BIOGRAPHICAL SKETCH

Provide the following information for the Senior/key personnel and other significant contributors.
Follow this format for each person. **DO NOT EXCEED FIVE PAGES.**

NAME: Tristan Bepler

eRA COMMONS USER NAME (credential, e.g., agency login): tbepler

POSITION TITLE: Postdoctoral Associate

EDUCATION/TRAINING *(Begin with baccalaureate or other initial professional education, such as nursing, include postdoctoral training and residency training if applicable. Add/delete rows as necessary.)*

INSTITUTION AND LOCATION	DEGREE *(if applicable)*	Completion Date MM/YYYY	FIELD OF STUDY
Duke University	B.S.	05/2014	Computer Science, Biology
Massachusetts Institute of Technology	Ph.D.	02/2020	Computational and Systems Biology
Massachusetts Institute of Technology	Postdoctoral	Ongoing	Computational Biology

A. Personal Statement

I have significant expertise in machine learning for protein biology and a broad background in computational biology, and sequence and structure analysis. I have published multiple first author papers in the top machine learning venues and biology journals. As a graduate student, I developed fundamental machine learning methods for better understanding proteins using sequence and structure information, where I proposed the first method for learning contextualized protein sequence embeddings. This work has since attracted significant interest due to its implications for protein homology search and protein engineering. The proposed work will build on this innovation to power short peptide therapeutic design. Furthermore, I have developed fundamental machine learning methods for understanding protein flexibility using cryo-electron microscopy as well as widely-used tools for automating the cryoEM analysis pipeline. Through my cryoEM work, I have established deep collaborations with technology developers and structural biologists and have taught me the importance of diverse expertise in collaborations and how to collaborate effectively.

1. Bepler, T. & Berger, B. (2019). Learning protein sequence embeddings using information from structure. International Conference on Learning Representations. Available from: https://openreview.net/pdf?id=SygLehCqtm
2. Bepler, T., Morin, A., Rapp, M., Brasch, J., Shapiro, L., Noble, A.J., Berger, B. (2019) Positive-unlabeled convolutional neural networks for particle picking in cryo-electron micrographs. *Nature Methods*, 16, 1153-1160.
3. Bepler, T., Zhong, E.D., Kelley, K., Brignole, E., Berger, B. (2019) Explicitly disentangling image content from translation and rotation with spatial-VAE. *Advances in Neural Information Processing Systems*. Available from: https://papers.nips.cc/paper/9677-explicitly-disentangling-image-content-from-translation-and-rotation-with-spatial-vae

B. Positions and Honors

Positions and Employment

2014-2020	Graduate Student, Computational and Systems Biology Program, Massachusetts Institute of Technology, Cambridge, MA
2017	Machine Learning Internship, Indigo Agriculture, Boston, MA
2018	Consultant, Flagship Pioneering VL56 (now Generate Biomedicines), Cambridge, MA
2019	Consultant, Indigo Agriculture, Boston, MA
2020-present	Consultant, Arbor Biotechnologies, Cambridge, MA
2020-present	Postdoctoral Associate, Research Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge, MA

Other Experience and Professional Memberships

2016-present	Member, International Society for Computational Biology

Honors

2014	Graduation with Distinction in Biology, Duke University, Durham, NC

C. Contributions to Science

1. A major thrust of my work is in developing fundamental machine learning (ML) tools for understanding protein sequence, structure, and function. The core problem that my work addresses is to leverage massive sequence datasets together with general protein knowledge contained in protein structure databases to learn better machine representations of proteins. These representations can be used to transfer this knowledge directly to new problems, by using them as features for specialized downstream ML models, and can also be used to search for structural homologues using only sequences. Historically, these problems have been solved by direct sequence homology and alignment algorithms, e.g., Needleman-Wunsch, BLAST, HMMER, HHalign. We developed the first contextual embedding method that incorporates sequence and structural information to generate protein sequence embeddings. We found that these embeddings enable rapid database search for structural homologues, from sequence alone, more accurately than prior sequence search methods. Remarkably, search with our embeddings was also more accurate than structure search. Applied to downstream problems, these embeddings improve performance on a host of protein function prediction problems. I conceived of the project, designed the model, designed and carried out the experiments, and wrote the paper on which I am first author. We have since applied these embeddings to develop specialized models for predicting ligand binding sites on proteins at residue resolution, a first author manuscript for which is in preparation. Now, I am developing active learning methods for protein engineering to improve and accelerate protein design. By leveraging protein embeddings and data-efficient models we can rapidly learn protein function predictors with little training data and accelerate the design process. Furthermore, we can incorporate prior knowledge through arbitrary black-box priors into this ML design procedure, which allows us to benefit from large existing datasets and also physics-, biochemical-, and simulation-based priors. I conceived, developed, and executed this project, which has been submitted to Neural Information Processing Systems (NeurIPS) 2020.
    a. Bepler, T. & Berger, B. (2019). Learning protein sequence embeddings using information from structure. *7th International Conference on Learning Representations*. Available from: https://openreview.net/pdf?id=SygLehCqtm
2. In addition to my machine learning efforts for large-scale data integration and protein engineering above, I have also made a number of contributions to better understanding protein structure using cryo-electron microscopy. In particular, I have developed fundamental ML algorithms for semi-supervised object detection, which we applied to better particle picking in cryoEM, and also for disentangled feature learning and generative modeling of protein structures with flexibility (i.e., continuous heterogeneity). Particle picking, a fundamental bottleneck in automation of the cryoEM analysis pipeline, is an object detection problem in which the many heterogeneous and low signal-to-noise ratio protein projections contained in cryo-electron micrographs must be detected and extracted. Historically, this procedure was either carried out by hand, with researchers hand labeling 10s-100s of thousands of particles, or using *ad hoc* image filters. At the same time, the computer vision and deep learning communities have made large strides in object detection in natural images. However, these innovations rely on enormous annotated datasets for the object classes of interest. Unfortunately, no such large, well-annotated datasets existed for cryoEM and the particle picking

problem is confounded by the fact that new proteins of interest often look dissimilar from those found in public repositories. This motivated us to approach the particle picking problem as a small data object detection problem. Specifically, we posed it as a positive-unlabeled learning problem (a special case of semi-supervised learning) where we are given a large number of unlabeled micrographs with a small number of labeled particles, but not labeled background, and we seek to learn a classifier that can distinguish particles from background and structured noise. We proposed a novel positive-unlabeled learning algorithm, which outperforms prior work in the field, and enabled unprecedented particle picking accuracy with minimal provided labeled data. I conceived of the problem formulation and solution, developed the software, designed and conducted the experiments, and wrote the paper on which I am first author. The particle picking software which I developed, Topaz, has been widely adopted by the cryoEM community. I have since extended this with micrograph denoising based on learned deep denoising models. I am first author on a manuscript for this that is under revision at *Nature Communications*. The heterogeneity problem is that we observe a large number of 2D protein projections and want to infer the continuous distribution over 3D structures from which these projections arise. Conventional reconstruction methods assume that the 2D projections arise from a single static 3D structure or a discrete mixture of static 3D structures. However, the 3D structure of each protein molecule is not actually static. To address this, I developed a fundamental machine learning algorithm for learning generative models of signals with continuous variability that enables efficient inference of the rotation and translation of the individual objects. The key insight is that the image generative function can be written as a function that maps spatial coordinates to pixel values (electron densities) at those positions in space. By formulating a generative model this way, transformations of the coordinate space (e.g., rotation and translation) are naturally modeled. I conceived this fundamental idea and developed the model framework. We first applied this to the problem of learning continuous distributions of 2D protein projections, for which I designed and conducted the experiments and wrote the paper on which I am first author. This was presented at NeurIPS 2019. My collaborators have since extended this framework into a complete 3D reconstruction framework. A paper on which I am second author was presented at the International Conference on Learning Representations (ICLR) 2020 and is under revision at *Nature Methods*.

a. Bepler, T., Morin, A., Rapp, M., Brasch, J., Shapiro, L., Noble, A.J., Berger, B. (2019) Positive-unlabeled convolutional neural networks for particle picking in cryo-electron micrographs. *Nature Methods*, 16, 1153-1160.
b. Bepler, T., Zhong, E.D., Kelley, K., Brignole, E., Berger, B. (2019) Explicitly disentangling image content from translation and rotation with spatial-VAE. *Advances in Neural Information Processing Systems*. Available from: https://papers.nips.cc/paper/9677-explicitly-disentangling-image-content-from-translation-and-rotation-with-spatial-vae
c. Zhong E.D. ,Bepler T., Davis J.H., Berger B. (2020) Reconstructing continuous distributions of 3D protein structure from cryo-EM images. *8th International Conference on Learning Representations*. Available from: https://openreview.net/pdf?id=SJxUjlBtwB

D. Additional Information: Research Support and/or Scholastic Performance

Scholastic Performance

YEAR	COURSE TITLE	GRADE
	MASSACHUSETTS INSTITUTE OF TECHNOLOGY	
2014	Principles of Synthetic Biology	A
2014	Machine Learning	A
2014	Advanced Computational Biology	A
2014	Topics in Computational and Systems Biology	A
2015	Inference and Information	B
2015	Molecular Biology	B
2015	Introduction to Numeric Methods	A
2015	Quantitative Genomics	A
2016	Topics in Computational Molecular Biology	A