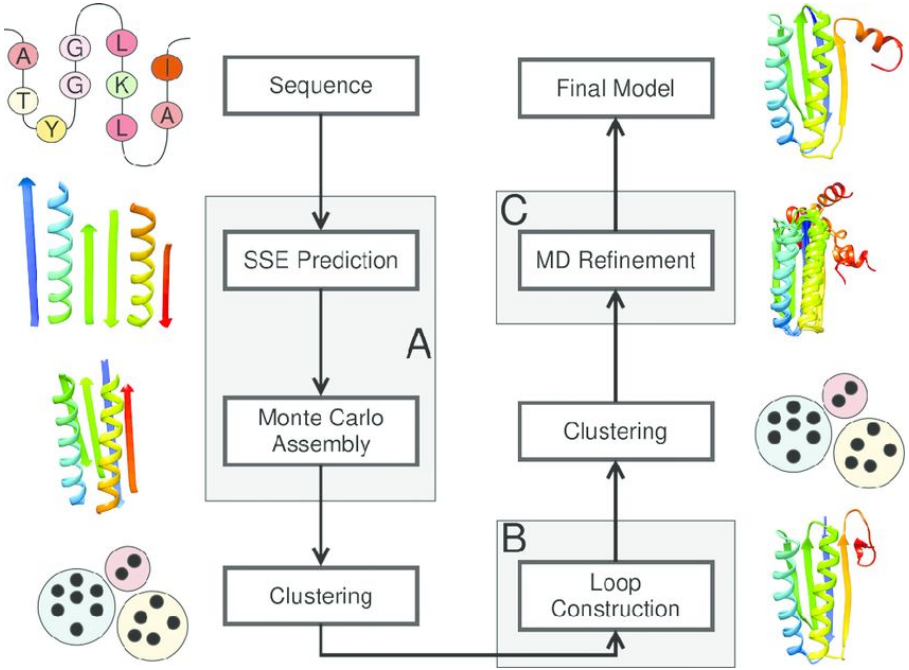
# Learning inverse folding from millions of predicted structures

Chloe Hsu et al.
Presented by Pascal Sturmfels
05/02/2022

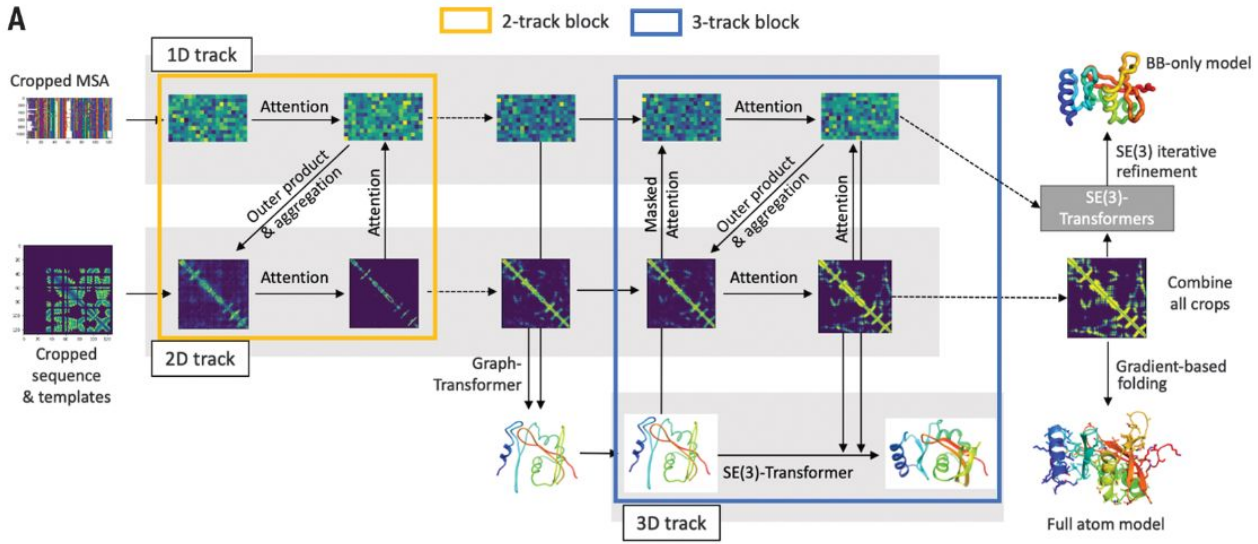
# Protein Structure Prediction

- Determining the structure of a protein given its amino acid sequence is a fundamental problem in computational biology
- Older modeling approaches used templates, underlying knowledge of molecular dynamics, or neural networks to predict contacts and bond angles and then refine the structures to final 3D coordinates



# The AlphaFold2 Era

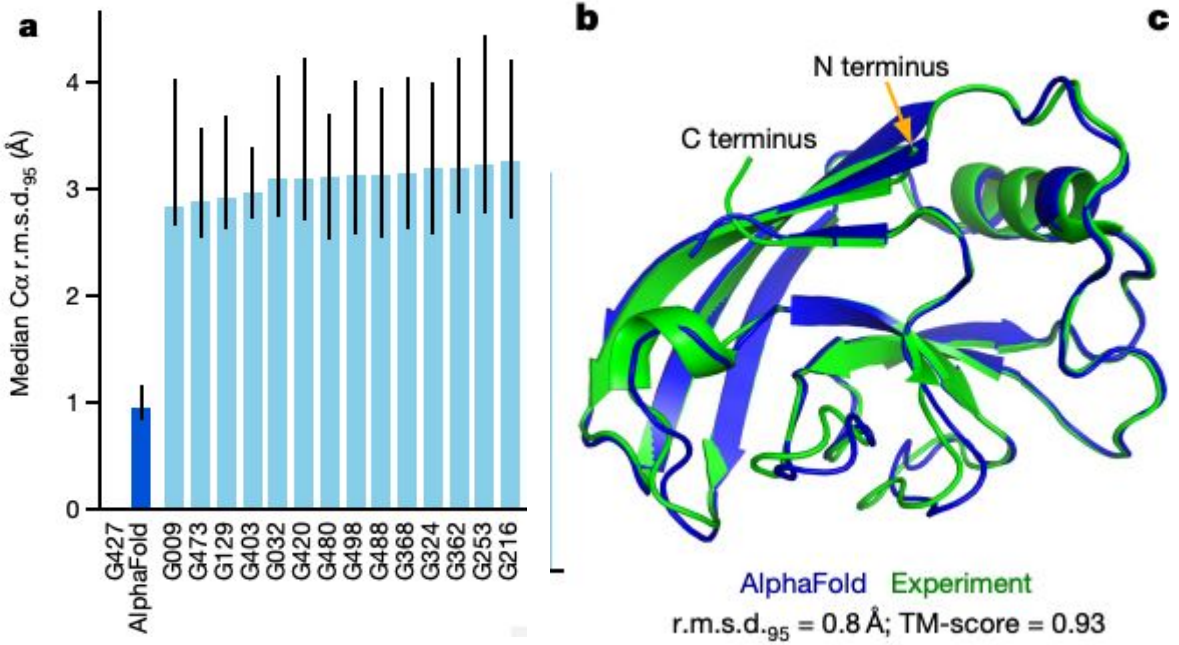
- AlphaFold2 and RoseTTAfold revolutionized structure prediction by predicting 3D coordinates directly using large transformer models




Baek, M., Dimaio, F., Anishchenko, I., Dauparas, J., Ovchinnikov, S., Lee, G. R., Wang, J., Cong, Q., Kinch, L. N., Schaeffer, R. D., Millán, C., Park, H., Adams, C., Glassman, C. R., Degiovanni, A., Pereira, J. H., Rodrigues, A. v, van Dijk, A. A., Ebrecht, A. C., … Baker, D. (2021). Accurate prediction of protein structures and interactions using a three-track neural network.

# Experimental Quality

- These new methods are close to experimental quality for many domains
- This enables a wide variety of applications in design and discovery



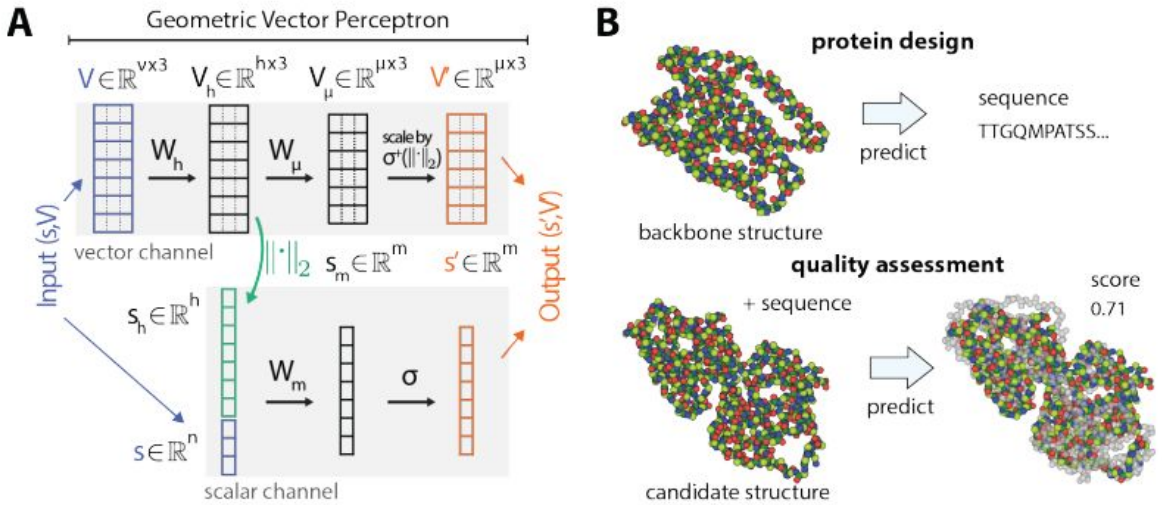

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., … Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold.

# The Inverse Problem

- A slightly less well covered problem is the inverse folding problem: given a structure, can we generate a sequence that folds to this structure?
- There are several different challenges in this tasks, including the fact that the map is one-to-many rather than many-to-one

# The General Approach to Inverse Design

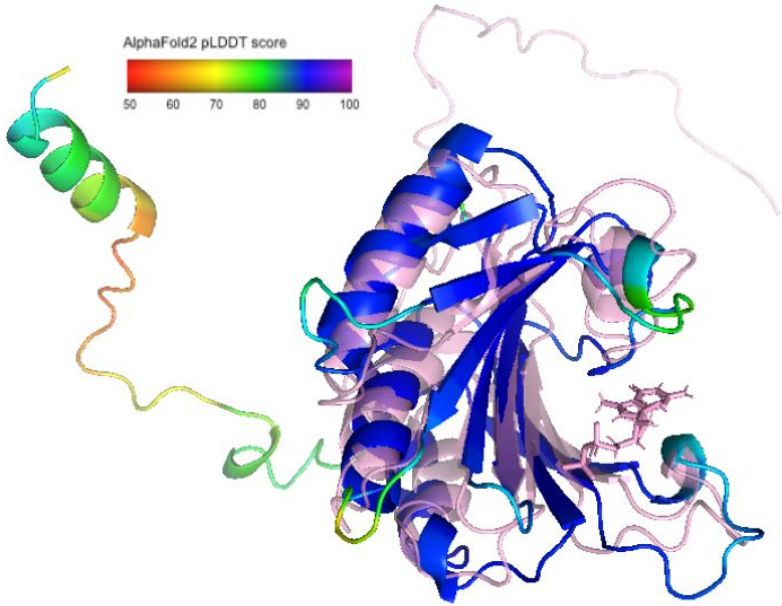
- Typically NN based
- The issue is lack of available data - generative NN are very data hungry




Jing, Bowen, et al. "Learning from Protein Structure with Geometric Vector Perceptrons." International Conference on Learning Representations. 2020.

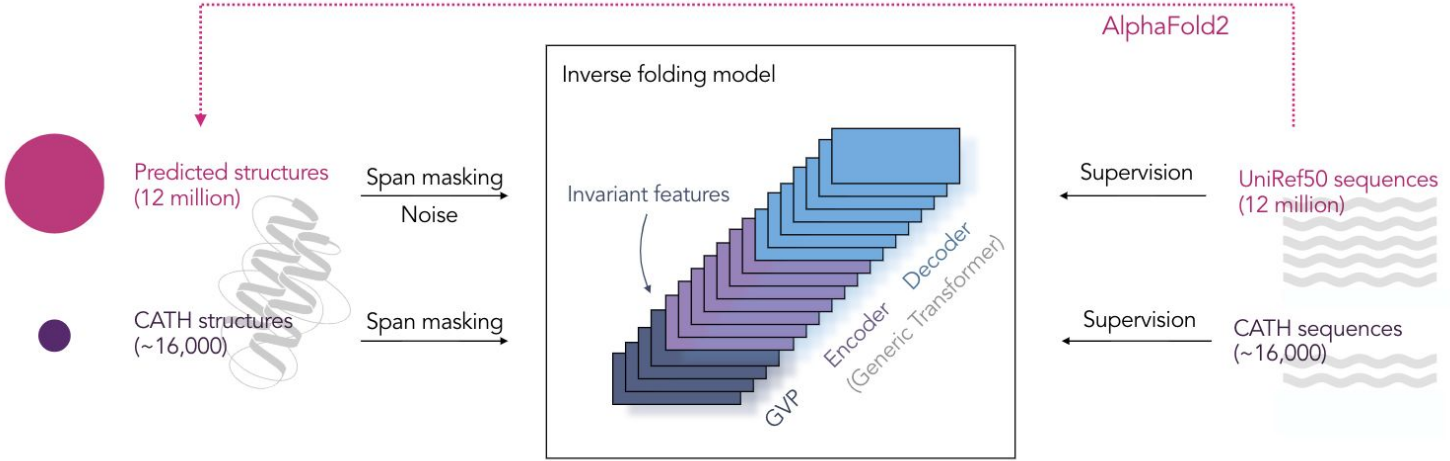
# The Data

- They augment experimental structure data (from CATH) with 12 million AlphaFold2 predicted structures
- Sequences are first passed through the MSA transformer and evaluated via pLDDT to determine predicted structure quality



# Architectures

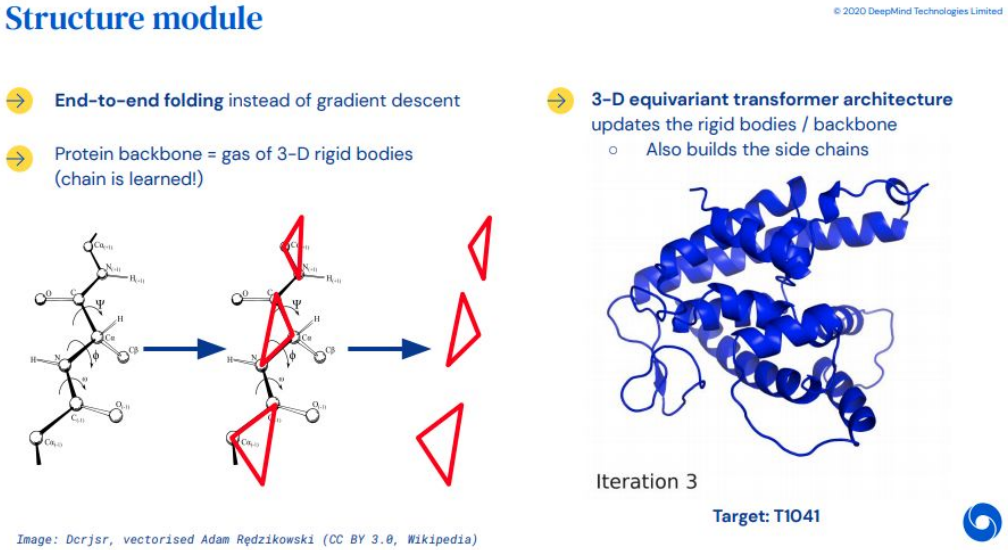
- Encoder is a Geometric Vector Perceptron, a type of neural network that is equivariant to the SE(3) group
- Generator is a standard transformer



# A Note on Equivariance

- Equivariance is a key concept in geometric deep learning
- An equivariant function is one such that:

$$h\big(\sigma(x)\big) = \sigma\big(h(x)\big) \qquad \forall \sigma, x$$



# A Note on Equivariance

- CNNs are translation equivariant
- Achieving SE(3) equivariance in general is… complicated
- It relies on the composition of equivariant functions and the fact that the norm of a vector is invariant to rotation

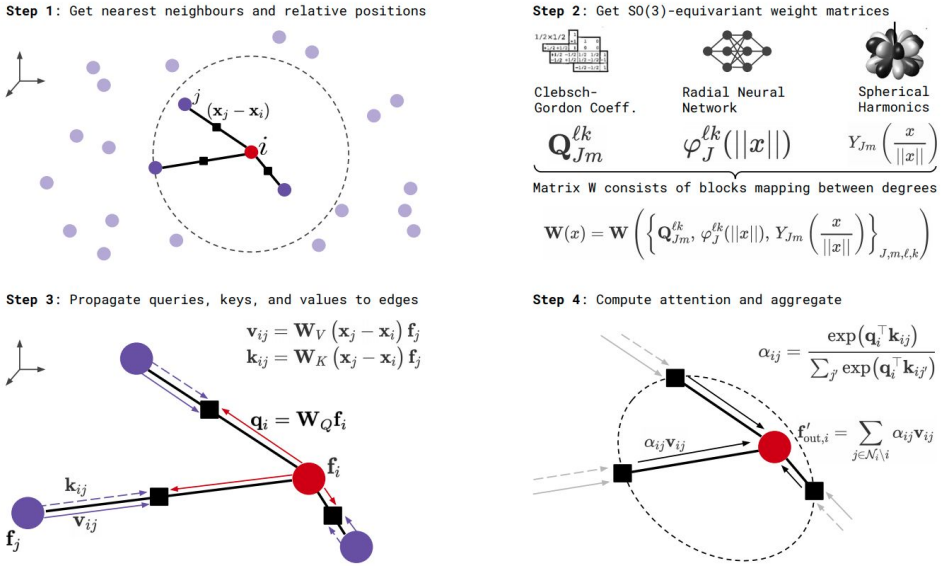
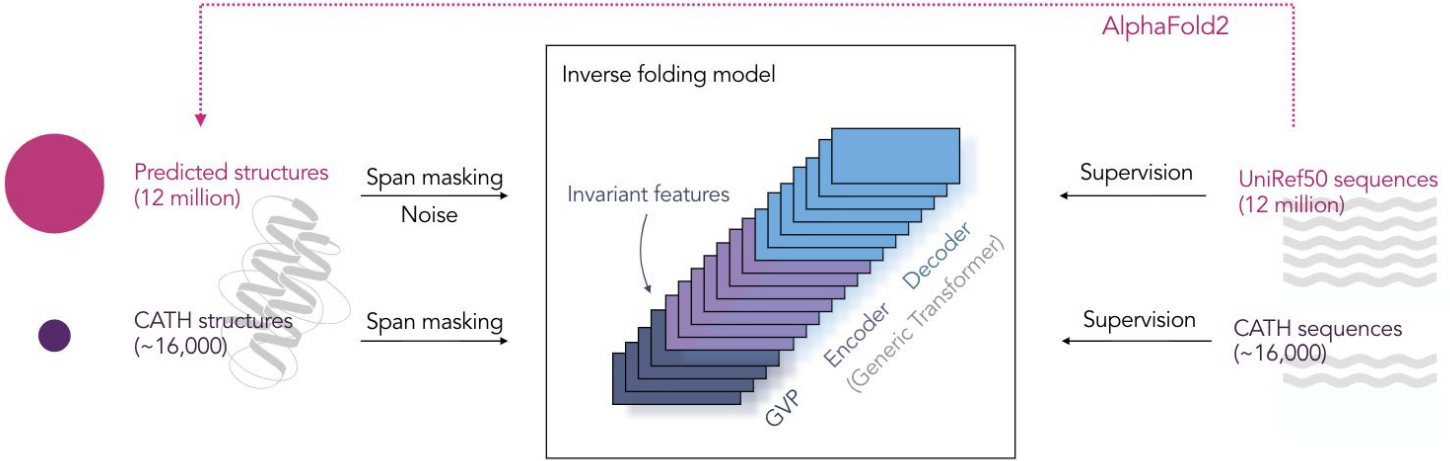



Figure 2: Updating the node features using our equivariant attention mechanism in four steps. A more detailed description, especially of step 2, is provided in the Appendix. Steps 3 and 4 visualise a graph network perspective: features are passed from nodes to edges to compute keys, queries and values, which depend both on features and relative positions in a rotation-equivariant manner.

# Training Tasks

- The network is trained to do autoregressive generation
- Additionally, portions of the structure are masked out via span masking and additionally gaussian noise is added to the predicted structure coordinates during training



# Evaluation Setting

- Perplexity is a standard NLP metric and is inversely proportional to the probability of the test set
- Native sequence recovery % is the percentage of amino acids in the native sequence recovered by predicted sequence
- Neither metric is perfect - assumption that the native sequence should be highly likely given the structure

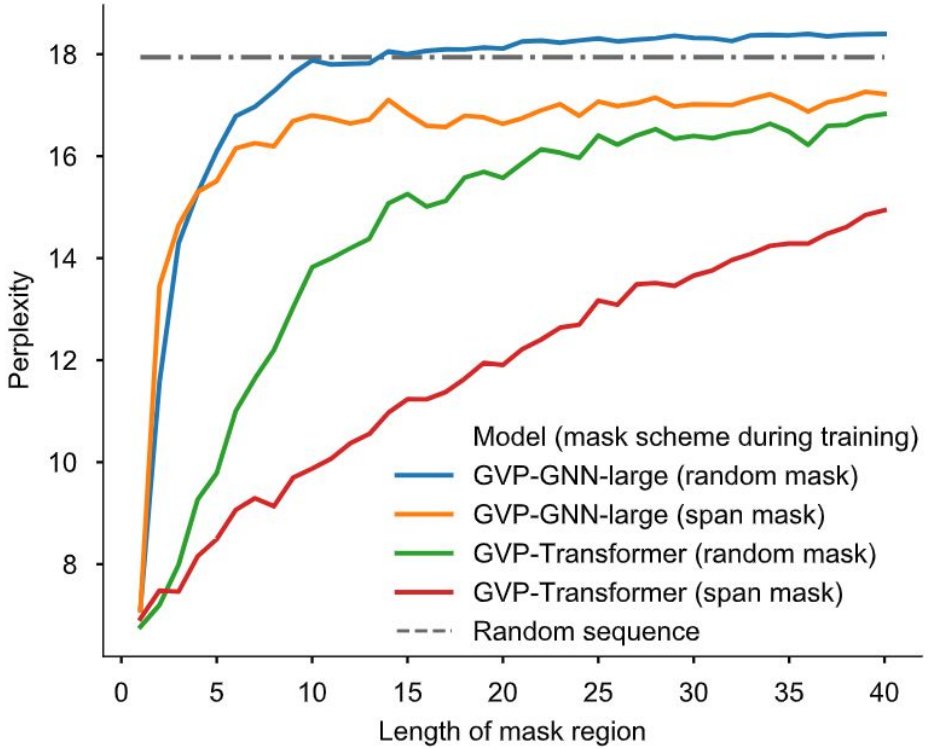
# Results

| Model | Data | Perplexity | | | Recovery % | | |
|---|---|---|---|---|---|---|---|
| | | Short | Single-chain | All | Short | Single-chain | All |
| Natural frequencies | | 18.12 | 18.03 | 17.97 | 9.6% | 9.0% | 9.5% |
| Structured GNN | CATH | 7.91 | 6.48 | 6.49 | 31.5% | 37.1% | 37.1% |
| GVP-GNN | CATH | 7.14 | 5.36 | 5.43 | 34.0% | 42.7% | 42.2% |
| | + AlphaFold2 | 8.55 | 6.17 | 6.06 | 29.5% | 38.2% | 38.6% |
| GVP-GNN-large | CATH | 7.68 | 6.12 | 6.17 | 32.6% | 39.4% | 39.2% |
| | + AlphaFold2 | 6.11 | 4.09 | 4.08 | **38.3%** | 50.8% | 50.8% |
| GVP-Transformer | CATH | 8.18 | 6.33 | 6.44 | 31.3% | 38.5% | 38.3% |
| | + AlphaFold2 | **6.05** | **4.00** | **4.01** | 38.1% | **51.5%** | **51.6%** |

*Table 1.* Fixed backbone sequence design. Evaluation on the CATH 4.3 topology split test set. Models are compared on the basis of per-residue perplexity (lower is better; lowest perplexity bolded) and sequence recovery (higher is better; highest sequence recovery bolded). Large models can make better use of the predicted UniRef50 structures. The best model trained with predicted structures (GVP-Transformer) improves sequence recovery by 8.9 percentage points over the best model (GVP-GNN) trained on CATH only.

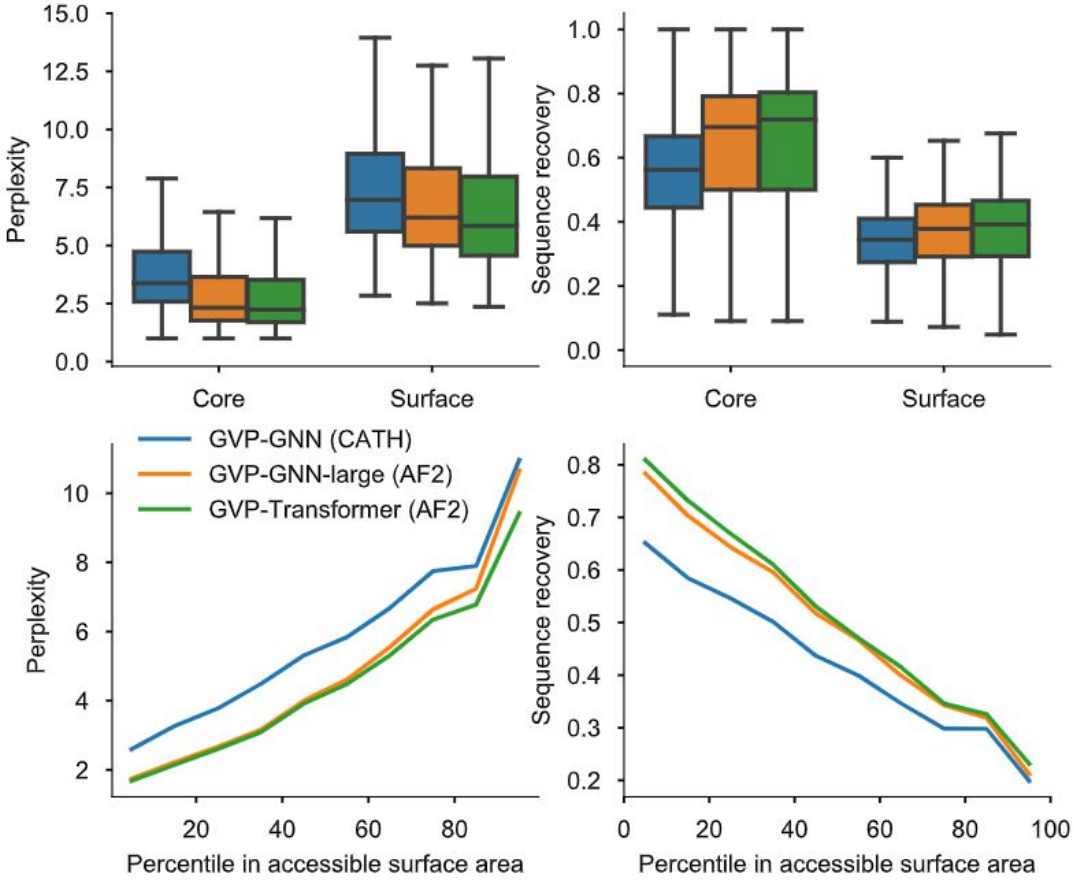
# Inpainting (span masking)

- Their architecture maintains impressive performance when masking large spans of structure
- The transformer clearly outperforms here



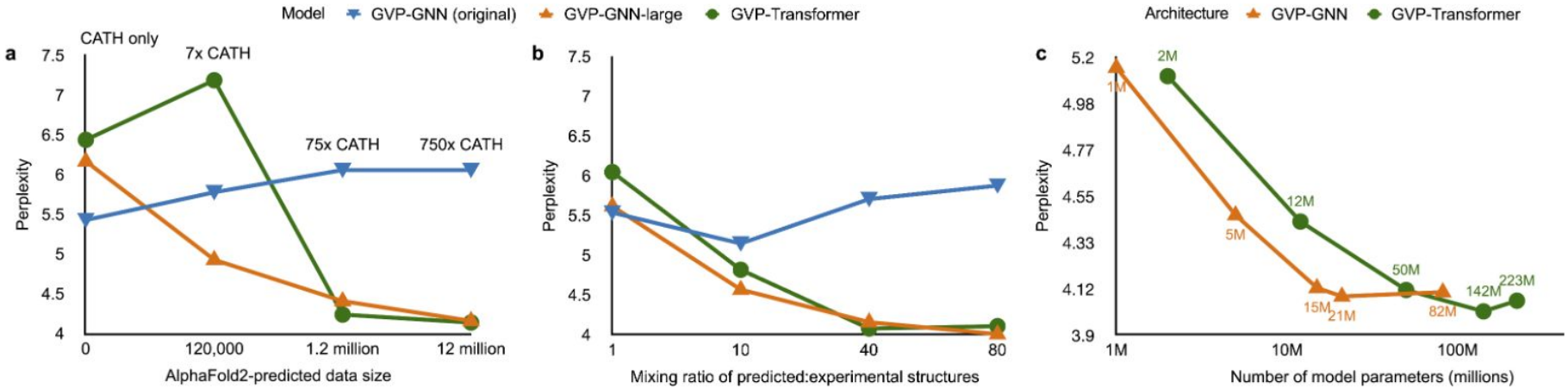
# Buried Residues

- Residues closer to the core of a protein are usually thought to be more evolutionarily constrained than those on the surface
- This is confirmed in their prediction results



# Ablation Studies

- Amount of data dominates, as long as the model is of sufficiently high capacity
- This holds true even if data is noisy/predicted



# Generalizing to Complexes

- It was observed that structure predictors can generalize out-of-domain to protein complex structures
- It would appear this model can do the same, even though it has never seen protein complexes in the training set

| Model | Perplexity | |
|---|---|---|
| | Chain | Complex |
| Natural frequencies | 17.93 | |
| GVP-GNN | 7.80 | 5.37 |
| GVP-GNN-large+AF2 | **6.32** | 3.90 |
| GVP-Transformer+AF2 | **6.32** | **3.81** |

*Table 2.* Sequence design performance on complexes in the CATH topology test split when given the backbone coordinates of only a chain ("Chain" column) and when given all backbone coordinates of the complex ("Complex" column). The perplexity is evaluated on the same chain in the complex for both columns.

# Multi-State Design

- Their results extend to proteins with multiple stable conformations: averaging over multiple stable structures decreases perplexity

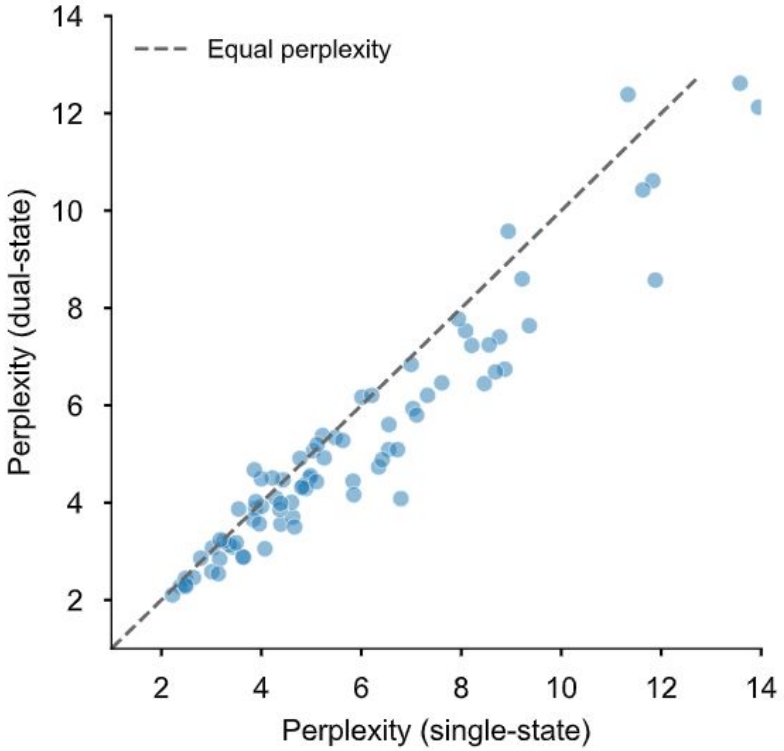
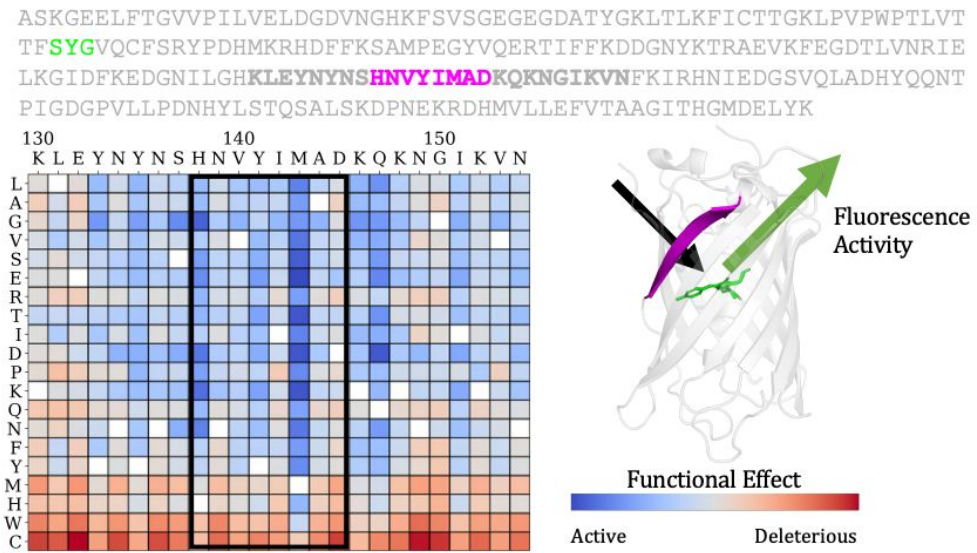



*Figure 7.* Dual-state design. GVP-Transformer conditioned on two conformations results in lower sequence perplexity at locally flexible residues than single-conformation conditioning for structurally held-out proteins in PDBFlex (see Appendix C for details).

# Predicting Variant Effects

- Language models can be used to predict variant effects
- The effect of a mutation is quantified by log-likelihood of the mutated sequence
- This was shown in previous work to be strongly predictive of mutational effects

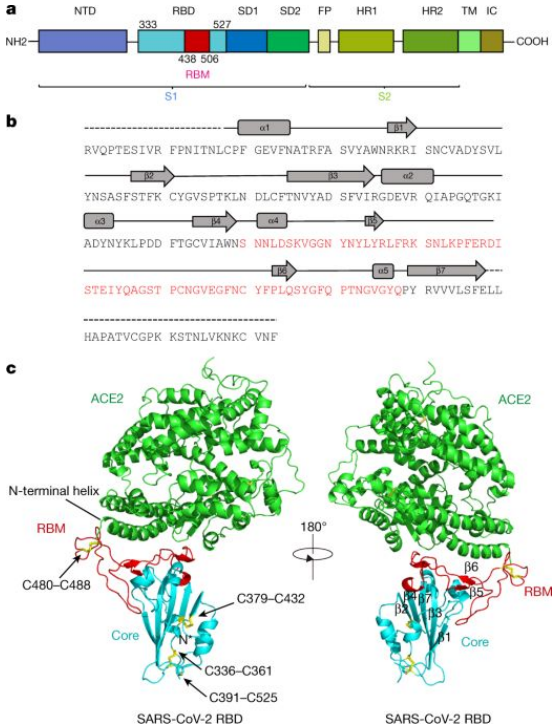



Meier, Joshua, et al. "Language models enable zero-shot prediction of the effects of mutations on protein function." Advances in Neural Information Processing Systems 34 (2021).

# COVID Receptor Binding Domain

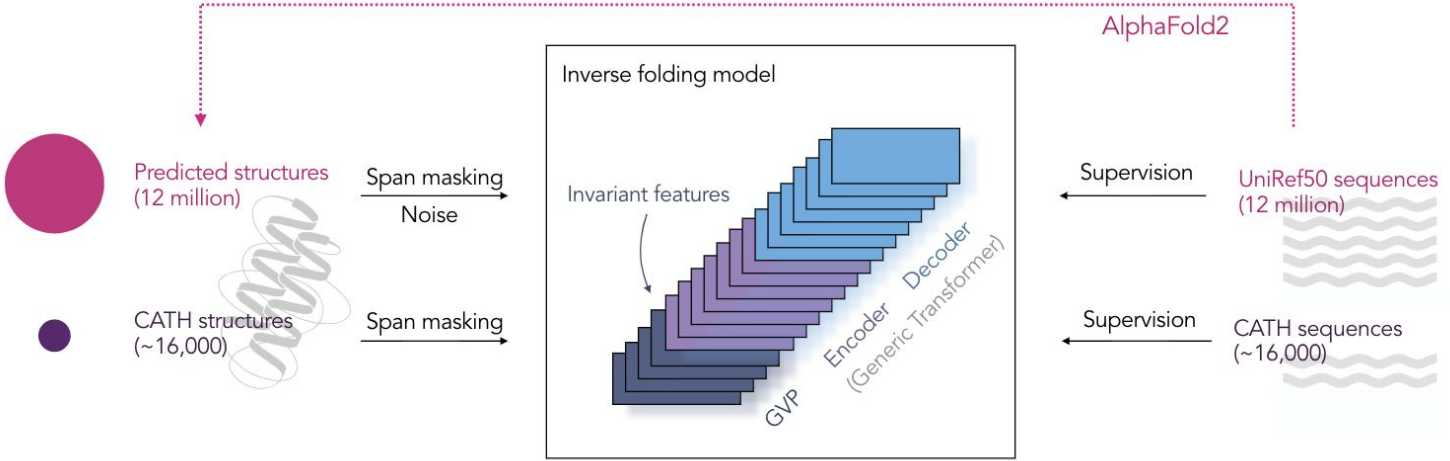
- Consider the portion of the COVID spike protein that binds to the ACE2 receptor
- Binding affinity of the various mutations is of particular interest
- Paper can “zero shot” predict binding affinity

| | Spearman correlation | | | |
|---|---|---|---|---|
| Model | No coords | No RBM coords | No ACE2 coords | All coords |
| ESM-1v | 0.03 | | | |
| ESM-1b | 0.02 | | | |
| ESM-MSA-1b (few-shot) | **0.51** | | | |
| GVP-GNN | | -0.10 | 0.50 | 0.60 |
| GVP-GNN-large+AF2 | | -0.05 | 0.52 | **0.69** |
| GVP-Transformer+AF2 | | -0.06 | **0.53** | 0.64 |



# Summary

- Paper provides a framework for fixed-backbone protein design, leveraging the computational predictions of modern structure predictors
- Their model generalizes beyond single-domain design to a variety of related problems



# Discussion

- The section on multi-state design feels limited - would be interested in follow up experiments showing fold of synthesized proteins in a wetlab
- How are big structure models able to generalize to protein complexes when they are only trained on single domains?
- The transformer architecture seems particularly suited to protein problem tasks given how we think about contacts

# Questions?