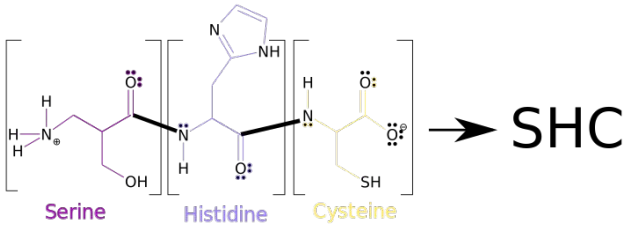
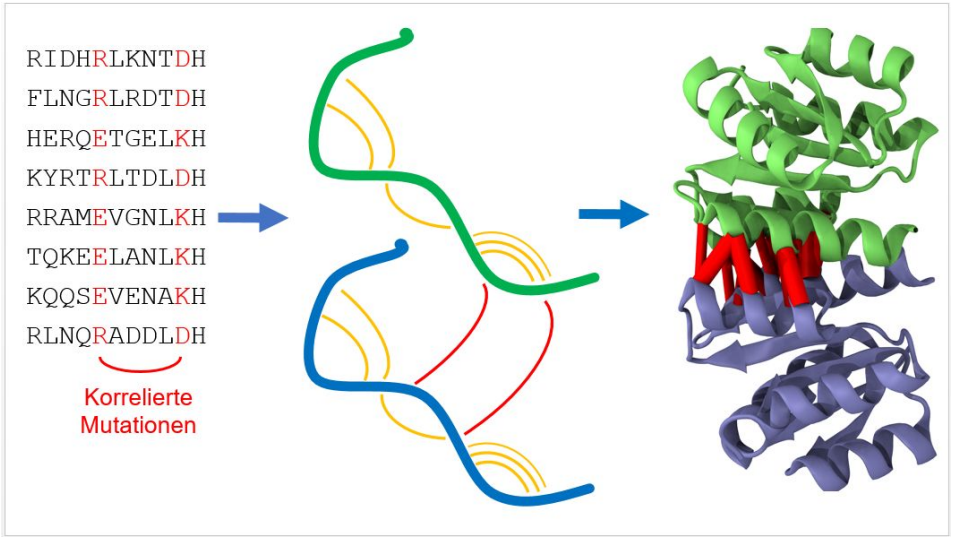
High Performance Computing for Protein Language Models

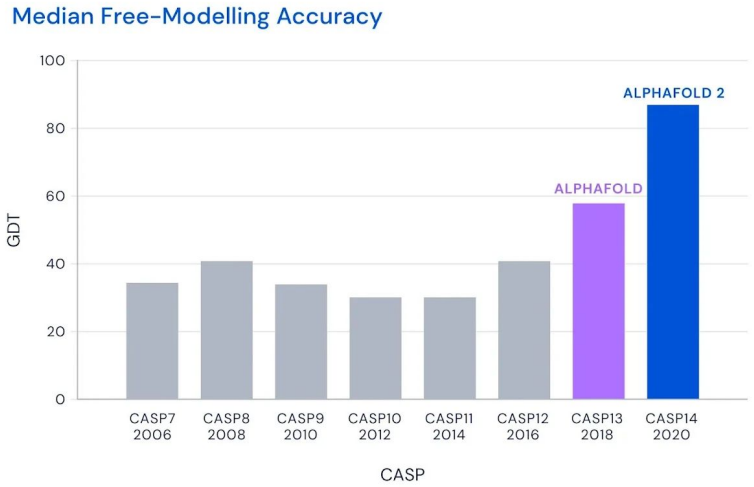
Anh T Nguyen
CS455

Proteins: Life's Essential Machines

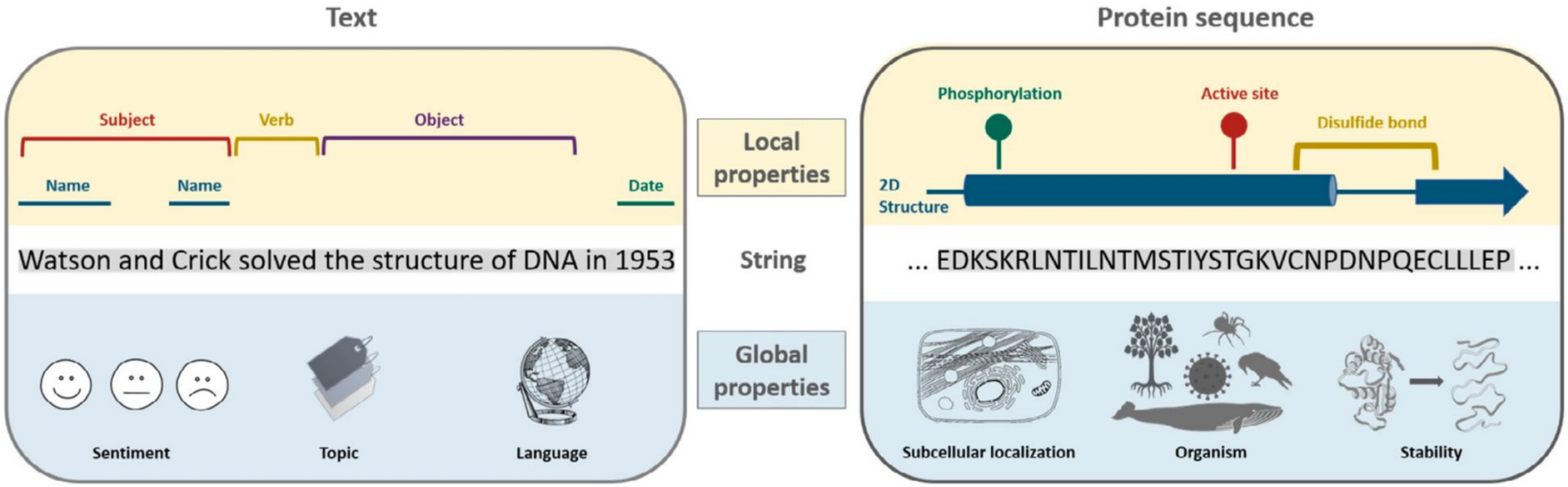


AlphaFold & Structure prediction





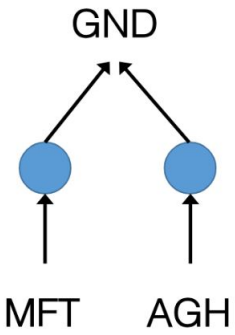
Proteins vs Natural Language



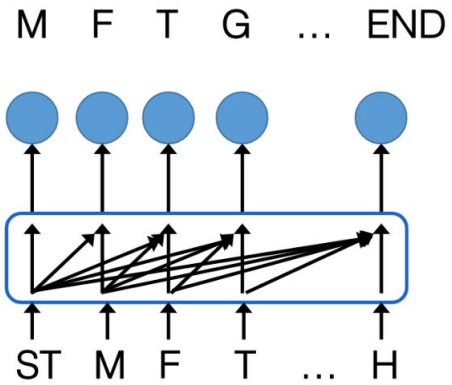
Ofer et al. (2021). The language of proteins: NLP, machine learning & protein sequences

Protein Language Model to the rescue

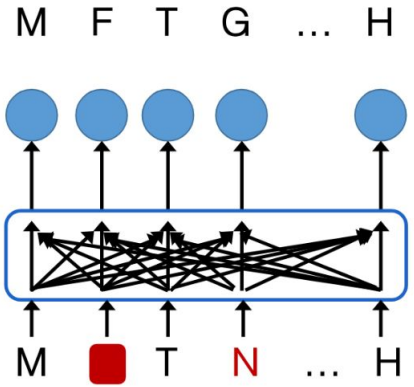
Word2Vec



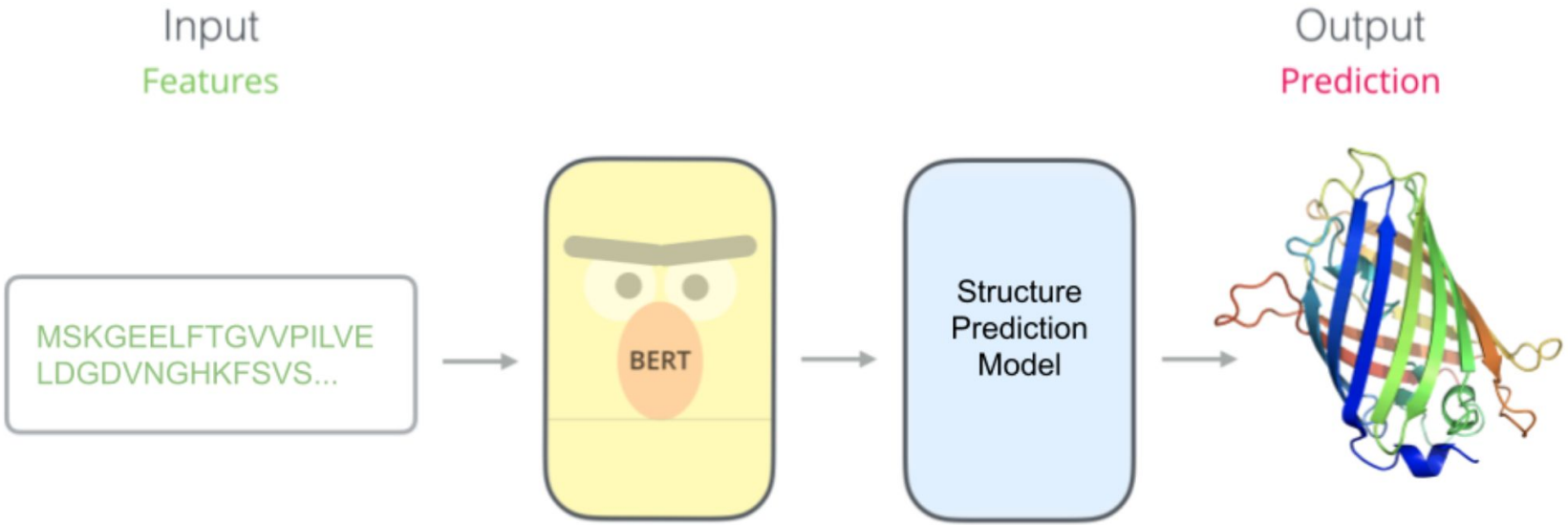
Autoregressive
language model



Masked
language model

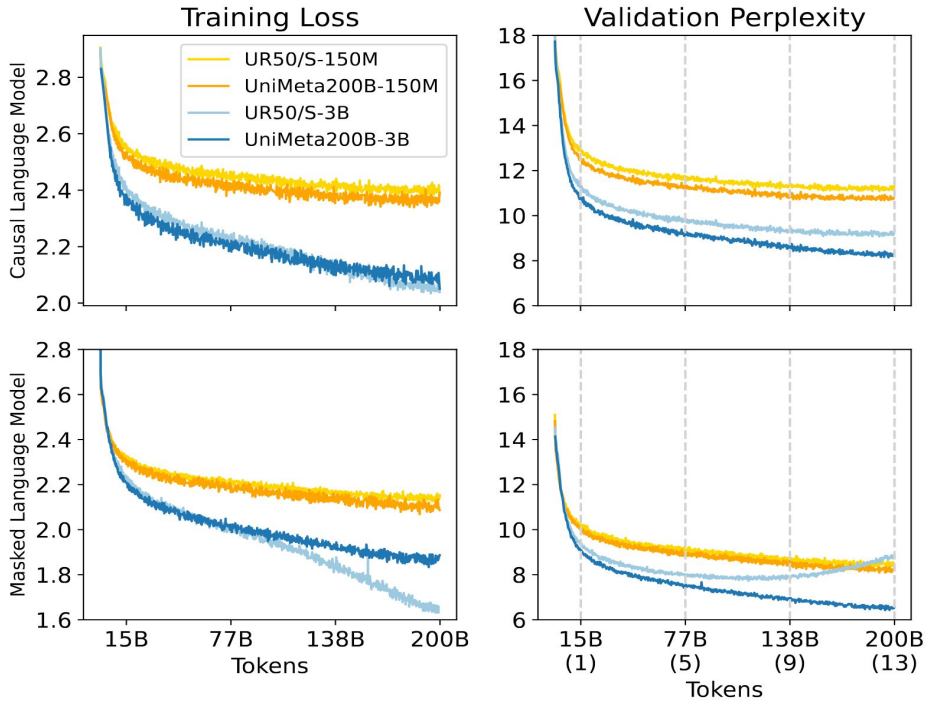


PLM Basics



https://bair.berkeley.edu

PLMs are data-hungry

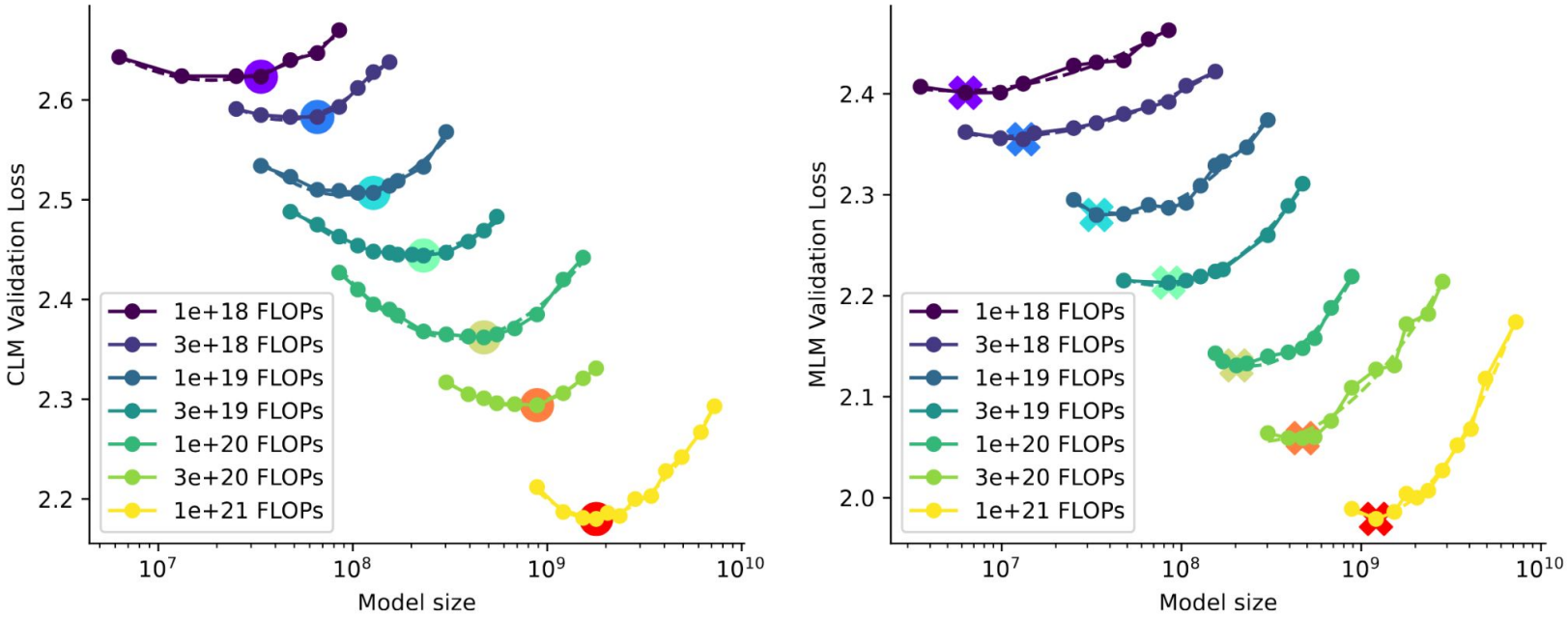


Using datasets with a broad range of unique tokens is essential

Using large-scale datasets for training is essential

Cheng et al. (2024) Training compute-optimal protein language models

Scaling laws for PLMs



Cheng et al. (2024) Training compute-optimal protein language models

Challenges of HPC for PLMs

- Scale of data
- Computational Demands
- Model Complexity
- Software and Infrastructure

Conclusion

- Training large-scale PLMs demands substantial resources, including powerful hardware like GPUs and specialized HPC infrastructure
- The development of novel hardware architectures, alongside innovative software and algorithmic approaches tailored for PLM training, will be essential for overcoming these computational hurdles.