Deep Learning in Structural Biology and Protein Design: Where, How, and Why

Deep Learning Theory Workshop and Summer School
Aug 5, 2022 @ Simons Institute
By Chloe Hsu (UC Berkeley)
Intro: A view of biology from living beings to molecules
“Where”: Advances in structural biology and protein design
Protein structures

Salmonella LP ring (It moves and swims!)
Structure prediction

Substantial improvement in structure prediction from AlphaFold2

Also notable in new capabilities for predicting protein-protein interactions (although not perfect yet)

Applications:

• Interprets complex experimental data in structural biology at a new speed
• Creates an unprecedented size of predicted structures for data mining & learning
• …
Structure prediction

Substantial improvement in structure prediction from AlphaFold2

Also notable in new capabilities for predicting protein-protein interactions (although not perfect yet)

Applications:

• Interprets complex experimental data in structural biology at a new speed
• **Creates an unprecedented size of predicted structures for data mining & learning**
• ...

Billions of known protein sequences

Experimentally determined structures (~0.1% of sequence space)
Billions of known protein sequences

New predicted structures (200+ millions)

Experimentally determined structures (~0.1% of sequence space)
Protein sequence models

- Modeling the distribution of protein sequences through language models or other density models

- Bigger is better??

- Applications
  - Predicts effects of genetic mutations
  - Guides sequence choices in protein engineering

![Chart showing model performance with different parameter sizes](chart.png)

- ProGen2-small: 151M params
- ProGen2-medium: 6.4B params
Sequence density model (language model) \[\Rightarrow\] Protein sequence \[\Rightarrow\] Sequence likelihood

Neutral variation

Disease variation

Enhanced variation
Generative models for protein design

Denoising diffusion probabilistic model for protein structure and sequence

Generative models for protein design

Conditional generative models for protein design, e.g. inverse folding models

Beyond naturally existing proteins: De novo proteins

“How”: The data and the learning algorithms
Protein structures: Protein Data Bank

(Today over 100,000)
Evolutionary data: observing the products of evolution
Evolutionary data: observing the products of evolution

Human PIF1 helicase (PDB: 6HPH)  
Thermophilic bacteria PIF1 helicase (PDB: 6S3E)  
(~30% sequence identity)
Evolutionary data: observing the products of evolution

Multiple sequence alignment (MSA)

PIF1 helicase

WALRKTRKRLLEEPFGGKVLLLGDTRQLEPVVPGGEALYIARTWGGPFFFQAHVWEER
R R+ +PFGG+++ GD QL PV G + F FQ+ W+
AVARAVRQQ--NKPFGGIQLIIICDFLQPLPPVTKGSQPR--------RFQFQSKSWKRCV

--VALRHVLWESQRQRERDPLFAELLKRLRQG--DPQALETLNRAAVRPDGEEPGTLLLT
VL+ ++W ++ D F LL+ +R G + L A G + L
PVTLELTKVW----RQADQTFISLLQAVRGLRSDEVTRQLQATASHKVGDRGIVATRLC

PRRKEADALNLKRLDELPKPLEYQAQVKG--EFAET---DFPTEAAALTKQGAVLRRN
+ + N +RL+ LPK ++A E A T P L K GAQV+L++N
THQDDVALTNERRLQPDELPGKVHRFELMDSNPLOLSTDLAQCPVSQLLQLGKGAQVMVKN

DPLGE--YNFDGLGWVDELAEALAVRLKR--NGRRVVRIPFVWEKIVTYDSEREEIKPQ
+ NG V EAE + R G VI W T + ++ +
LVSVRGLNVNGARGVVVGEAEGRLPQVRFLCGVTEIYHADRW----TVQATGGQLLSR

VVGTFRQVPVRLLAWALTWHKAQGLTLKDHVLELRGGLFAHGQLYVALTRVRRLQDL
+Q+P++LAWA+++HK+QG+TLD V + LGR +FA GQ YVAL+R R LQ L
-------QQLPLQLAWAMSITHKSQGMULTDCVEISLGR--VFASGQAYVALSRARSQGL

180
168
237
224
293
284
350
339
406
389
Evolutionary density model

High density area = likely functional proteins

Before deep learning: Simple models on evolutionary data

Hidden Markov Models (HMMs) only model the site-specific amino acid features.

Potts models, also known as Markov Random Fields (MRFs), model the site-specific amino acid features and the pairwise interactions between sites.

They work quite well, too
(when used on specific protein families in combination with search algorithms)
Replacing multiple sequence alignments with deep learning models
(Protein language models)

PIF1 helicase

Insulin

Casein

New ability to jointly model all protein families regardless of “alignments”
AlphaFold2 also relies on evolutionary data (multiple sequence alignments)

Recycling iteration 0, block 01
Secondary structure assigned from the final prediction
“Why”: Common themes and open questions
“Search & learn” vs. “universal learning”

Search for relevant sequences → Learn a “local” model

Learning a “universal protein model”
"Generalization" vs "memorization"

Are language models / single-sequence folding models effectively memorizing the evolutionary database?