learning. *Nature Methods*, *16*(12), 1315–1322. https://doi.org/10.1038/s41592-019-0598-1

AlQuraishi, M. (2019). End-to-End Differentiable Learning of Protein Structure. *Cell Systems*, *8*(4), 292-301.e3. https://doi.org/10.1016/j.cels.2019.03.006

Armenteros, J. J. A., Johansen, A. R., Winther, O., & Nielsen, H. (2020). Language modelling for biological sequences – curated datasets and baselines. *BioRxiv*, 2020.03.09.983585. https://doi.org/10.1101/2020.03.09.983585

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). Language Models are Few-Shot Learners. *ArXiv:2005.14165 [Cs]*. http://arxiv.org/abs/2005.14165

Chelba, C., Mikolov, T., Schuster, M., Ge, Q., Brants, T., Koehn, P., & Robinson, T. (2014). One Billion Word Benchmark for Measuring Progress in Statistical Language Modeling. *ArXiv:1312.3005 [Cs]*. http://arxiv.org/abs/1312.3005

Dawson, N. L., Lewis, T. E., Das, S., Lees, J. G., Lee, D., Ashford, P., Orengo, C. A., & Sillitoe, I. (2017). CATH: An expanded resource to predict protein function through structure and sequence. *Nucleic Acids Research*, *45*(D1), D289–D295. https://doi.org/10.1093/nar/gkw1098

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *ArXiv:1810.04805 [Cs]*. http://arxiv.org/abs/1810.04805

Elnaggar, A., Heinzinger, M., Dallago, C., Rehawi, G., Wang, Y., Jones, L., Gibbs, T., Feher, T., Angerer, C., Steinegger, M., Bhowmik, D., & Rost, B. (2020). ProtTrans: Towards Cracking the Language of Life's Code Through Self-Supervised Deep Learning and High Performance Computing. *BioRxiv*, 2020.07.12.199554. https://doi.org/10.1101/2020.07.12.199554

Hatos, A., Hajdu-Soltész, B., Monzon, A. M., Palopoli, N., Álvarez, L., Aykac-Fas, B., Bassot, C., Benítez, G. I., Bevilacqua, M., Chasapi, A., Chemes, L., Davey, N. E., Davidović, R., Dunker, A. K., Elofsson, A., Gobeill, J., Foutel, N. S. G., Sudha, G., Guharoy, M., … Piovesan, D. (2020). DisProt: Intrinsic protein disorder annotation in 2020. *Nucleic Acids Research*, *48*(D1), D269–D276. https://doi.org/10.1093/nar/gkz975

Heinzinger, M., Elnaggar, A., Wang, Y., Dallago, C., Nechaev, D., Matthes, F., & Rost, B. (2019). Modeling aspects of the language of life through transfer-learning protein sequences. *BMC Bioinformatics*, *20*(1), 723. https://doi.org/10.1186/s12859-019-3220-8

Littmann, M., Heinzinger, M., Dallago, C., Olenyi, T., & Rost, B. (2020). Embeddings from deep learning transfer GO annotations beyond homology. *BioRxiv*, 2020.09.04.282814. https://doi.org/10.1101/2020.09.04.282814

Lu, A. X., Zhang, H., Ghassemi, M., & Moses, A. (2020). Self-Supervised Contrastive Learning of Protein Representations By Mutual Information Maximization. *BioRxiv*, 2020.09.04.283929. https://doi.org/10.1101/2020.09.04.283929

Madani, A., McCann, B., Naik, N., Keskar, N. S., Anand, N., Eguchi, R. R., Huang, P.-S., & Socher, R. (2020). ProGen: Language Modeling for Protein Generation. *BioRxiv*, 2020.03.07.982272. https://doi.org/10.1101/2020.03.07.982272

Manning, C. D. (2011). Part-of-Speech Tagging from 97% to 100%: Is It Time for Some Linguistics? In A. F. Gelbukh (Ed.), *Computational Linguistics and Intelligent Text Processing* (pp. 171–189). Springer. https://doi.org/10.1007/978-3-642-19400-9_14

Min, S., Park, S., Kim, S., Choi, H.-S., & Yoon, S. (2020). Pre-Training of Deep Bidirectional Protein Sequence Representations with Structural Information. *ArXiv:1912.05625 [Cs, q-Bio, Stat]*. http://arxiv.org/abs/1912.05625

Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep Contextualized Word Representations. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 2227–2237. https://doi.org/10.18653/v1/N18-1202

Raina, R., Battle, A., Lee, H., Packer, B., & Ng, A. Y. (2007). Self-taught learning: Transfer learning from unlabeled data. *Proceedings of the 24th International Conference on Machine Learning*, 759–766. https://doi.org/10.1145/1273496.1273592

Rao, R., Bhattacharya, N., Thomas, N., Duan, Y., Chen, P., Canny, J., Abbeel, P., & Song, Y. (2019). Evaluating Protein Transfer Learning with TAPE. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d\textquotesingle Alché-Buc, E. Fox, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems 32* (pp. 9689–9701). Curran Associates, Inc. http://papers.nips.cc/paper/9163-evaluating-protein-transfer-learning-with-tape.pdf

Rives, A., Goyal, S., Meier, J., Guo, D., Ott, M., Zitnick, C. L., Ma, J., & Fergus, R. (2019). Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *BioRxiv*, 622803. https://doi.org/10.1101/622803

The HDF Group. (2000, 2010). *Hierarchical data format version 5*. http://www.hdfgroup.org/HDF5

The UniProt Consortium. (2019). UniProt: A worldwide hub of protein knowledge. *Nucleic Acids Research*, *47*(D1), D506–D515. https://doi.org/10.1093/nar/gky1049

Villegas-Morcillo, A., Makrodimitris, S., van Ham, R. C. H. J., Gomez, A. M., Sanchez, V., & Reinders, M. J. T. (2020). Unsupervised protein embeddings outperform hand-crafted sequence and structure features at predicting molecular function. *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btaa701