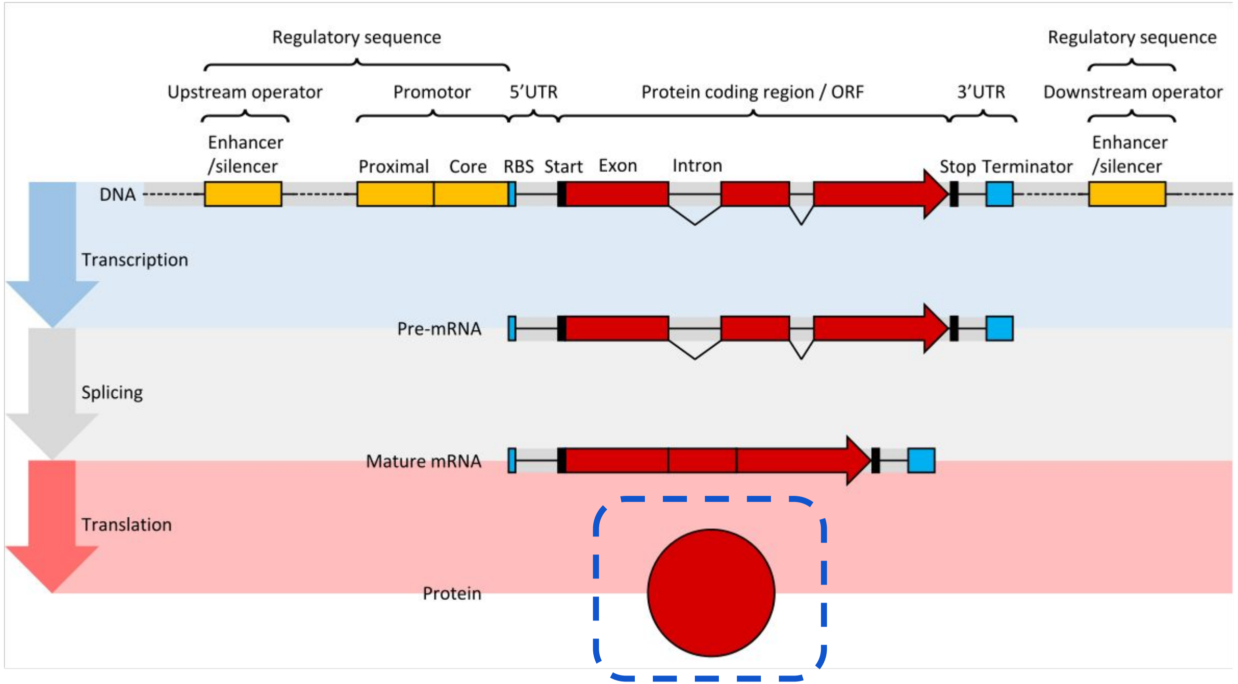
Improved protein structure prediction using potentials from deep learning

Andrew W. Senior, Richard Evans, John Jumper, James Kirkpatrick, Laurent Sifre, Tim Green, Chongli Qin, Augustin Žídek, Alexander W. R. Nelson, Alex Bridgland, Hugo Penedones, Stig Petersen, Karen Simonyan, Steve Crossan, Pushmeet Kohli, David T. Jones, David Silver, Koray Kavukcuoglu & Demis Hassabis

NOTE: CASP14 (2020) results not published - this is CASP13 (2018)

Background 1 of 2: Proteins

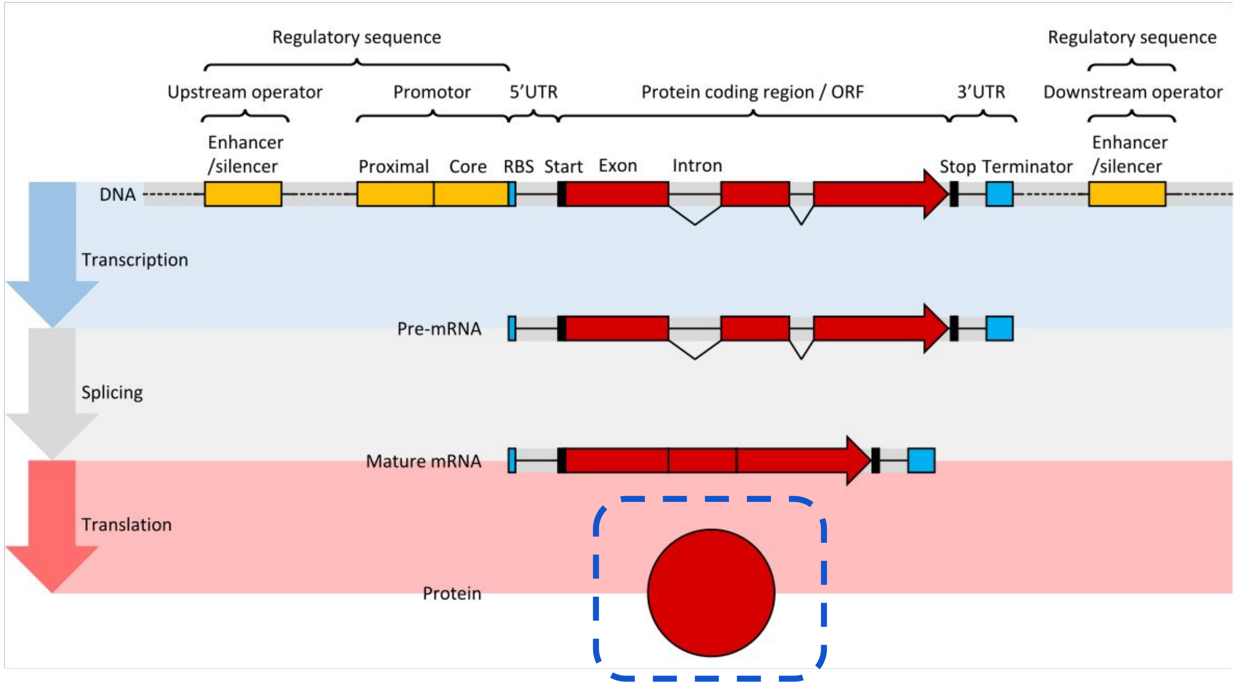
Recall: DNA → RNA → Protein



"Folded" Protein

Brief Intro to Genetics and Bioinformatics

Recall: DNA → RNA → Protein

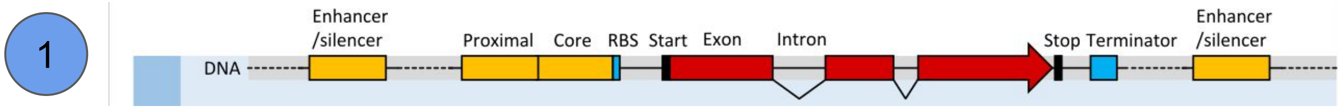


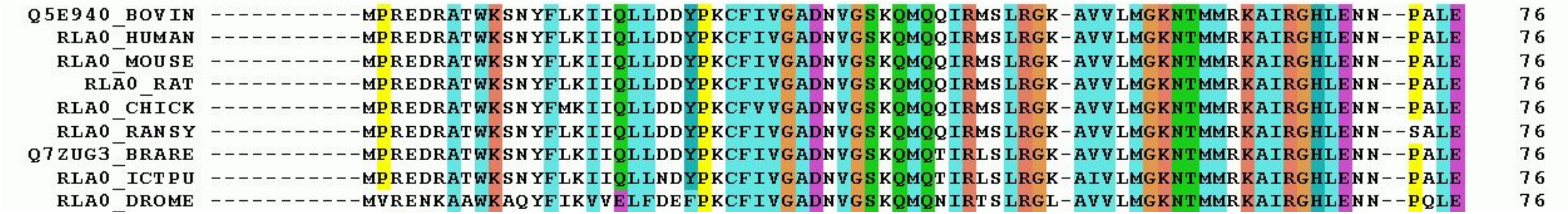
"Folded" Protein

"Folded" Cow

Brief Intro to Genetics and Bioinformatics

Multiple Sequence Alignment



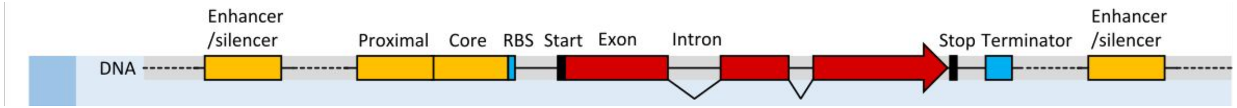


```
Q5E940_BOVIN -----------MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKQMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE   76
 RLA0_HUMAN -----------MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKQMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE   76
 RLA0_MOUSE -----------MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKQMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE   76
   RLA0_RAT -----------MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKQMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE   76
 RLA0_CHICK -----------MPREDRATWKSNYFMKIIQLLDDYPKCFVVGADNVGSKQMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE   76
 RLA0_RANSY -----------MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKQMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--SALE   76
Q7ZUG3_BRARE -----------MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKQMQTIRLSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE   76
 RLA0_ICTPU -----------MPREDRATWKSNYFLKIIQLLNDYPKCFIVGADNVGSKQMQTIRLSLRGK-AIVLMGKNTMMRKAIRGHLENN--PALE   76
 RLA0_DROME -----------MVRENKAAWKAQYFIKVVELFDEFPKCFIVGADNVGSKQMQNIRTSLRGL-AVVLMGKNTMMRKAIRGHLENN--PQLE   76
```

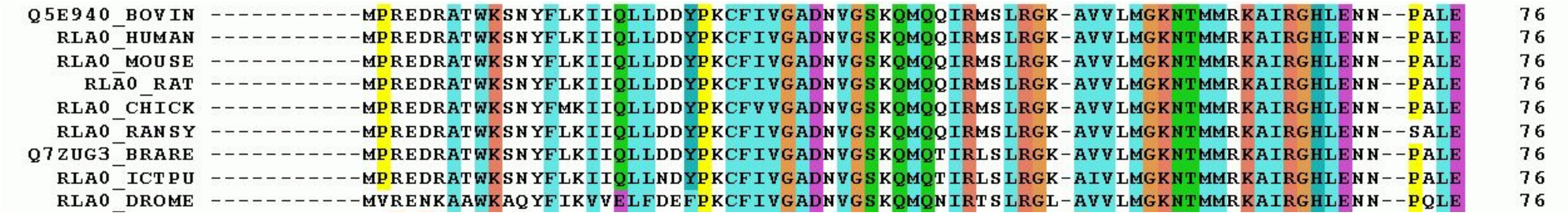
1) Predict gene structure (E.g. using known start sequences, splice sequences, etc). Now we have linear AA sequence of 1 organism.

Multiple Sequence Alignment

1



2



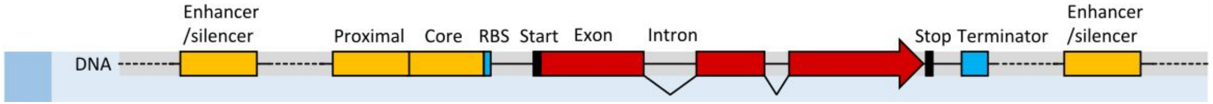
```
Q5E940_BOVIN -----------MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKQMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE    76
  RLA0_HUMAN -----------MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKQMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE    76
  RLA0_MOUSE -----------MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKQMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE    76
    RLA0_RAT -----------MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKQMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE    76
  RLA0_CHICK -----------MPREDRATWKSNYFMKIIQLLDDYPKCFVVGADNVGSKQMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE    76
  RLA0_RANSY -----------MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKQMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--SALE    76
Q7ZUG3_BRARE -----------MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKQMQTIRLSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE    76
  RLA0_ICTPU -----------MPREDRATWKSNYFLKIIQLLNDYPKCFIVGADNVGSKQMQTIRLSLRGK-AIVLMGKNTMMRKAIRGHLENN--PALE    76
  RLA0_DROME -----------MVRENKAAWKAQYFIKVVELFDEFPKCFIVGADNVGSKQMQNIRTSLRGL-AVVLMGKNTMMRKAIRGHLENN--PQLE    76
```

1) Predict gene structure (E.g. using known start sequences, splice sequences, etc). Now we have linear AA sequence of 1 organism.

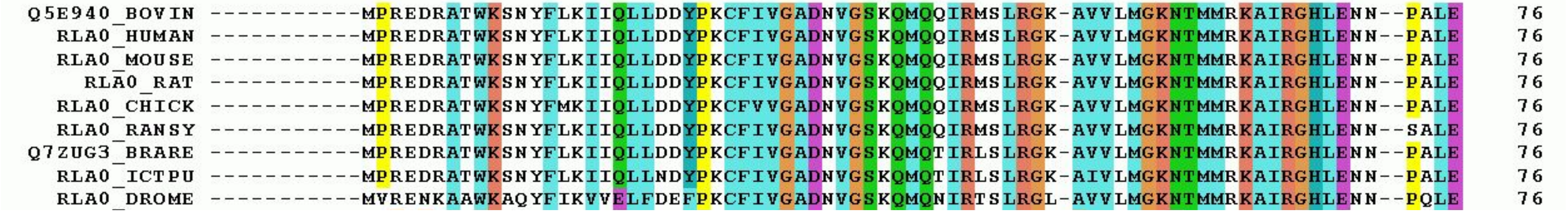
2) Find substring matches across AA sequence of multiple organisms. ("homologous regions")

Multiple Sequence Alignment

1



2



```
Q5E940_BOVIN ------------MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKQMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE   76
  RLA0_HUMAN ------------MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKQMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE   76
  RLA0_MOUSE ------------MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKQMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE   76
    RLA0_RAT ------------MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKQMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE   76
  RLA0_CHICK ------------MPREDRATWKSNYFMKIIQLLDDYPKCFVVGADNVGSKQMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE   76
  RLA0_RANSY ------------MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKQMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--SALE   76
Q7ZUG3_BRARE ------------MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKQMQTIRLSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE   76
  RLA0_ICTPU ------------MPREDRATWKSNYFLKIIQLLNDYPKCFIVGADNVGSKQMQTIRLSLRGK-AIVLMGKNTMMRKAIRGHLENN--PALE   76
  RLA0_DROME ------------MVRENKAAWKAQYFIKVVELFDEFPKCFIVGADNVGSKQMQNIRTSLRGL-AVVLMGKNTMMRKAIRGHLENN--PQLE   76
```

1) Predict gene structure (E.g. using known start sequences, splice sequences, etc). Now we have linear AA sequence of 1 organism.

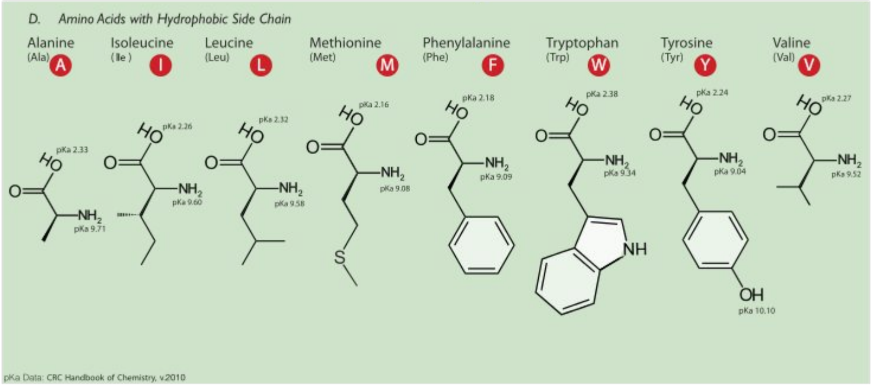
2) Find substring matches across AA sequence of multiple organisms. ("homologous regions")

3) Make deductions, such as:
   a) If we know the structure of one, we can guess structure of its matches.
   b) If a region is highly conserved, except two positions whose changes are correlated, those positions may be in contact.

Potts Hamiltonian models of protein co-variation, free energy landscapes, and evolutionary fitness

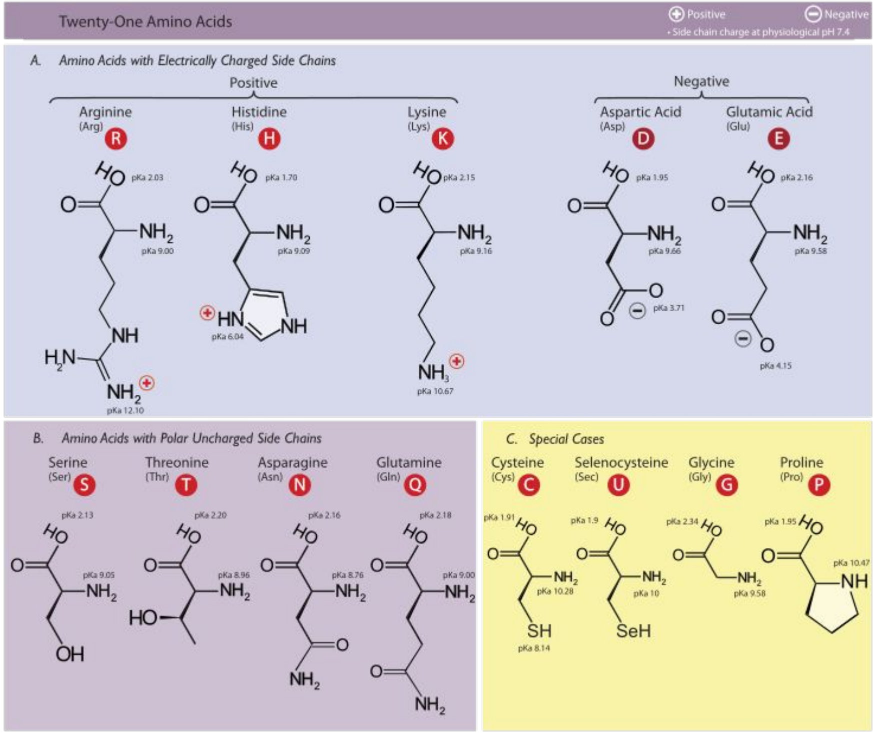
Amino Acid Structure

Side chains determine chemical properties



pKa Data: CRC Handbook of Chemistry, v.2010

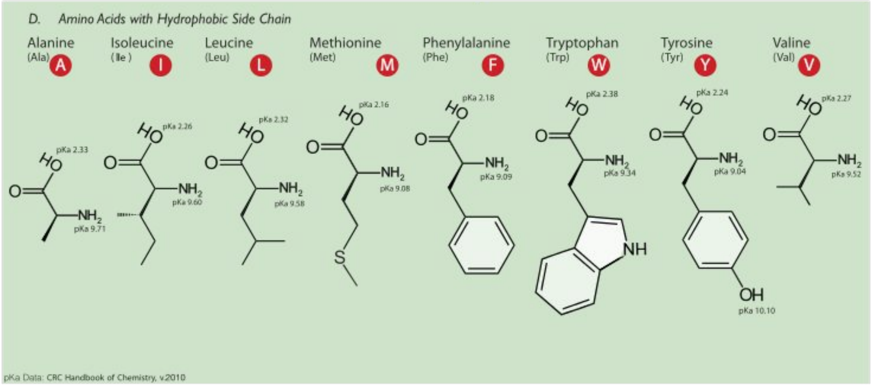
Dan Cojocari, Department of Medical Biophysics, **University of Toronto**, 2010



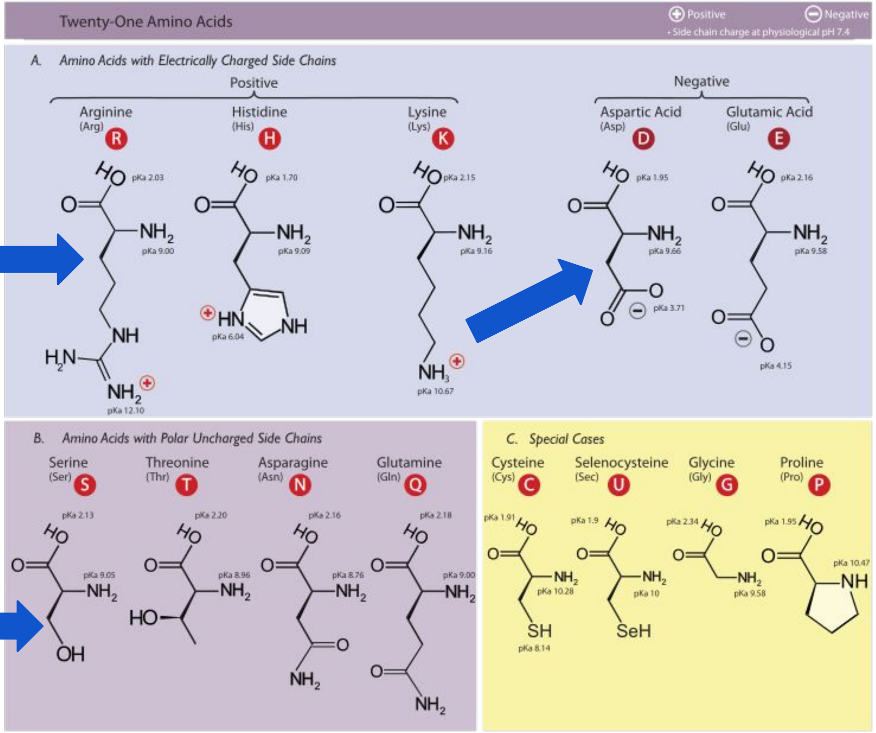
https://en.wikipedia.org/wiki/Proteinogenic_amino_acid

Amino Acid Structure

Side chains determine chemical properties



Dan Cojocari, Department of Medical Biophysics, **University of Toronto**, 2010



https://en.wikipedia.org/wiki/Proteinogenic_amino_acid

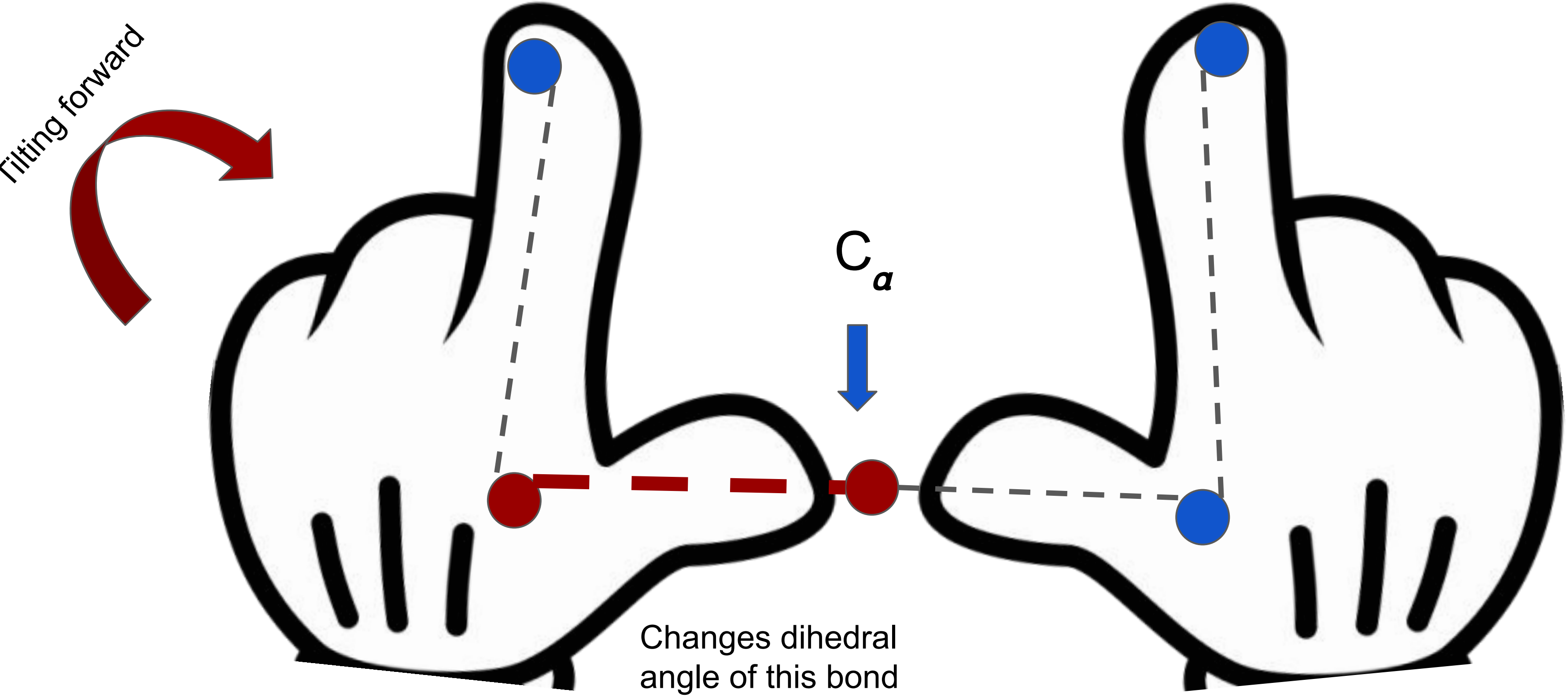
Describing Protein Structure: Phi-Psi Angles

Torsion Angles (aka Phi-Psi Angles) - http://bioinformatics.org/molvis/phipsi/

- Looking along each bond of the protein backbone, we can describe a dihedral angle. (Considering 4 atoms at a time).

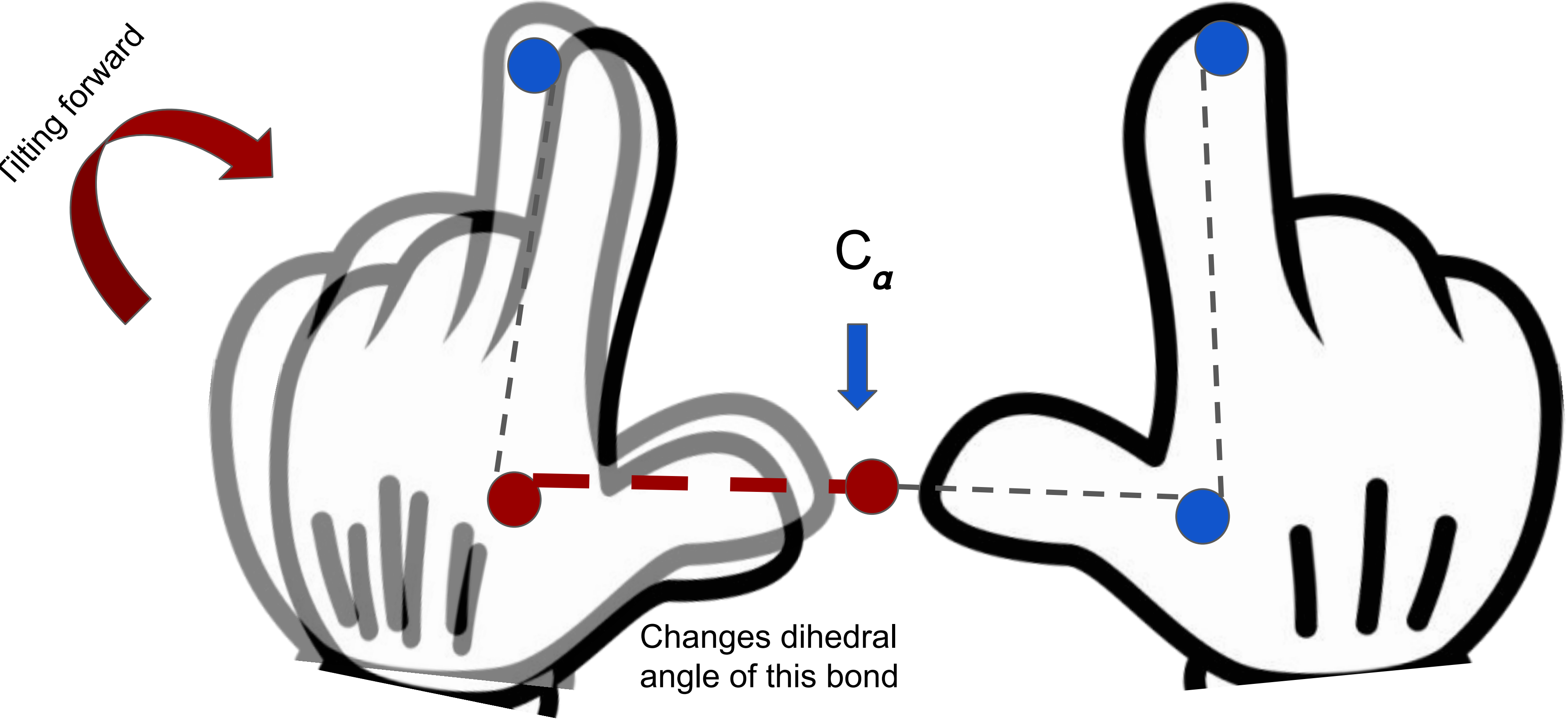
- There are 2 backbone bonds inside each AA, and then 1 peptide bond linking to the next AA.
    - Thus there are actually 3 dihedral angles to consider, but the peptide bond's dihedral angle ("Omega") is constrained by electron structure to be flat

Describing Protein Structure: Phi-Psi Angles



Changes dihedral angle of this bond

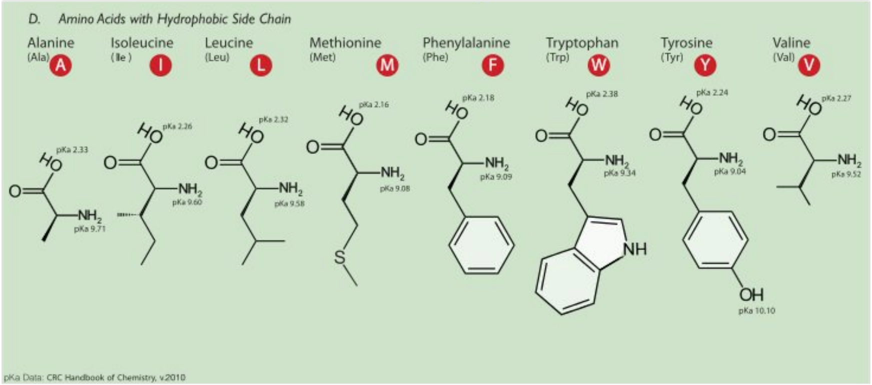
Describing Protein Structure: Phi-Psi Angles



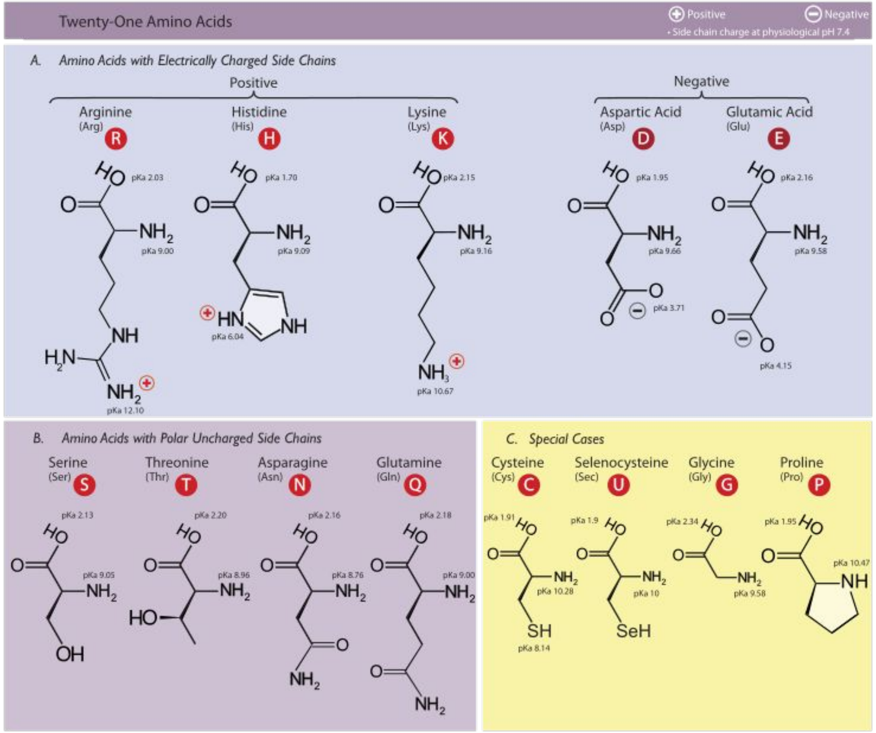
Changes dihedral angle of this bond

Amino Acid Structure

Side chains determine chemical properties

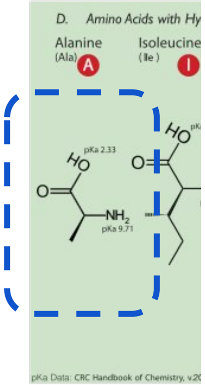


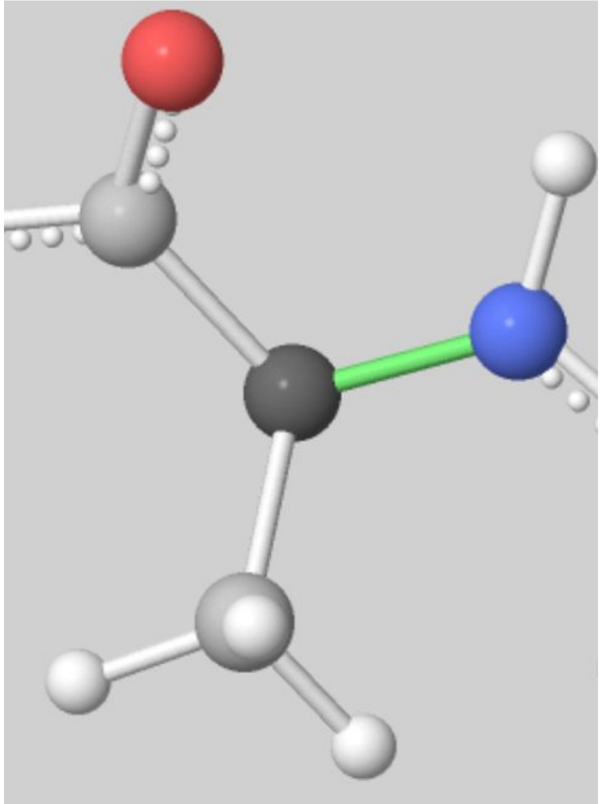
Dan Cojocari, Department of Medical Biophysics, **University of Toronto**, 2010

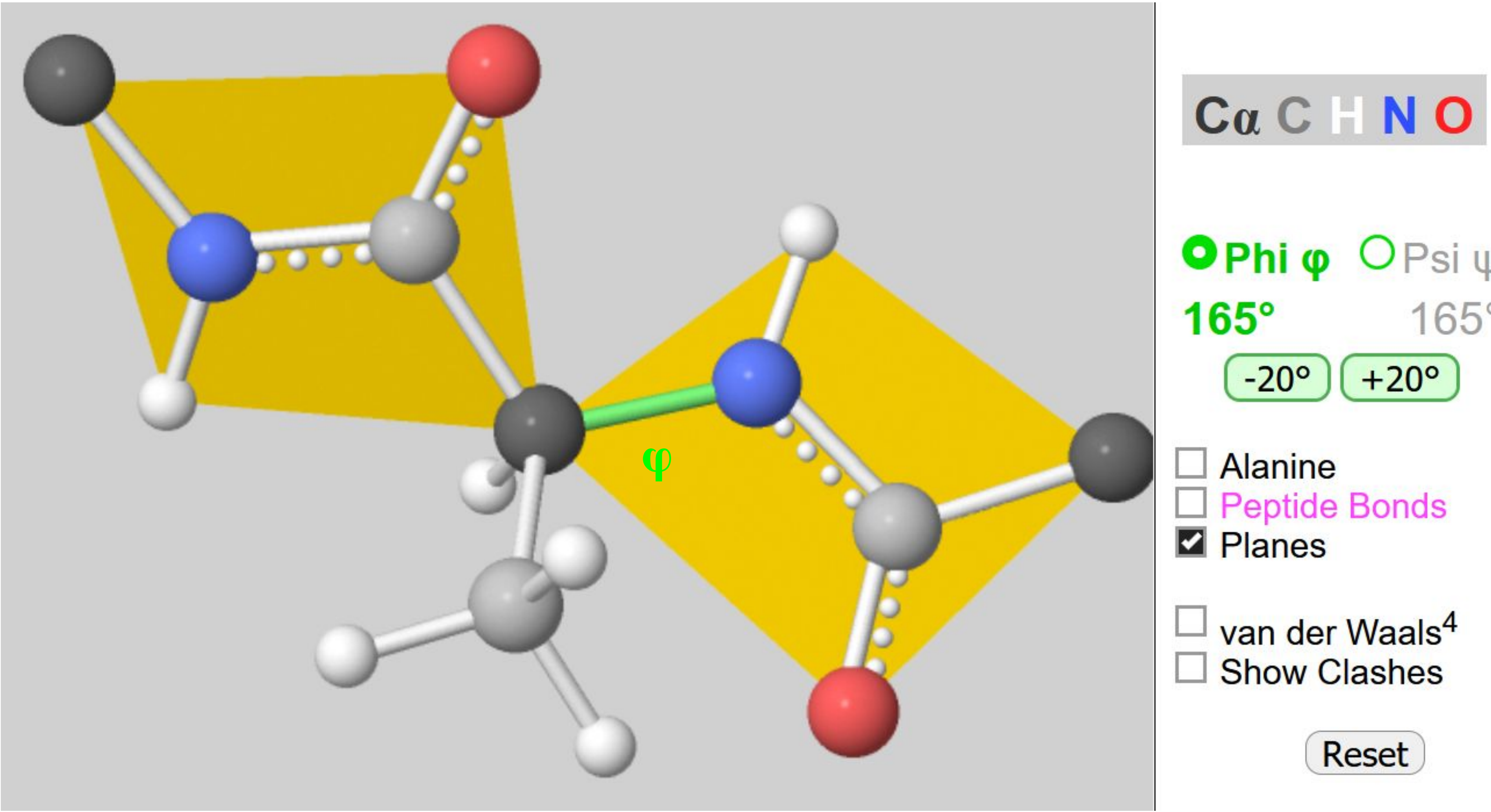


https://en.wikipedia.org/wiki/Proteinogenic_amino_acid

3D Amino Acid Structure







● Phi φ ○ Psi ψ

165° 165°

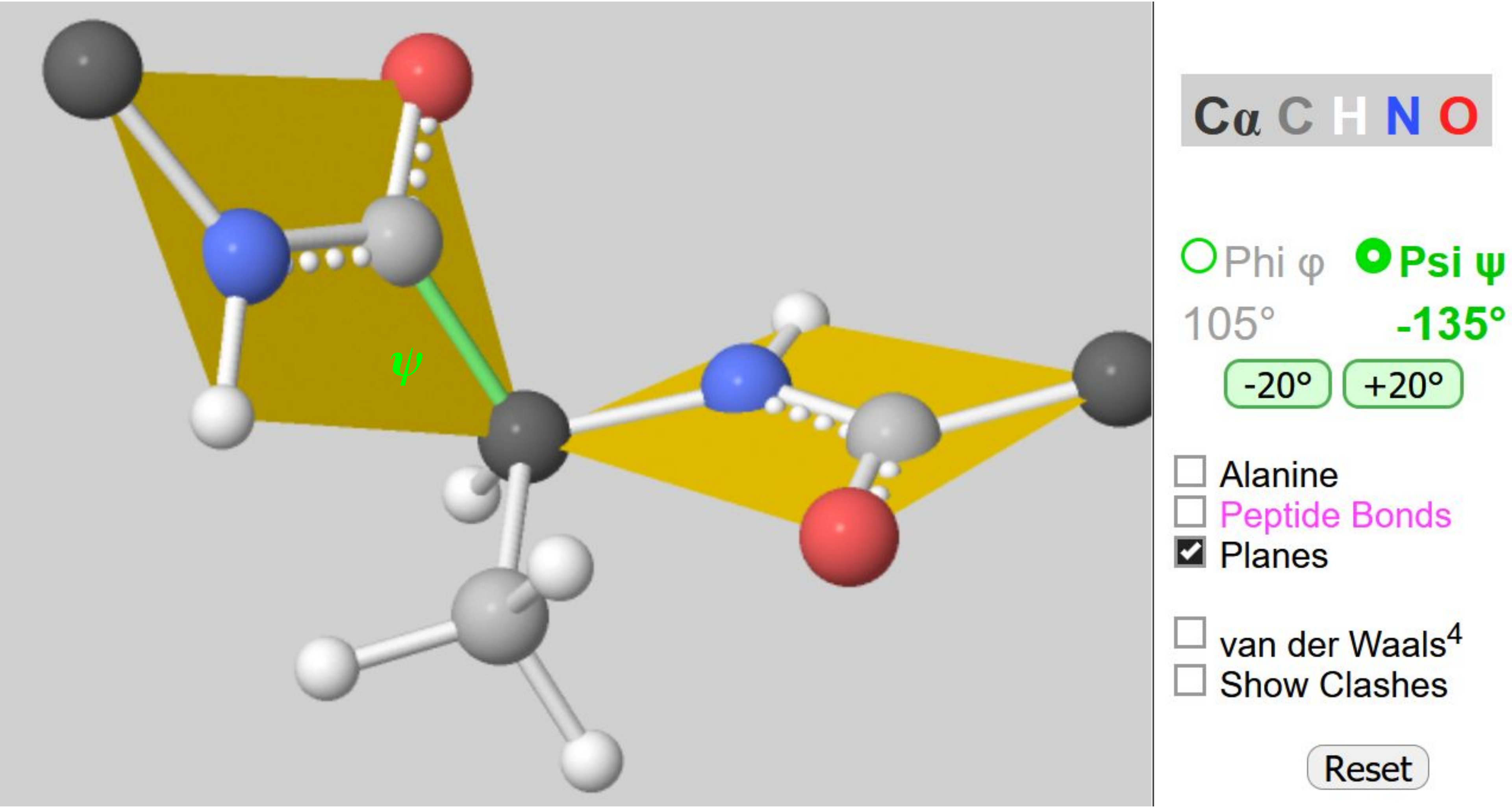
□ Alanine

□ Peptide Bonds

☑ Planes

□ van der Waals[4]

□ Show Clashes



○ Phi φ ● Psi ψ

105° -135°

□ Alanine

□ Peptide Bonds

☑ Planes

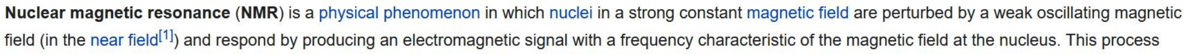
□ van der Waals[4]

□ Show Clashes

Experimental Methods for Finding Protein Structure

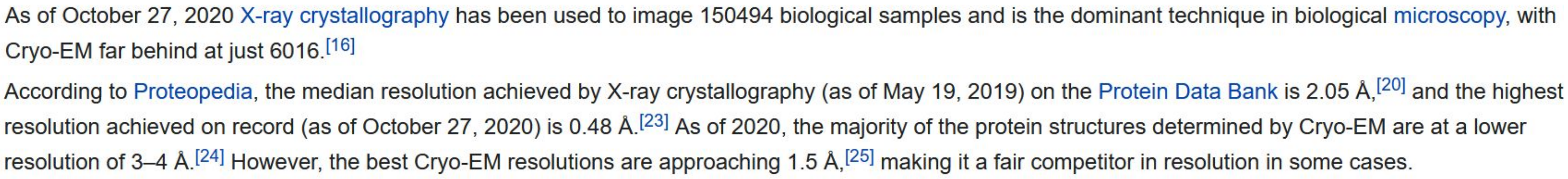
X-ray crystallography - requires crystallizing a protein (very hard, sometimes impossible, may alter protein shape, will only capture a single conformation)

Protein NMR - generate distance, angle, and orientation constraints using NMR



Nuclear magnetic resonance (**NMR**) is a physical phenomenon in which nuclei in a strong constant magnetic field are perturbed by a weak oscillating magnetic field (in the near field[1]) and respond by producing an electromagnetic signal with a frequency characteristic of the magnetic field at the nucleus. This process

Cryo Electron Microscopy - 2017 nobel prize in chemistry! Proteins in solution, even in motion… but lower resolution and not yet common:



As of October 27, 2020 X-ray crystallography has been used to image 150494 biological samples and is the dominant technique in biological microscopy, with Cryo-EM far behind at just 6016.[16]

According to Proteopedia, the median resolution achieved by X-ray crystallography (as of May 19, 2019) on the Protein Data Bank is 2.05 Å,[20] and the highest resolution achieved on record (as of October 27, 2020) is 0.48 Å.[23] As of 2020, the majority of the protein structures determined by Cryo-EM are at a lower resolution of 3–4 Å.[24] However, the best Cryo-EM resolutions are approaching 1.5 Å,[25] making it a fair competitor in resolution in some cases.

Background 2 of 2: CASP Competition

CASP - Critical Assessment of Structure Prediction

- Experimentalists submit experimental structures from upcoming publications, competitors predict structure from sequence.

https://predictioncenter.org/casp14/

CASP - Modeling Tasks

- Template-based Modeling (TBM)
    - Homologous protein domains (templates) can be used to guide prediction - we can find regions conserved across evolution and use their known structure in other contexts

- Free Modeling (FM)

- Contact Prediction
    - Two residues contact when predicted distance between beta carbons < threshold

- Biological Relevance

- Others

AlphaFold1

Extended Data Fig. 1 (Workflow)

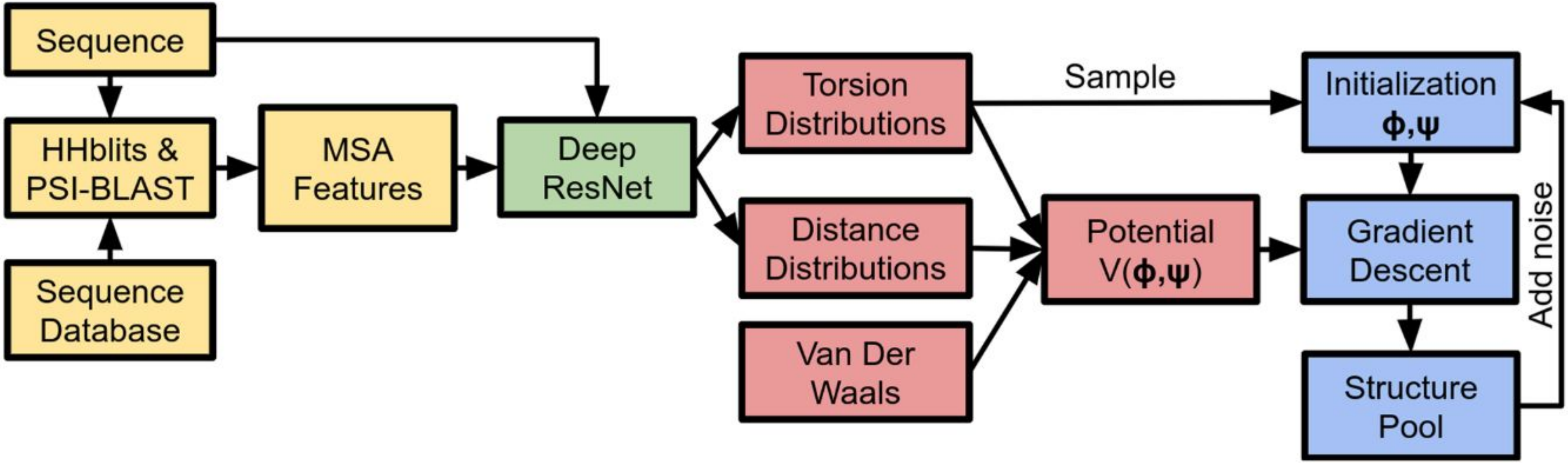


Fig. 2 (Workflow)

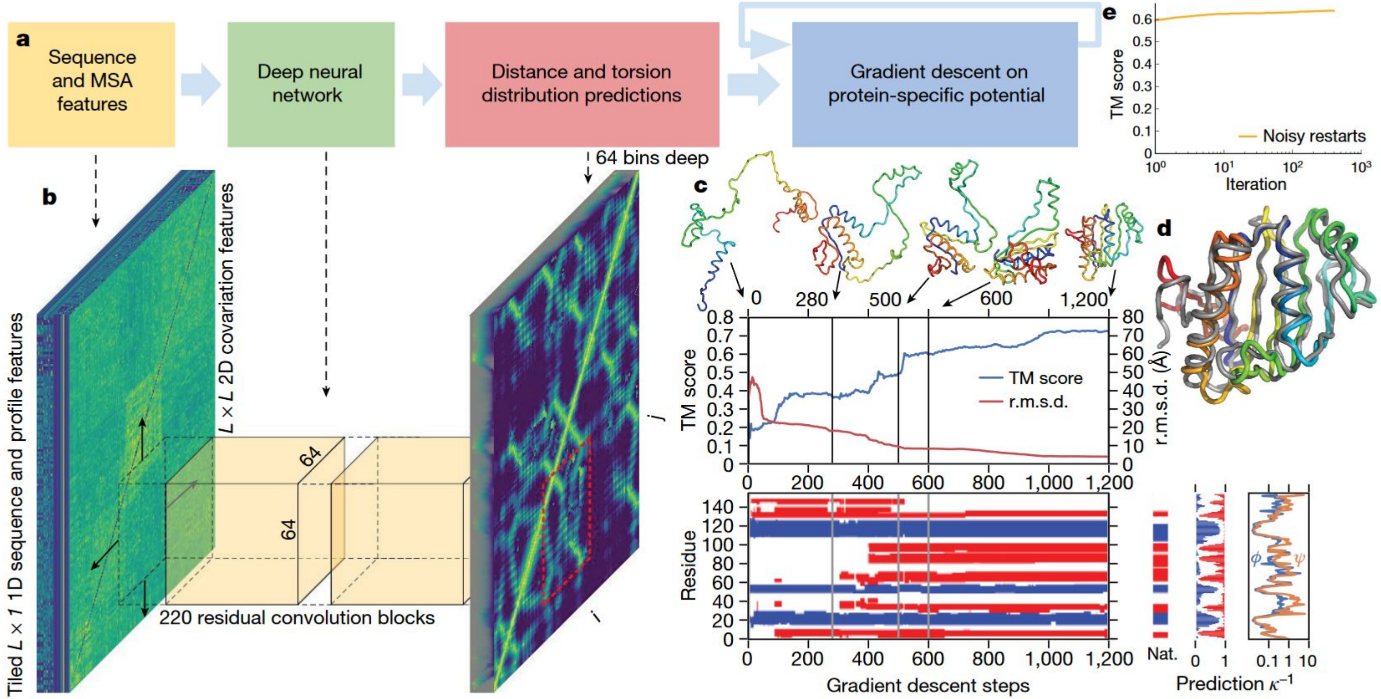
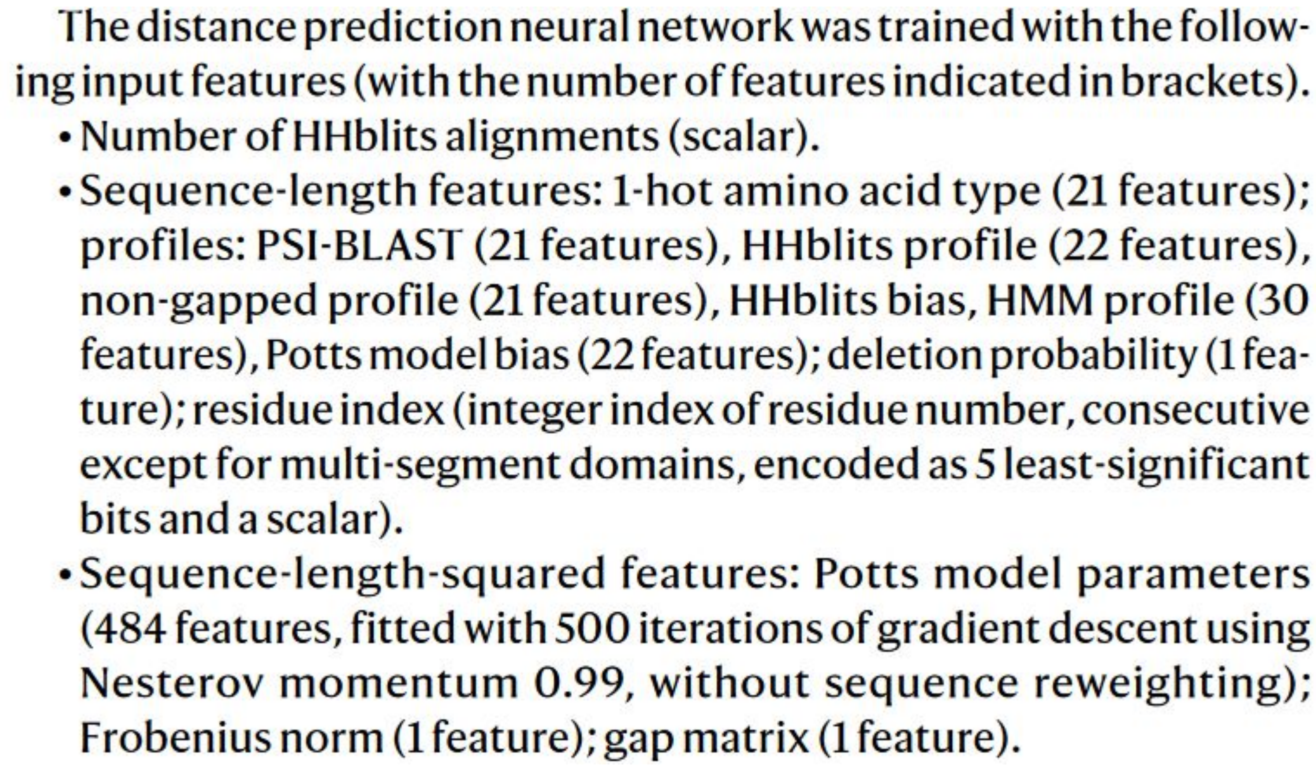


Fig. 2 | The folding process illustrated for CASP13 target T0986s2. CASP target T0986s2, $L = 155$, PDB: 6N9V. **a**, Steps of structure prediction. **b**, The neural network predicts the entire $L \times L$ distogram based on MSA features, accumulating separate predictions for 64 × 64-residue regions. **c**, One iteration of gradient descent (1,200 steps) is shown, with the TM score and root mean square deviation (r.m.s.d.) plotted against step number with five snapshots of the structure. The secondary structure (from SST[33]) is also shown (helix in blue, strand in red) along with the native secondary structure (Nat.), the secondary structure prediction probabilities of the network and the uncertainty in torsion angle predictions (as κ^{-1} of the von Mises distributions fitted to the predictions for φ and ψ). While each step of gradient descent greedily lowers the potential, large global conformation changes are effected, resulting in a well-packed chain. **d**, The final first submission overlaid on the native structure (in grey). **e**, The average (across the test set, $n = 377$) TM score of the lowest-potential structure against the number of repeats of gradient descent per target (log scale).

Input Features



The distance prediction neural network was trained with the following input features (with the number of features indicated in brackets).

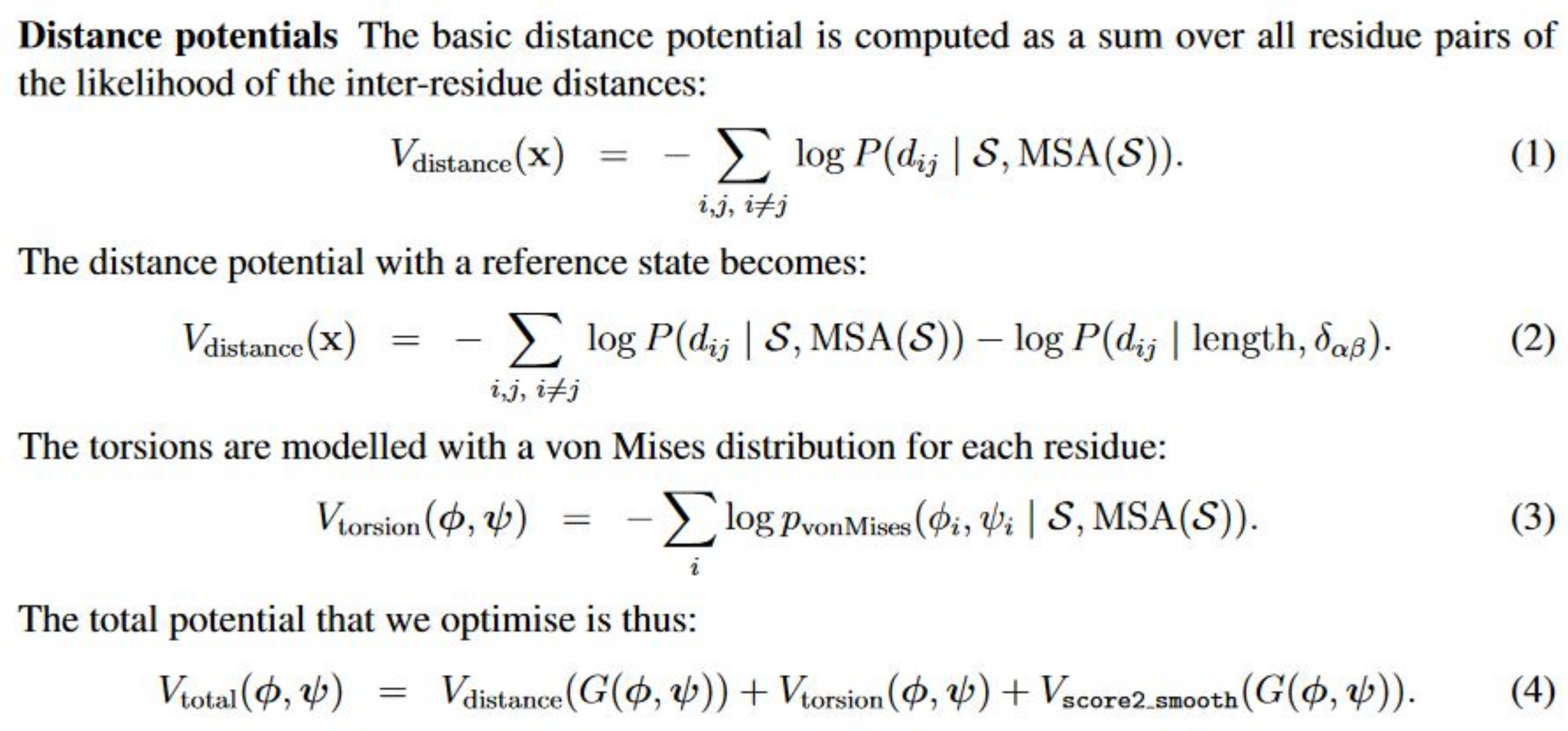
- Number of HHblits alignments (scalar).
- Sequence-length features: 1-hot amino acid type (21 features); profiles: PSI-BLAST (21 features), HHblits profile (22 features), non-gapped profile (21 features), HHblits bias, HMM profile (30 features), Potts model bias (22 features); deletion probability (1 feature); residue index (integer index of residue number, consecutive except for multi-segment domains, encoded as 5 least-significant bits and a scalar).
- Sequence-length-squared features: Potts model parameters (484 features, fitted with 500 iterations of gradient descent using Nesterov momentum 0.99, without sequence reweighting); Frobenius norm (1 feature); gap matrix (1 feature).

Potential Function

Recall from EBM: $p(x) = \exp(- E(x)) / Z$

Thus given $p(x)$, we can obtain: $E(x) = - \log p(x)$

Potential Function



Distance potentials The basic distance potential is computed as a sum over all residue pairs of the likelihood of the inter-residue distances:

$$V_{\text{distance}}(\mathbf{x}) \quad = \quad -\sum_{i,j,\; i \neq j} \log P(d_{ij} \mid \mathcal{S}, \text{MSA}(\mathcal{S})). \qquad (1)$$

The distance potential with a reference state becomes:

$$V_{\text{distance}}(\mathbf{x}) \quad = \quad -\sum_{i,j,\; i \neq j} \log P(d_{ij} \mid \mathcal{S}, \text{MSA}(\mathcal{S})) - \log P(d_{ij} \mid \text{length}, \delta_{\alpha\beta}). \qquad (2)$$

The torsions are modelled with a von Mises distribution for each residue:

$$V_{\text{torsion}}(\phi, \psi) \quad = \quad -\sum_{i} \log p_{\text{vonMises}}(\phi_i, \psi_i \mid \mathcal{S}, \text{MSA}(\mathcal{S})). \qquad (3)$$

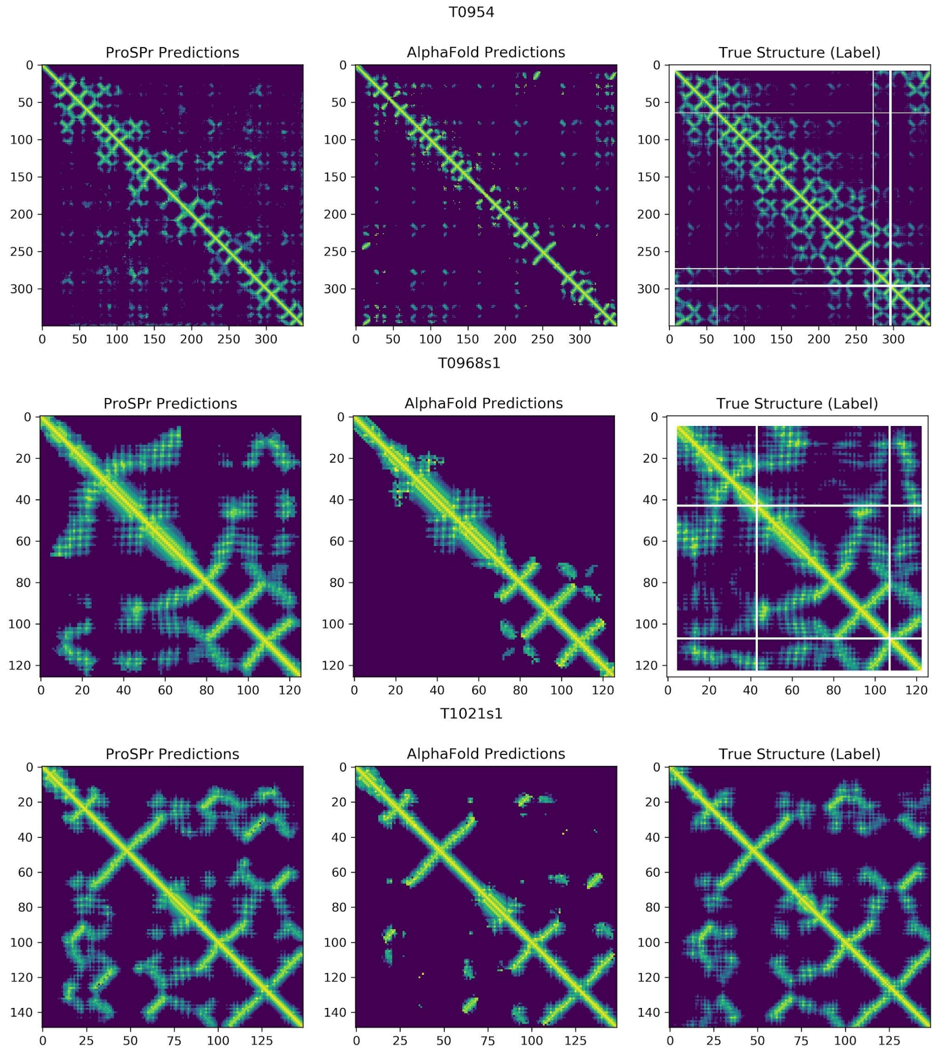
The total potential that we optimise is thus:

$$V_{\text{total}}(\phi, \psi) \quad = \quad V_{\text{distance}}(G(\phi, \psi)) + V_{\text{torsion}}(\phi, \psi) + V_{\texttt{score2\_smooth}}(G(\phi, \psi)). \qquad (4)$$

Distograms

$D_{ij} = \text{pred\_distance}(\text{residue}_i, \text{residue}_j)$

(bright means close)



https://github.com/dellacortelab/prospr

Fig. 1 (Compare to Other Methods)

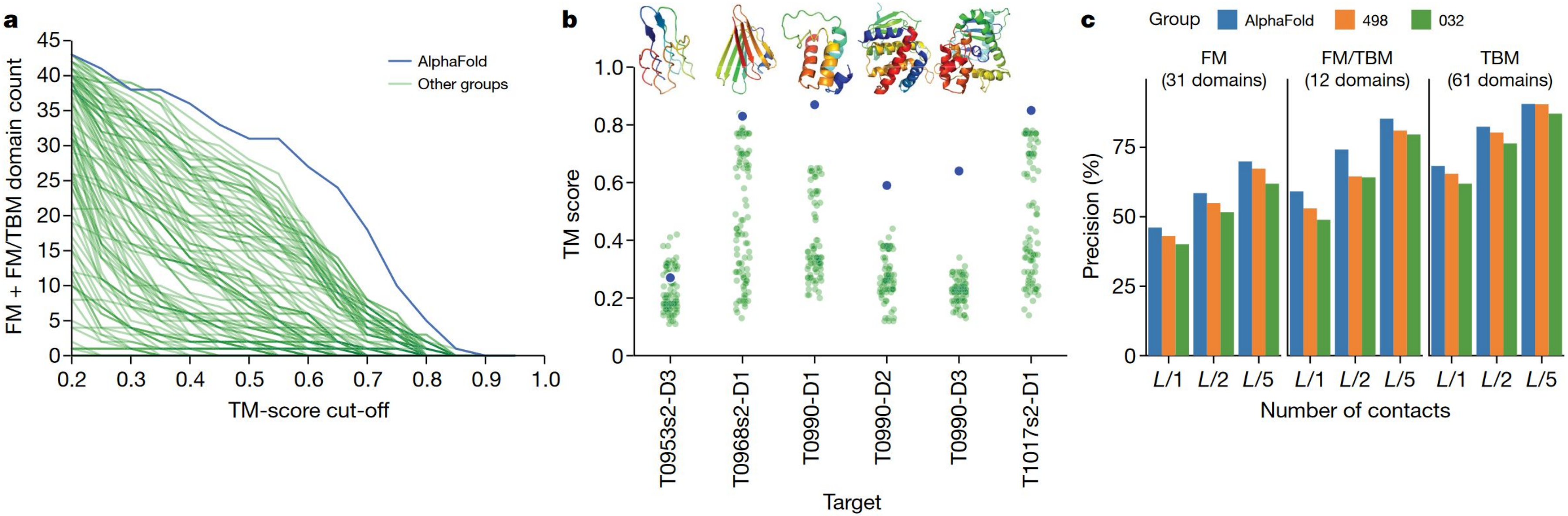


Fig. 1 | The performance of AlphaFold in the CASP13 assessment. a, Number of FM (FM + FM/TBM) domains predicted for a given TM-score threshold for AlphaFold and the other 97 groups. **b**, For the six new folds identified by the CASP13 assessors, the TM score of AlphaFold was compared with the other groups, together with the native structures. The structure of T1017s2-D1 is not available for publication. **c**, Precisions for long-range contact prediction in CASP13 for the most probable *L*, *L*/2 or *L*/5 contacts, where *L* is the length of the domain. The distance distributions used by AlphaFold in CASP13, thresholded to contact predictions, are compared with the submissions by the two best-ranked contact prediction methods in CASP13: 498 (RaptorX-Contact[26]) and 032 (TripletRes[32]) on 'all groups' targets, with updated domain definitions for T0953s2.

Fig. 3 (Compare to Ground Truth, and Uncertainty)

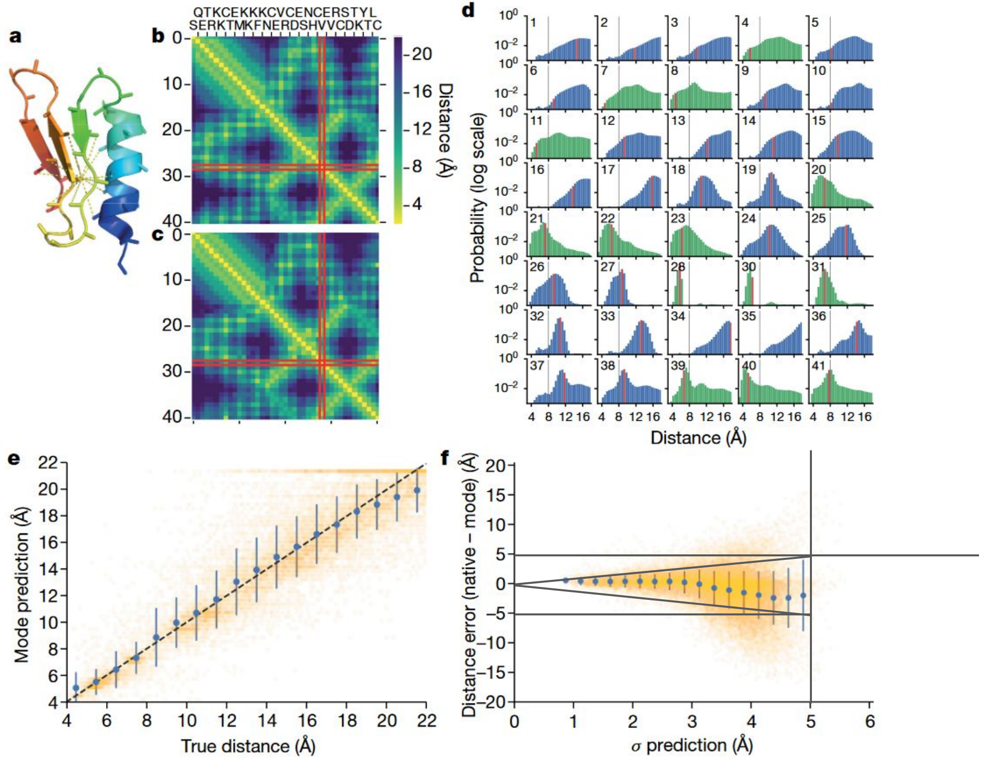


Fig. 3 | Predicted distance distributions compared with true distances. **a**–**d**, CASP target T0955, $L = 41$, PDB 5W9F. **a**, Native structure showing distances under 8 Å from the C_β of residue 29. **b**, **c**, Native inter-residue distances (**b**) and the mode of the distance predictions (**c**), highlighting residue 29. **d**, The predicted probability distributions for distances of residue 29 to all other residues. The bin corresponding to the native distance is highlighted in red, 8 Å is drawn in black. The distributions of the true contacts are plotted in green, non-contacts in blue. **e**, **f**, CASP target T0990, $L = 552$, PDB 6N9V. **e**, The mode of the predicted distance plotted against the true distance for all residue pairs with distances ≤22 Å, excluding distributions with s.d. > 3.5 Å ($n = 28{,}678$). Data are mean ± s.d. calculated for 1 Å bins. **f**, The error of the mode distance prediction versus the s.d. of the distance distributions, excluding pairs with native distances >22 Å ($n = 61{,}872$). Data are mean ± s.d. are shown for 0.25 Å bins. The true distance matrix and distogram for T0990 are shown in Extended Data Fig. 2b, c.

Fig. 4 (Ablation of Terms in Potential Function)

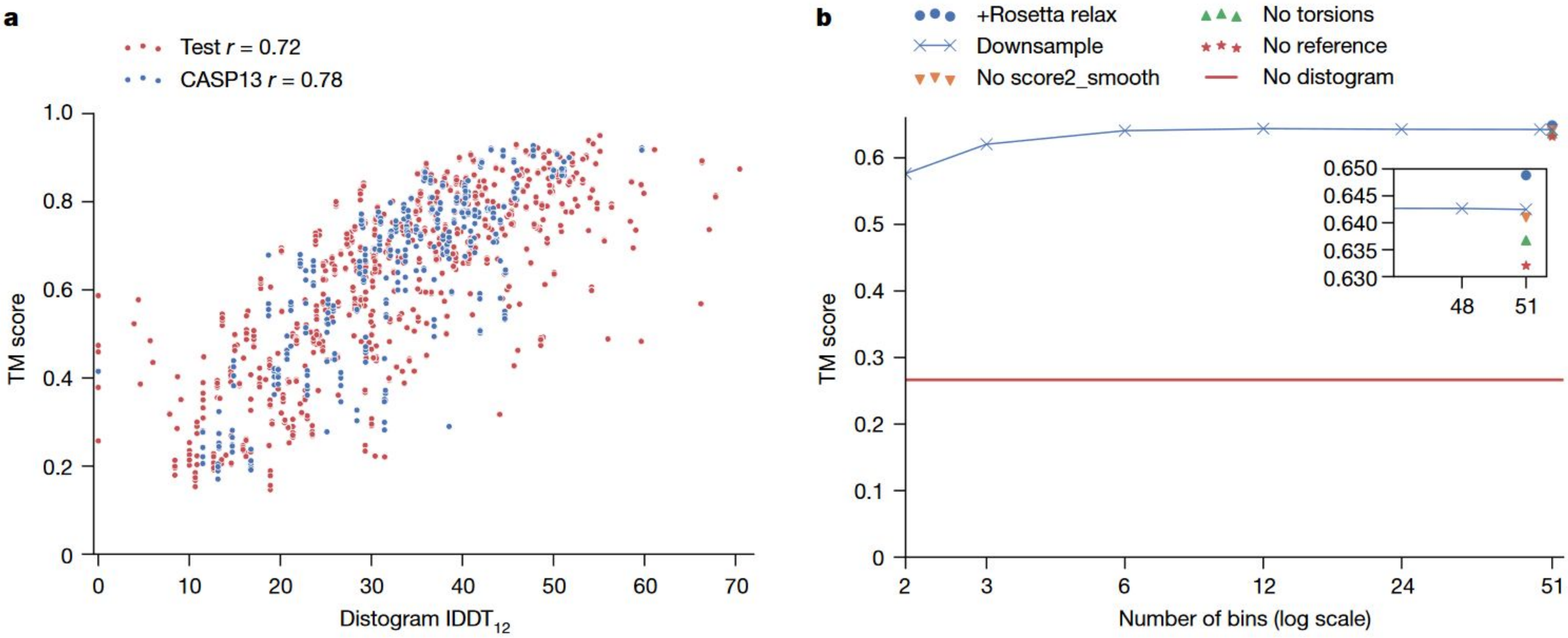


Fig. 4 | TM scores versus the accuracy of the distogram, and the dependency of the TM score on different components of the potential. a, TM score versus distogram lDDT$_{12}$ with Pearson's correlation coefficients, for both CASP13 ($n = 500$: 5 decoys for all domains, excluding T0999) and test ($n = 377$) datasets. **b**, Average TM score over the test set ($n = 377$) versus the number of histogram bins used when downsampling the distogram, compared with removing different components of the potential, or adding Rosetta relaxation.

Editorializing

Douglas Hofstadter's "Location of Meaning"

Hot take: Protein folding is an *ill-posed problem*.

In Gödel, Escher, Bach, Hofstadter asks: Can an alien who discovers a phonograph record space can ever hope to hear the music it contains?

- No; some of the information is contained in the record player!

Protein Folding in Biology

Proteins function by moving, changing conformation, associating/dis-associating with partners. (**They have multiple structures**)

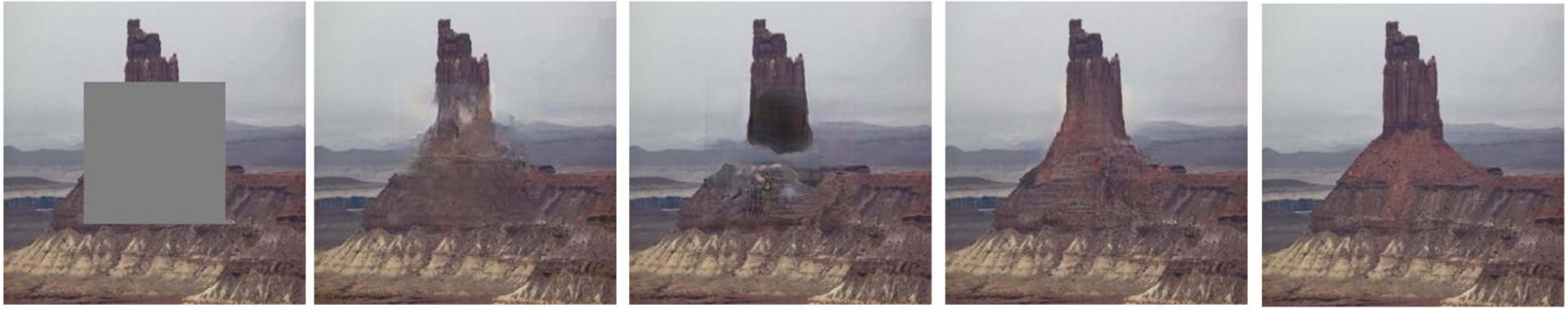
See Proteins that switch folds

Folding occurs in a 4D biochemical **context**, aided by chaperone proteins, in response to ligands/pH/solute concentrations, etc.

So structure is **mostly but not fully** determined by linear sequence. E.g. see

Protein Folding and Processing - The Cell - NCBI Bookshelf
Molecular chaperone functions in protein folding and proteostasis

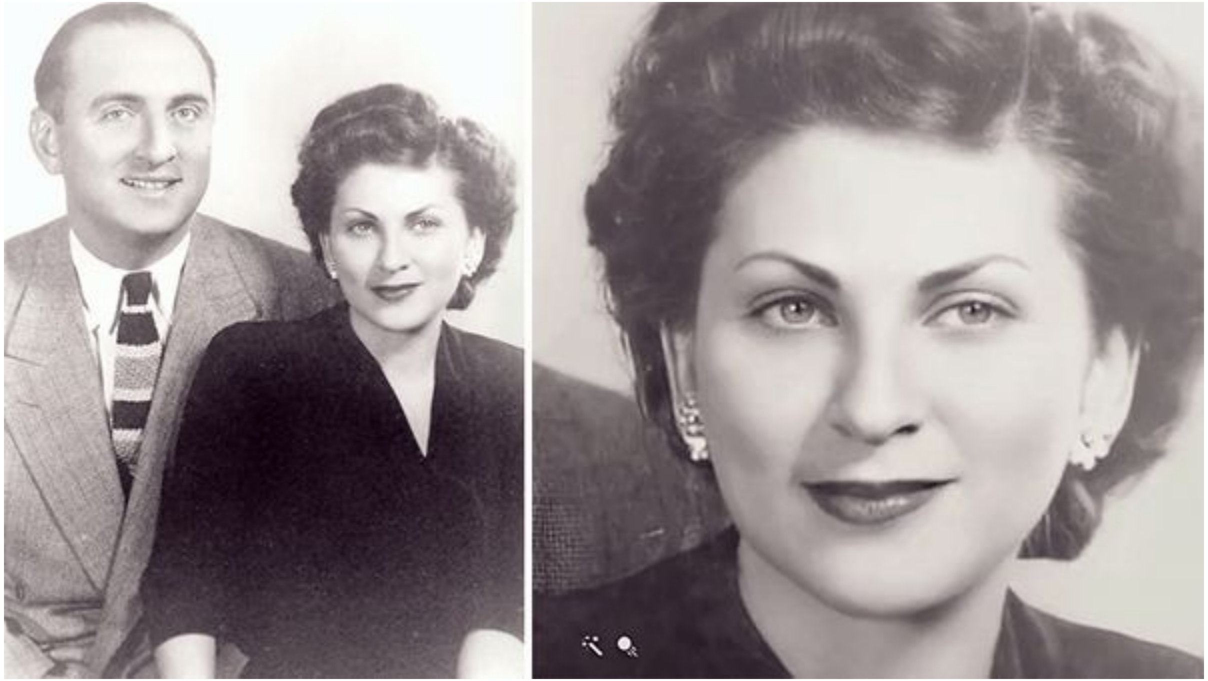
ML Loves Ill Posed Problems



Input (a) Shift-net (b) Contextual Attention (c) Our Result (d) Ground Truth (e)

https://arxiv.org/pdf/1905.12384v3.pdf

ML Loves Ill Posed Problems



https://i.redd.it/yygdsodyvwj61.gif

Links and Appendix

~~Open Source~~

https://github.com/deepmind/deepmind-research/tree/master/alphafold_casp13

"""This code can't be used to predict structure of an arbitrary protein sequence. It can be used to predict structure only on the CASP13 dataset (links below). The feature generation code is tightly coupled to our internal infrastructure as well as external tools, **hence we are unable to open-source it**."""

Open Source

ProSPr: Democratized Implementation of Alphafold Protein Distance Prediction Network - https://www.biorxiv.org/content/10.1101/830273v2

- https://github.com/dellacortelab/prospr

- """This repository currently contains a democratized implementation of the **AlphaFold1** distance prediction network."""

Also: https://github.com/Urinx/alphafold_pytorch

AlphaFold2 (at CASP14)

https://deepmind.com/blog/article/alphafold-a-solution-to-a-50-year-old-grand-challenge-in-biology

"""It uses approximately 16 TPUv3s (which is 128 TPUv3 cores or roughly equivalent to ~100-200 GPUs) run over a few weeks, **a relatively modest amount of compute** in the context of most large state-of-the-art models used in machine learning today. """

Deepmind Slides: https://predictioncenter.org/casp14/doc/presentations/2020_12_01_TS_predictor_AlphaFold2.pdf

Fabian Fuchs writeup: AlphaFold 2 & Equivariance (Good find by Robin!)