Melbourne Bioinformatics

BIOINFORMATICS + DATA SERVICES + INFRASTRUCTURE, FOR LIFE SCIENCES TODAY

THE UNIVERSITY OF MELBOURNE

COMP90014

Algorithms for Bioinformatics

Week 6A - Genomic Features & Regions

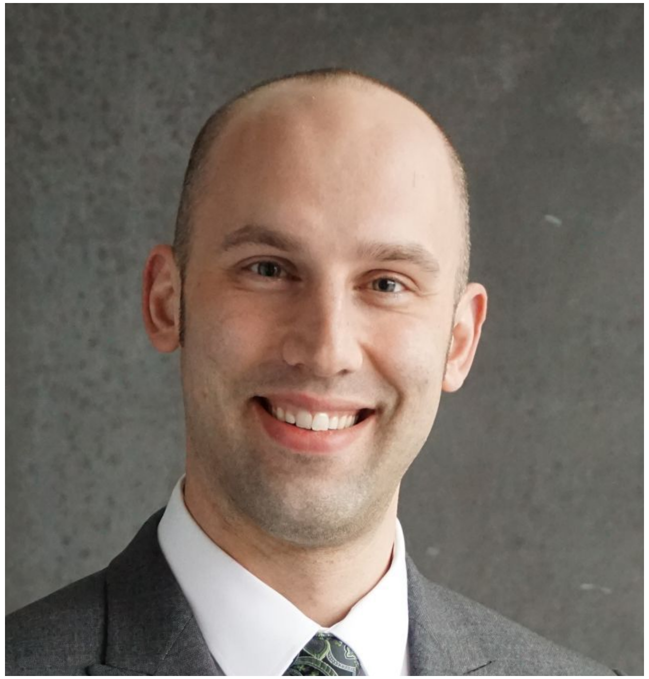
Assignment 2 Released Tomorrow

(Wednesday 30th @ midnight)

Upcoming Guest Speakers!

Adam Taranto

This Thursday! (31st August)



Ryan Wick

Next Tuesday! (5th September)

Genomic Features & Regions

Introduction

Leveraging Data

Prokka: Annotation Pipeline

Training Based Approaches

- Markov Models
- Deep Learning: AlphaFold 2.0

Introduction

Introduction

Genome assembly just the beginning!

- Want to understand genomic features / regions

Introduction

在幻灯片的上半部分，提到“ 基因组组装只是开始 ”，这意味着完成一个物种基因组的测序和组装后，科学家的工作远未结束。接下来的目标是“ 理解基因组特征/区域 ”，也就是要弄清楚基因组中各个部分的功能和意义。

基因组注释是一个重要的后续步骤，定义为“ 在基因组序列上识别并标记相关特征 ”。这些特征包括：

Genome assembly just the beginning!

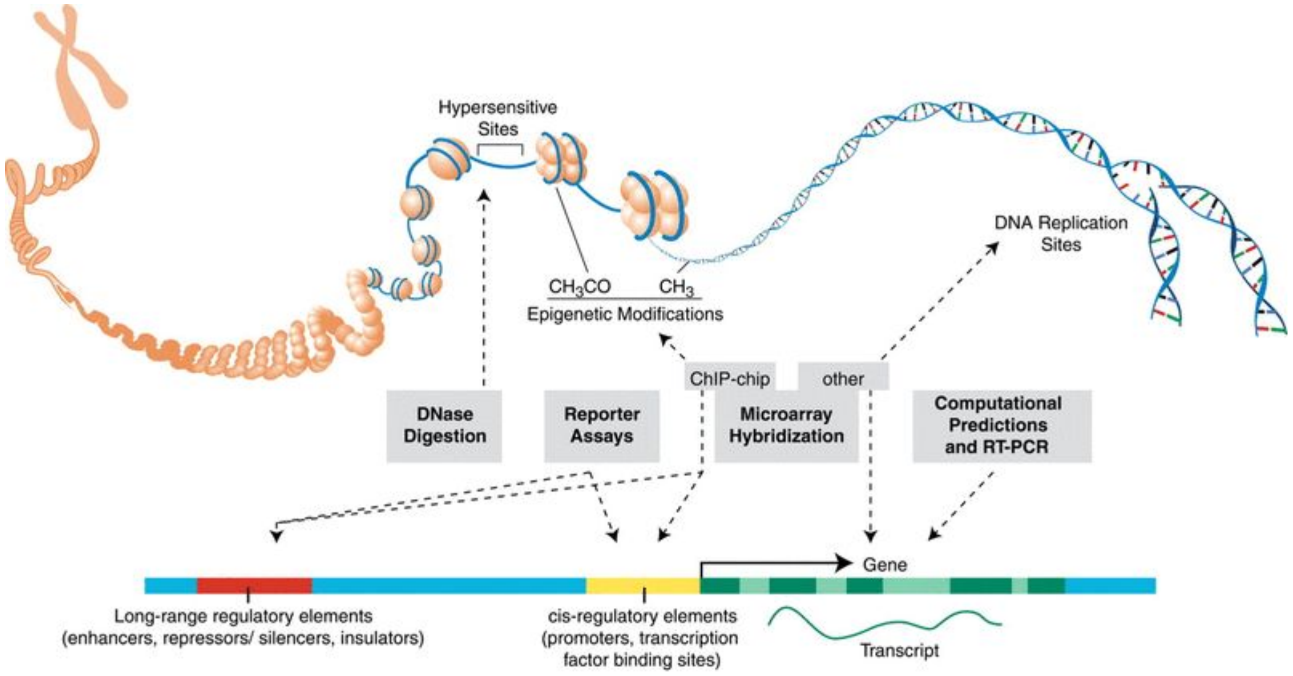
Want to understand genomic features / regions

Genome annotation

“Identifying and labeling the relevant features on a genome sequence”

Coding genes and predicted AA products (most important)

Promoter / Enhancers, Non-coding RNAs
Signal Peptides, etc
(also important)


Hypersensitive Sites
CH_3CO
CH_3
Epigenetic Modifications
DNA Replication Sites
ChIP-chip
other
DNase Digestion
Reporter Assays
Microarray Hybridization
Computational Predictions and RT-PCR
Gene
Long-range regulatory elements (enhancers, repressors/ silencers, insulators)
cis-regulatory elements (promoters, transcription factor binding sites)
Transcript

The ENCODE (ENCyclopedia Of DNA Elements) Project. DOI:10.1126/science.1105136

Genomic Features & Regions

Too broad to cover implementations

- Repeat masking
- Identifying promotors / enhancers
- Identifying signal peptides (SPs)
- Identifying genes
- Gene annotation
- Assembly QC: BUSCO
- Identifying structural variation
- Identifying sequence variation
- Characterisation

Will instead cover

- Some main ideas
- Leveraging Data
- Prokka: Annotation Pipeline (main ideas)
- Training Based Approaches
- Markov Models
- Deep Learning: AlphaFold 2.0

Leveraging Data

Leveraging Data

Over last 50 years, accumulated lots of data!

This data can be helpful for tasks we wish to perform.

- Some tasks purely isolated
- Some tasks partially leverage existing data
- Some tasks fully use existing data

Leveraging Data

Over last 50 years, accumulated lots of data!

This data can be helpful for tasks we wish to perform.

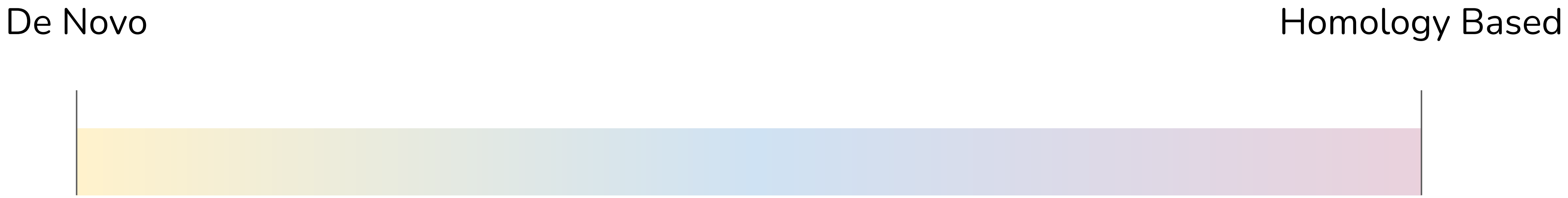
- Some tasks purely isolated
- Some tasks partially leverage existing data
- Some tasks fully use existing data

Leveraging data is a spectrum.

For a given task, often have different software / algorithms which sit at different points along the spectrum.

These perform better / worse for a particular set of input data. Why?

幻灯片底部有一个从“De Novo”（从头开始）到“Homology Based”（同源性基础）的渐变条，暗示了任务执行可以从完全不依赖现有数据到完全依赖与其他数据的相似性。
“De Novo”（从头开始）意味着在没有任何先前知识的情况下构建数据或模型。而“Homology Based”（基于同源性）意味着使用已知的数据来预测或理解未知的数据，这是通过比较它们与已知数据的相似性来实现的。



De Novo

Existing data not used.

Existing data:

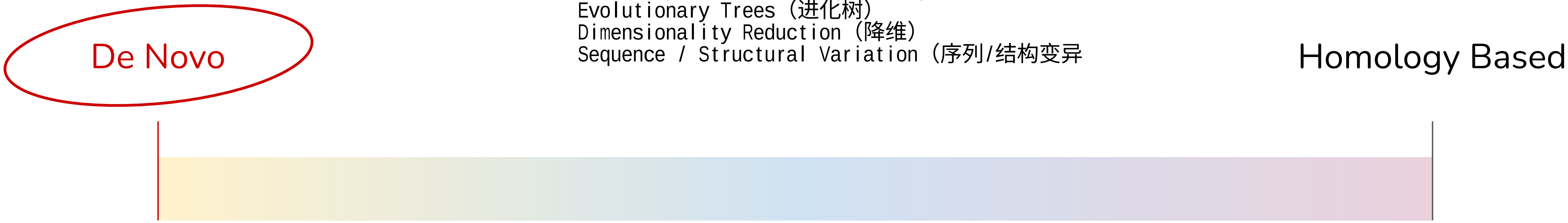
- May not be applicable, or
- May not help

- Indexing & Alignment
- Genome Assembly (De Novo)
- Transcriptome Assembly (De Novo)
- Evolutionary Trees
- Dimensionality Reduction
- Sequence / Structural Variation

不使用现有数据。
现有数据可能不适用，或者可能没有帮助。

幻灯片的右侧列出了与“De Novo”方法相关的几个任务或过程：

Indexing & Alignment（索引和比对）
Genome Assembly (De Novo)（基因组组装，从头开始）
Transcriptome Assembly (De Novo)（转录组组装，从头开始）
Evolutionary Trees（进化树）
Dimensionality Reduction（降维）
Sequence / Structural Variation（序列/结构变异



Homology Based

Existing data used from large sequence databases.

Far end of spectrum:

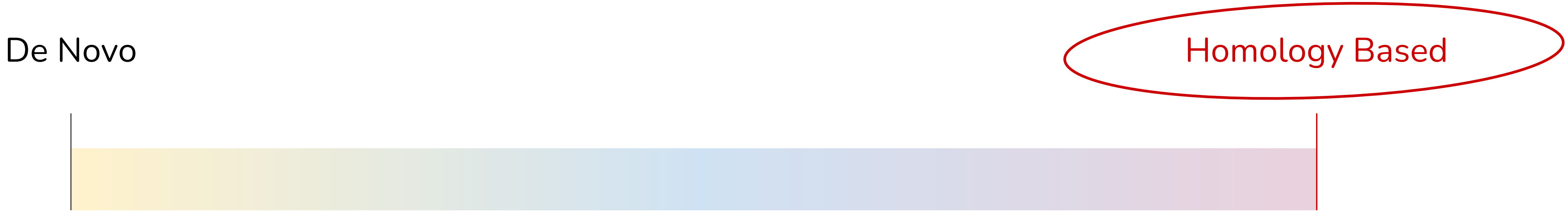
- Existing data fully answers our question.

Depending on task, may require strong evidence:

- Large set of corroborating data
- Experimentally validated data

For sequences: Homology

- Characterisation (MetaPhLan)
- Gene Annotation (BLAST)
- Identifying Homologues
- Conservation



Training Based

Existing data can be used as model.

Combines existing data with new data.

Statistical approaches (often machine learning)

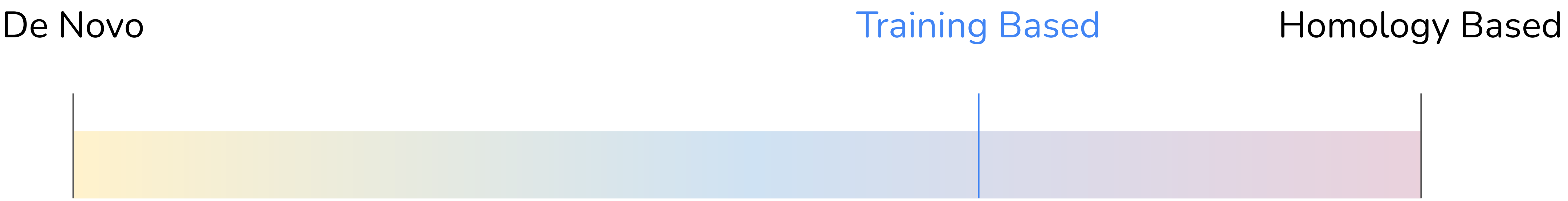
Large group.

使用来自大型序列数据库的现有数据。
在谱系的远端，现有数据能完全回答我们的问题。
根据任务的不同，可能需要强有力的证据：
大量支持的数据
实验验证的数据

- Repeat Masking
- Signal Peptide Identification
- Gene Identification
- Gene Annotation
- Variant Effect Prediction

幻灯片的下半部分列出了一些具体的基于同源性的序列分析任务：

Characterisation (MetaPhlAn)（特征化，例如使用MetaPhlAn软件）
Gene Annotation (BLAST)（基因注释，例如使用BLAST工具）
Identifying Homologues（识别同源基因）
Conservation（保存性分析）



Leveraging Data

Task	De Novo	Training Based	Homology Based
Genome QC	✖	✖	✓
Repeat Masking	✓	✓	✖
Gene Identification	✖	✓	✖
Gene Annotation			
Promoter / Enhancer Identification	✖	✓	✓
Signal Peptide Identification	✖	✓	✓
Sequence / Structural Variation	✓	✓	✖
Variant Effect Prediction	✓	✓	✖
Characterisation	✖	✖	✓

Leveraging Data

Task	De Novo	Training Based	Homology Based
Genome QC	✖	✖	✓
Repeat Masking	✓	✓	✖
Gene Identification	✖	✓	✖
Gene Annotation	✖		
Promoter / Enhancer Identification	✖	✓	✓
Signal Peptide Identification	✖	✓	✓
Sequence / Structural Variation	✓	✓	✖
Variant Effect Prediction	✓	✓	✖
Characterisation	✖	✖	✓

Leveraging Data

Task	De Novo	Training Based	Homology Based
Genome QC	✖	✖	✓
Repeat Masking	✓	✓	✖
Gene Identification	✖	✓	✖
Gene Annotation	✖	✓	
Promoter / Enhancer Identification	✖	✓	✓
Signal Peptide Identification	✖	✓	✓
Sequence / Structural Variation	✓	✓	✖
Variant Effect Prediction	✓	✓	✖
Characterisation	✖	✖	✓

Leveraging Data

Task	De Novo	Training Based	Homology Based
Genome QC	✖	✖	✓
Repeat Masking	✓	✓	✖
Gene Identification	✖	✓	✖
Gene Annotation	✖	✓	✓
Promoter / Enhancer Identification	✖	✓	✓
Signal Peptide Identification	✖	✓	✓
Sequence / Structural Variation	✓	✓	✖
Variant Effect Prediction	✓	✓	✖
Characterisation	✖	✖	✓

Leveraging Data

Task	De Novo	Training Based	Homology Based
Genome QC	✖	✖	✓
Repeat Masking	✓	✓	✖
Gene Identification	✖	✓	✖
Gene Annotation	✖	✓	✓
Promoter / Enhancer Identification	✖	✓	✓
Signal Peptide Identification	✖	✓	✓
Sequence / Structural Variation	✓	✓	✖
Variant Effect Prediction	✓	✓	✖
Characterisation	✖	✖	✓

Particular software may perform better / worse for a particular set of input data. Why?

Prokka: Annotation Pipeline

Prokka

Prokka 是一个软件管道（pipeline），用于自动化原核生物基因组的注释过程，并整合了多个现有的生物信息学工具。

输入：以FASTA格式提供的DNA序列（contigs/ scaffolds）。理想情况下是每条染色体或质粒一个序列，且没有间隙。但实际上，可能会是片段化且存在间隙的序列。
输出：带有起始/停止位置的注释特征。注释过程需要识别并注释这些特征

Prokka: Torsten Seemann

Prokaryote genome annotation pipeline.
Coordinates existing tools in a pipeline.

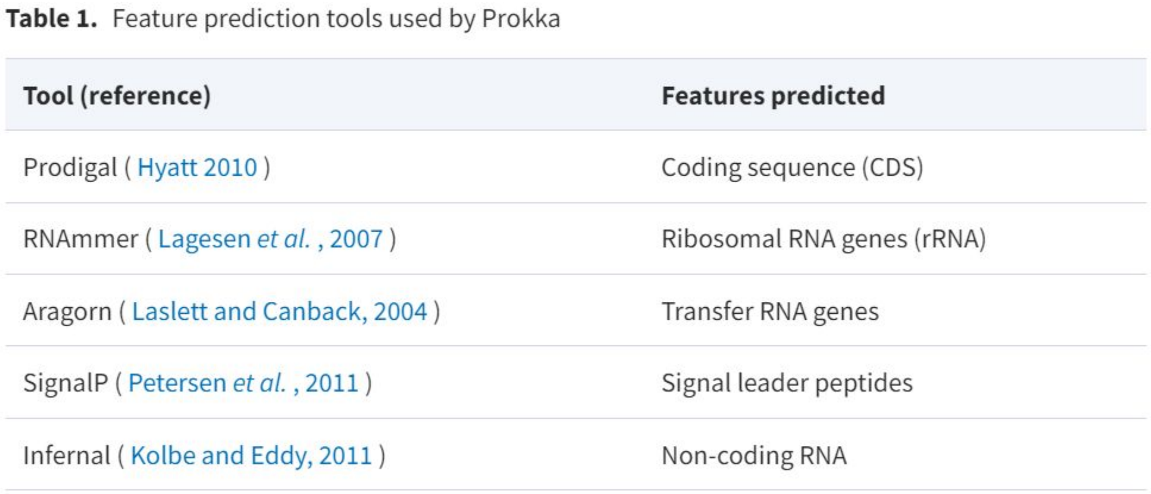


Table 1. Feature prediction tools used by Prokka

Tool (reference)	Features predicted
Prodigal (Hyatt 2010)	Coding sequence (CDS)
RNAmmer (Lagesen *et al.* , 2007)	Ribosomal RNA genes (rRNA)
Aragorn (Laslett and Canback, 2004)	Transfer RNA genes
SignalP (Petersen *et al.* , 2011)	Signal leader peptides
Infernal (Kolbe and Eddy, 2011)	Non-coding RNA

Input

DNA sequences in Fasta format (contigs / scaffolds)
Ideal: 1 sequence per chromosome / plasmid, no gaps
Real: Fragmented, gaps

Output:

Annotated features with start / stop positions
Need to both identify, and annotate these features.

Prokka

Prokka’s use of data: It’s all about trust.

Prokka

Prokka’s use of data: It’s all about trust.

Hierarchical approach (coding sequence annotation):

1. Optional user-provide set of annotated proteins.
   Expected to be curated datasets - high confidence

Prokka

Prokka’s use of data: It’s all about trust.

Hierarchical approach (coding sequence annotation):

1. Optional user-provide set of annotated proteins.
   Expected to be curated datasets - high confidence

2. All bacterial proteins in UniProt that have real protein or transcript evidence
   Experimentally validated - high confidence
   Typically covers >50% of the core genes in most genomes

Prokka

Prokka's use of data: It's all about trust.

Hierarchical approach (coding sequence annotation):

1. Optional user-provide set of annotated proteins.
   Expected to be curated datasets - high confidence

2. All bacterial proteins in UniProt that have real protein or transcript evidence
   Experimentally validated - high confidence
   Typically covers >50% of the core genes in most genomes

3. All proteins from finished bacterial genomes in RefSeq for a specified genus
   Proteins this organism is reasonably likely to possess

Prokka

Prokka’s use of data: It’s all about trust.

Hierarchical approach (coding sequence annotation):

1. Optional user-provide set of annotated proteins.
   Expected to be curated datasets - high confidence
2. All bacterial proteins in UniProt that have real protein or transcript evidence
   Experimentally validated - high confidence
   Typically covers >50% of the core genes in most genomes
3. All proteins from finished bacterial genomes in RefSeq for a specified genus
   Proteins this organism is reasonably likely to possess
4. A series of hidden Markov model profile databases
   “Putative / predicted” proteins

强调了“信任”在注释数据中的重要性。幻灯片描述了Prokka在进行编码序列注释时采用的分层方法：

用户提供的注释蛋白集（可选）：预期这些数据集是经过策展的，因此具有高度的可信度。

UniProt数据库中所有细菌蛋白：这些是有真实蛋白或转录本证据支持的数据，经过实验验证，可信度高。通常这些蛋白能覆盖大多数基因组中超过50%的核心基因。

RefSeq数据库中已完成细菌基因组的所有蛋白：针对指定属的细菌，这些蛋白被认为该细菌种类有合理的可能性拥有。

使用隐马尔可夫模型（Hidden Markov Model，HMM）概况数据库来识别一系列的“推定/预测”蛋白质：这种方法可以帮助预测那些尚未在实验中得到证实的蛋白质。

总的来说，这张幻灯片概述了Prokka注释编码序列时所依赖的数据来源和策略，重点是使用高可信度的数据源以确保注释结果的准确性。

Training Based Approaches

Training Based Approaches

Homology approaches

- Rely on relevant & high confidence data.
- What if your organism is too different from those in databases?

Training Based Approaches

Homology approaches

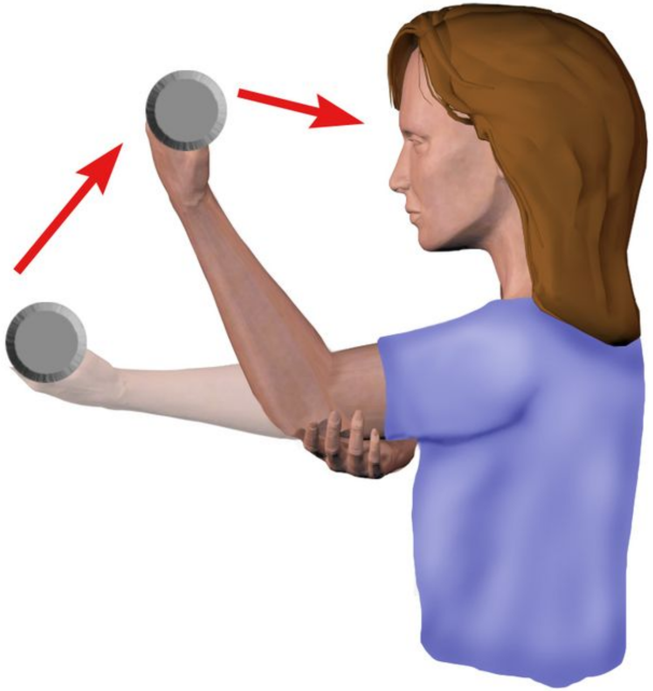
Rely on relevant & high confidence data.

What if your organism is too different from those in databases?

Training based approaches

If no direct comparisons can be made, can turn to generalised features / patterns.

Underlying structure.

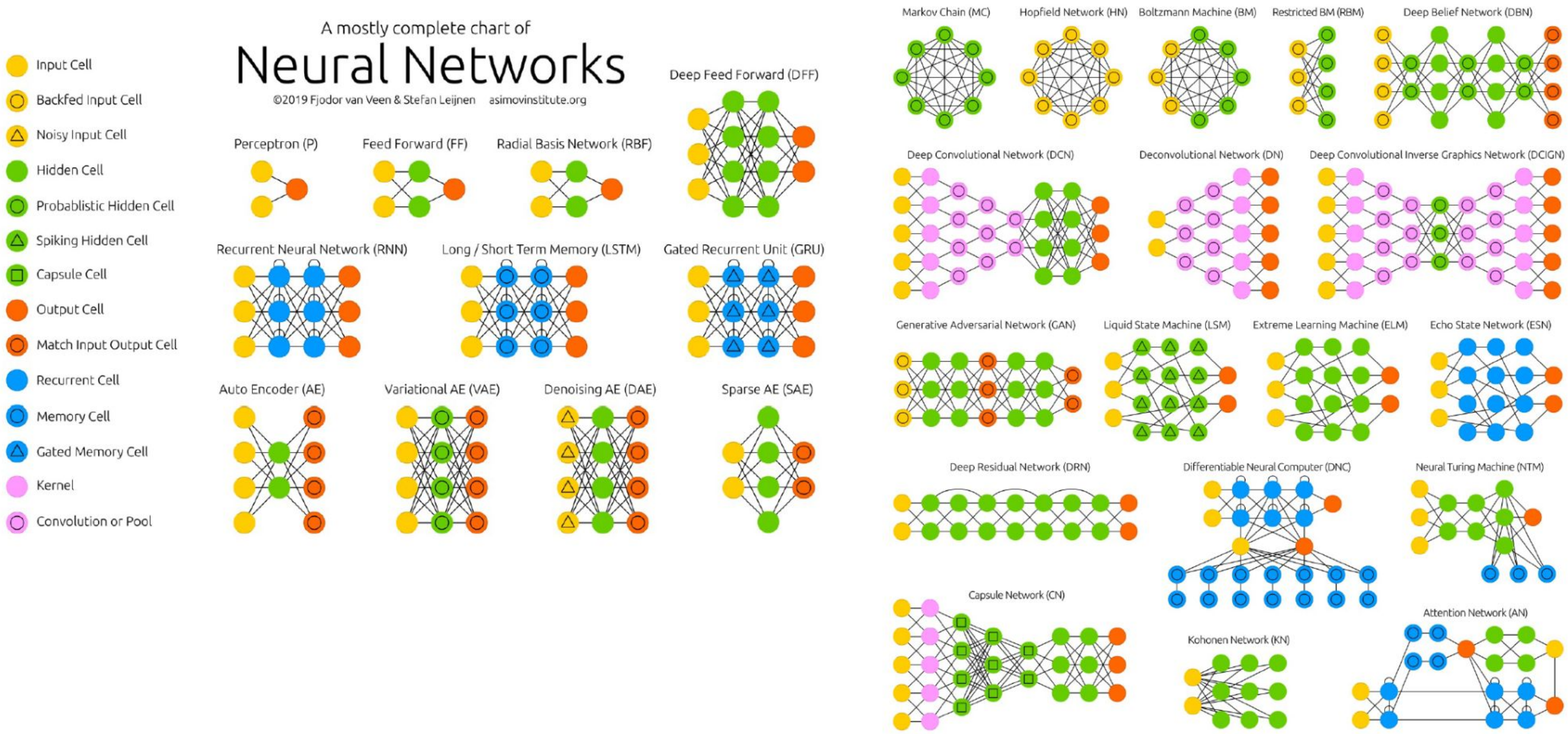


Bicep Curls

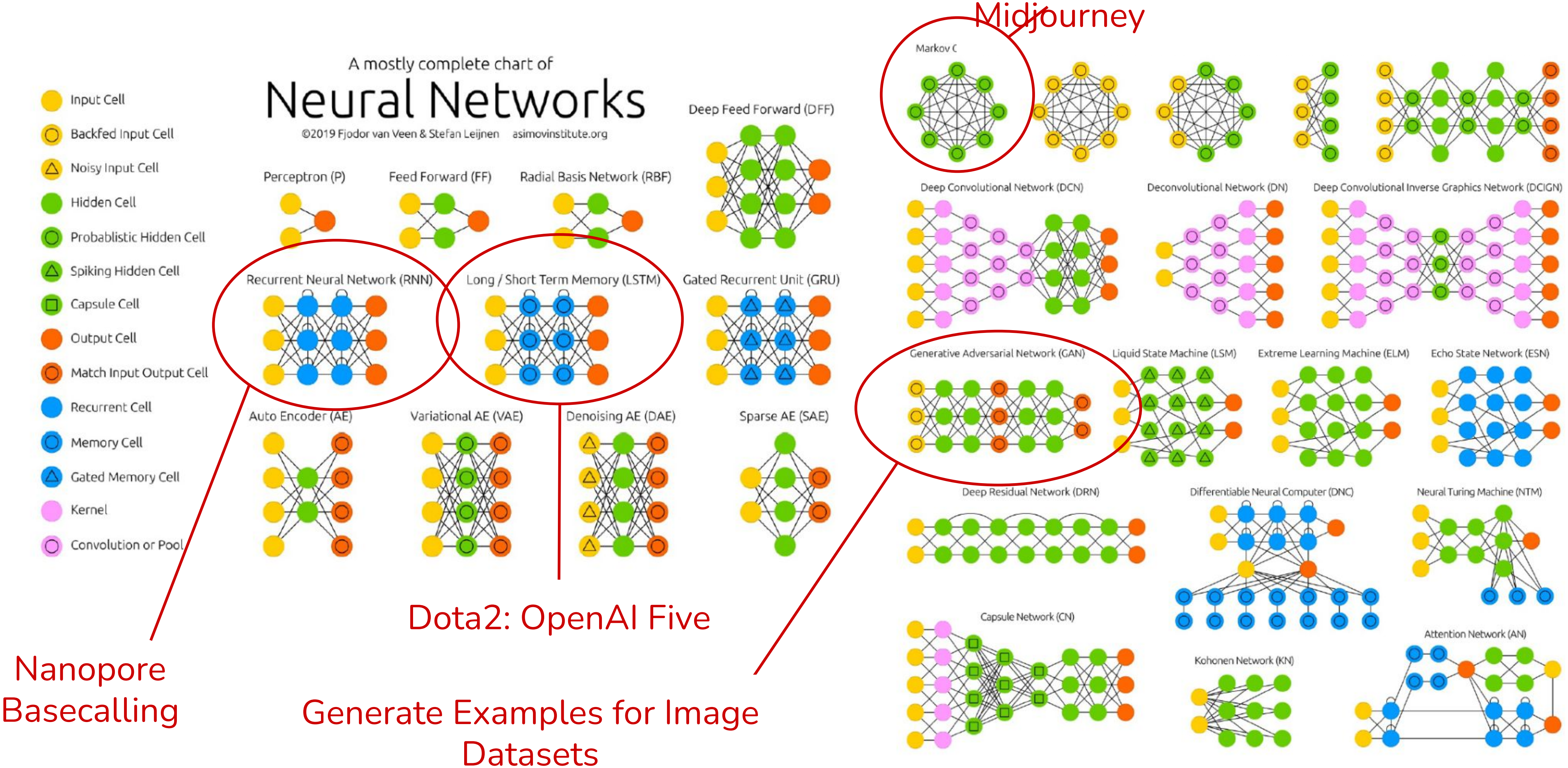
同源性方法（Homology approaches）：
这种方法依赖于相关和高置信度的数据。
如果你的研究对象与数据库中的有显著差异，即提出了一个问题：如果你的生物体与数据库中的生物体相差太远怎么办？

基于训练的方法（Training based approaches）：
如果无法进行直接比较，则可以转向一般性特征/模式。
提到了“基础结构”（Underlying structure），可能指的是用于识别或预测的基本模式和特征。

Training Based Approaches


A mostly complete chart of
Neural Networks
©2019 Fjodor van Veen & Stefan Leijnen asimovinstitute.org
Input Cell
Backfed Input Cell
Noisy Input Cell
Hidden Cell
Probablistic Hidden Cell
Spiking Hidden Cell
Capsule Cell
Output Cell
Match Input Output Cell
Recurrent Cell
Memory Cell
Gated Memory Cell
Kernel
Convolution or Pool
Perceptron (P)
Feed Forward (FF)
Radial Basis Network (RBF)
Deep Feed Forward (DFF)
Recurrent Neural Network (RNN)
Long / Short Term Memory (LSTM)
Gated Recurrent Unit (GRU)
Auto Encoder (AE)
Variational AE (VAE)
Denoising AE (DAE)
Sparse AE (SAE)
Markov Chain (MC)
Hopfield Network (HN)
Boltzmann Machine (BM)
Restricted BM (RBM)
Deep Belief Network (DBN)
Deep Convolutional Network (DCN)
Deconvolutional Network (DN)
Deep Convolutional Inverse Graphics Network (DCIGN)
Generative Adversarial Network (GAN)
Liquid State Machine (LSM)
Extreme Learning Machine (ELM)
Echo State Network (ESN)
Deep Residual Network (DRN)
Differentiable Neural Computer (DNC)
Neural Turing Machine (NTM)
Capsule Network (CN)
Kohonen Network (KN)
Attention Network (AN)

Training Based Approaches


A mostly complete chart of
Neural Networks
©2019 Fjodor van Veen & Stefan Leijnen asimovinstitute.org
Input Cell
Backfed Input Cell
Noisy Input Cell
Hidden Cell
Probablistic Hidden Cell
Spiking Hidden Cell
Capsule Cell
Output Cell
Match Input Output Cell
Recurrent Cell
Memory Cell
Gated Memory Cell
Kernel
Convolution or Pool
Perceptron (P)
Feed Forward (FF)
Radial Basis Network (RBF)
Deep Feed Forward (DFF)
Recurrent Neural Network (RNN)
Long / Short Term Memory (LSTM)
Gated Recurrent Unit (GRU)
Auto Encoder (AE)
Variational AE (VAE)
Denoising AE (DAE)
Sparse AE (SAE)
Markov C
Deep Convolutional Network (DCN)
Deconvolutional Network (DN)
Deep Convolutional Inverse Graphics Network (DCIGN)
Generative Adversarial Network (GAN)
Liquid State Machine (LSM)
Extreme Learning Machine (ELM)
Echo State Network (ESN)
Deep Residual Network (DRN)
Differentiable Neural Computer (DNC)
Neural Turing Machine (NTM)
Capsule Network (CN)
Kohonen Network (KN)
Attention Network (AN)

Common in bioinformatics (many flavors)
Latent diffusion: Stable Diffusion, Midjourney

Nanopore Basecalling

Dota2: OpenAI Five

Generate Examples for Image Datasets

Markov Models

Markov Models

Widely used in bioinformatics.

Visible data (eg a sequence) modelled as a system with states.

Will explore these flavors:

- Markov Chain
- Interpolated Markov Models
- Hidden Markov Models

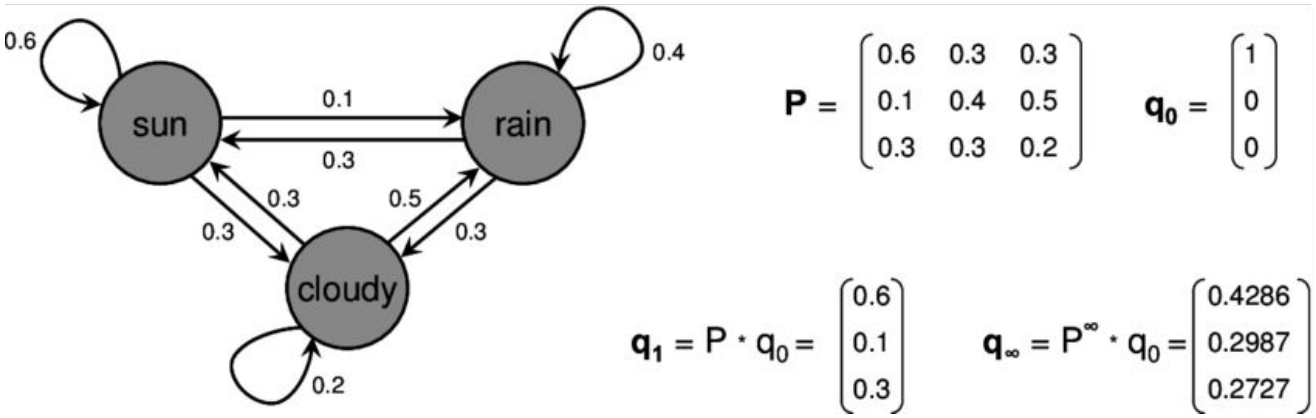
Markov Chain

Simplest example.

Stochastically model of a series of events or ‘states’

Current state only depends on previous state (memorylessness)

Transition matrix models the transition probabilities from one state to another.



$$P = \begin{pmatrix} 0.6 & 0.3 & 0.3 \\ 0.1 & 0.4 & 0.5 \\ 0.3 & 0.3 & 0.2 \end{pmatrix} \qquad q_0 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$$

$$q_1 = P \cdot q_0 = \begin{pmatrix} 0.6 \\ 0.1 \\ 0.3 \end{pmatrix} \qquad q_\infty = P^\infty \cdot q_0 = \begin{pmatrix} 0.4286 \\ 0.2987 \\ 0.2727 \end{pmatrix}$$

Arnold, Ruedi. (2023). Interactive Learning Environments for Mathematical Topics.

马尔可夫链是一种统计模型，它用于预测一系列事件（或“状态”）的随机过程。它的核心假设是，当前状态只取决于前一个状态，这也被称为无记忆性（memorylessness）。在数学上，这种状态转换由一个叫做转移矩阵（transition matrix）的概念来描述，它包含了从一个状态转移到另一个状态的概率。

幻灯片中的第一个图显示了一个简单的天气模型，其中的状态有“晴天”、“下雨”和“多云”。每个状态之间的箭头代表了转移概率，比如从晴天到下雨的概率是0.1，从多云到晴天的概率是0.3等。

第二个图是一个转移矩阵P，它数学上表示了上述状态转换的概率。

右侧的计算展示了如何使用转移矩阵来预测下一个状态。q0代表一个状态的初始概率分布，q1则是经过一次转换后的概率分布，计算为P * q0。q∞代表状态分布最终稳定时的概率分布，计算为P的多次幂乘以q0直到收敛。

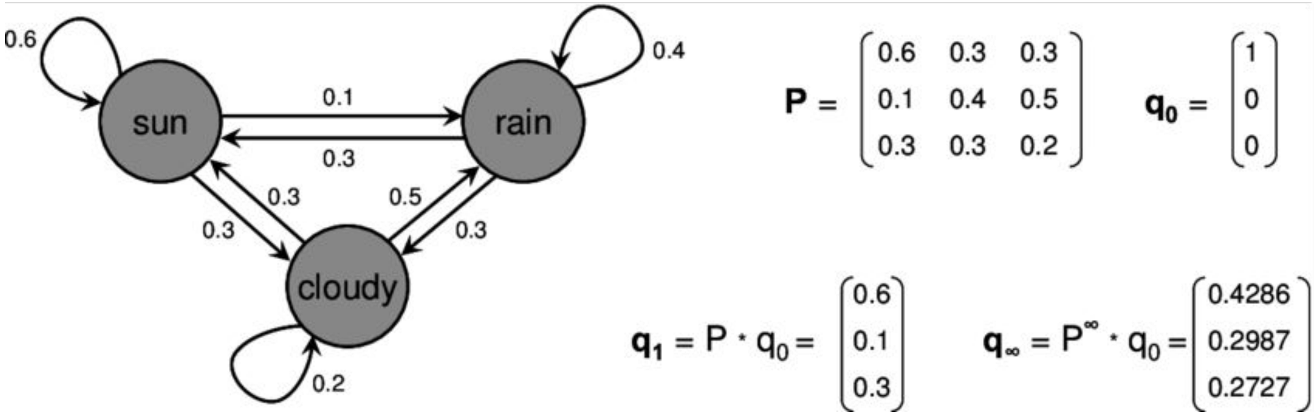
Markov Chain

Simplest example.

Stochastically model of a series of events or ‘states’

Current state only depends on previous state (memorylessness)

Transition matrix models the transition probabilities from one state to another.



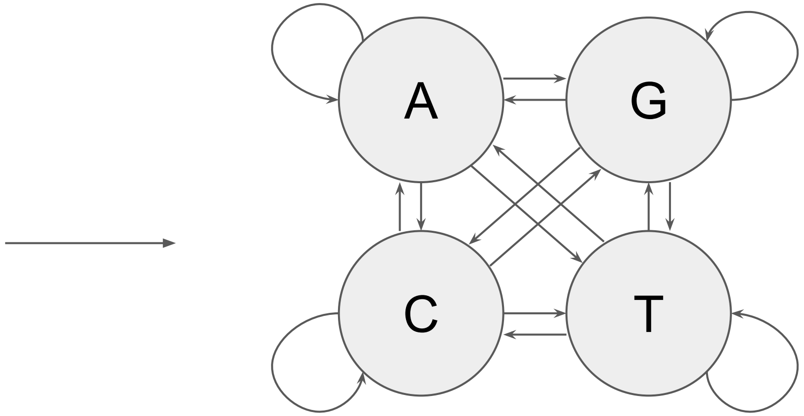
$$\mathbf{P} = \begin{pmatrix} 0.6 & 0.3 & 0.3 \\ 0.1 & 0.4 & 0.5 \\ 0.3 & 0.3 & 0.2 \end{pmatrix} \qquad \mathbf{q_0} = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$$

$$\mathbf{q_1} = P \cdot q_0 = \begin{pmatrix} 0.6 \\ 0.1 \\ 0.3 \end{pmatrix} \qquad \mathbf{q_\infty} = P^\infty \cdot q_0 = \begin{pmatrix} 0.4286 \\ 0.2987 \\ 0.2727 \end{pmatrix}$$

Arnold, Ruedi. (2023). Interactive Learning Environments for Mathematical Topics.

A C A A G C A T T G A G …

… C T T C T A A G C G A T



How does this apply to gene prediction?

Goal: Predict a genomic feature.

Underlying structure of coding regions is different to non-coding regions.

How does this apply to gene prediction?

编码区的结构与非编码区不同：基因组中的编码区（coding regions）是指那些可以被转录和翻译成蛋白质的DNA序列。非编码区（non-coding regions）通常不直接编码蛋白质，但它们可能包含调控基因表达的序列。

Goal: Predict a genomic feature.

编码蛋白质对序列施加了不同的压力：在进化过程中，为了有效地编码蛋白质，一些特定的碱基序列变得更常见。例如，某些氨基酸在蛋白质中出现的频率比其他氨基酸高，这会影响它们对应的密码子（codons）在基因序列中的分布

Underlying structure of coding regions is different to non-coding regions.

Encoding a protein puts different pressures on the sequence.

Certain amino acids are more common

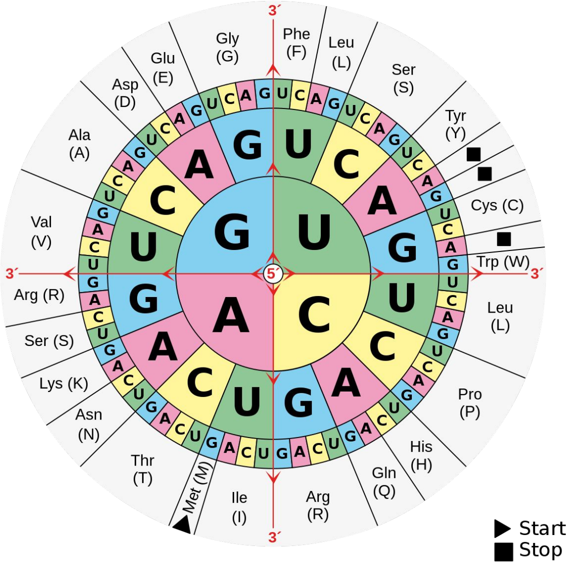
Leucine used more commonly than Cysteine

Codon wobble

Alanine has 4 codons, Methionine has 1 codon.

Codon preference

For a given species, usually one codon is used more often than others



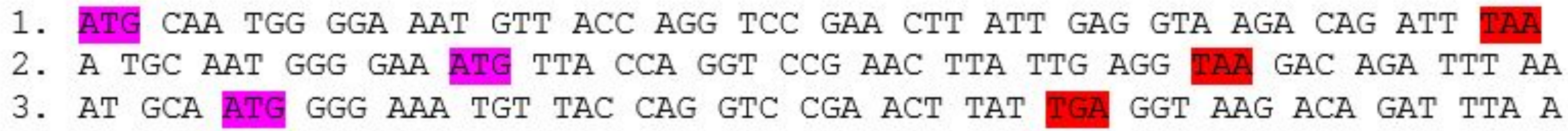
How does this apply to gene prediction?

1. Generate two models:
   - One from known coding sequence
   - One from known non-coding sequence

How does this apply to gene prediction?

1. Generate two models:
   - One from known coding sequence
   - One from known non-coding sequence

2. Identify open reading frames (ORFs)
   - Start / stop codons in the same frame



```
1. ATG CAA TGG GGA AAT GTT ACC AGG TCC GAA CTT ATT GAG GTA AGA CAG ATT TAA
2. A TGC AAT GGG GAA ATG TTA CCA GGT CCG AAC TTA TTG AGG TAA GAC AGA TTT AA
3. AT GCA ATG GGG AAA TGT TAC CAG GTC CGA ACT TAT TGA GGT AAG ACA GAT TTA A
```

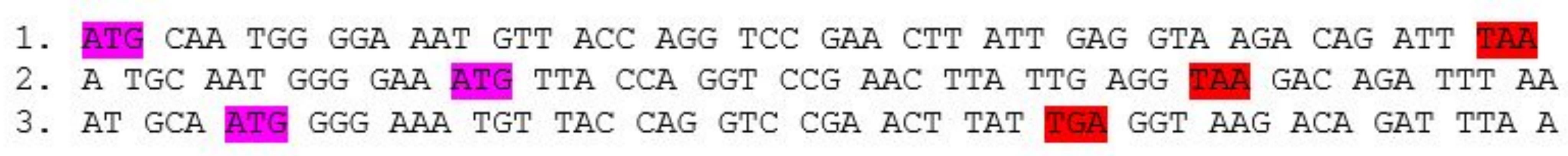
All 3 possible ORFS for +ve strand, indicated by start & stop codon

How does this apply to gene prediction?

1. Generate two models:
   - One from known coding sequence
   - One from known non-coding sequence

2. Identify open reading frames (ORFs)
   - Start / stop codons in the same frame

3. For each ORF, score the sequence by:
   - Calculating the probability that it was generated by the coding model.
   - Calculating the probability that it was generated by the non-coding model.



```
1. ATG CAA TGG GGA AAT GTT ACC AGG TCC GAA CTT ATT GAG GTA AGA CAG ATT TAA
2. A TGC AAT GGG GAA ATG TTA CCA GGT CCG AAC TTA TTG AGG TAA GAC AGA TTT AA
3. AT GCA ATG GGG AAA TGT TAC CAG GTC CGA ACT TAT TGA GGT AAG ACA GAT TTA A
```

All 3 possible ORFS for +ve strand, indicated by start & stop codon

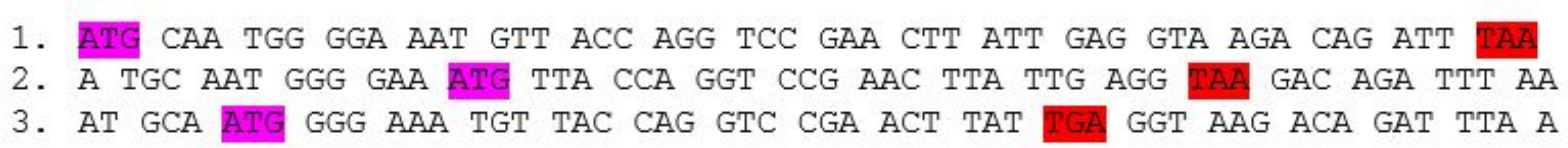
How does this apply to gene prediction?

1. Generate two models:
   - One from known coding sequence
   - One from known non-coding sequence

2. Identify open reading frames (ORFs)
   - Start / stop codons in the same frame

3. For each ORF, score the sequence by:
   - Calculating the probability that it was generated by the coding model.
   - Calculating the probability that it was generated by the non-coding model.

4. Higher probability for coding model indicates possible gene
   - Confidence can be measured using some form of ratio.



```
1. ATG CAA TGG GGA AAT GTT ACC AGG TCC GAA CTT ATT GAG GTA AGA CAG ATT TAA
2. A TGC AAT GGG GAA ATG TTA CCA GGT CCG AAC TTA TTG AGG TAA GAC AGA TTT AA
3. AT GCA ATG GGG AAA TGT TAC CAG GTC CGA ACT TAT TGA GGT AAG ACA GAT TTA A
```

All 3 possible ORFS for +ve strand, indicated by start & stop codon

Interpolated Markov Model (IMM)

Previously: 1st order model

- States: ‘A’, ‘G’, ‘T’, ‘C’

Interpolated Markov Model (IMM)

Previously: 1st order model

- States: ‘A’, ‘G’, ‘T’, ‘C’

Would be nice if we could include more “history”

- Eg. Predict the next word:
- “... what” {“do”, “is”, “the”, “to”}

Interpolated Markov Model (IMM)

Previously: 1st order model

States: ‘A’, ‘G’, ‘T’, ‘C’

Would be nice if we could include more “history”

Eg. Predict the next word:

“... what” {“do”, “is”, “the”, “to”}

Interpolated Markov Model (IMM)

Previously: 1st order model

States: ‘A’, ‘G’, ‘T’, ‘C’

Would be nice if we could include more “history”

Eg. Predict the next word:

“... what” {“do”, “is”, “the”, “to”}

“... I don’t know what” {“do”, “is”, “the”, “to”}

Interpolated Markov Model (IMM)

Previously: 1st order model

States: ‘A’, ‘G’, ‘T’, ‘C’

Would be nice if we could include more “history”

Eg. Predict the next word:

“... what” {“do”, “is”, “the”, “to”}

“... I don’t know what” {“do”, “is”, “the”, “to”}

Interpolated Markov Model (IMM)

Previously: 1st order model

States: ‘A’, ‘G’, ‘T’, ‘C’

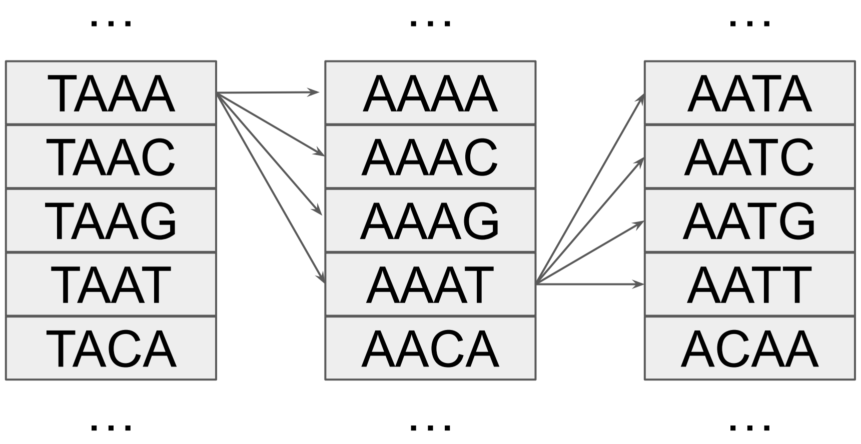
Would be nice if we could include more “history”

Eg. Predict the next word:

“... what” {“do”, “is”, “the”, “to”}

“... I don’t know what” {“do”, “is”, “the”, “to”}

4th order model



…

TAAA

TAAC

TAAG

TAAT

TACA

…

…

AAAA

AAAC

AAAG

AAAT

AACA

…

…

AATA

AATC

AATG

AATT

ACAA

…

Interpolated Markov Model (IMM)

Issue: not enough data!

Number of parameters (transition probabilities) we must estimate grows exponentially with order

n^{th} order: 4^{n+1}

1^{st} order: $4^2 = 16$

4^{th} order: $4^5 = 1024$

Interpolated Markov Model (IMM)

Issue: not enough data!

Number of parameters (transition probabilities) we must estimate grows exponentially with order

- n^{th} order: 4^{n+1}
- 1^{st} order: $4^2 = 16$
- 4^{th} order: $4^5 = 1024$

Imagining we have 200k bases of sequence data to estimate parameters:

- For 2^{nd} order Markov chain, expect to see each state ~12k times
- For 8^{th} order Markov chain, expect to see each state ~3 times

Interpolated Markov Model (IMM)

Issue: not enough data!

Number of parameters (transition probabilities) we must estimate grows exponentially with order

n^th order: 4^{n+1}

1st order: $4^2 = 16$

4th order: $4^5 = 1024$

Imagining we have 200k bases of sequence data to estimate parameters:

For 2nd order Markov chain, expect to see each state ~12k times

For 8th order Markov chain, expect to see each state ~3 times

Interpolated Markov Models (eg in GLIMMA)

- Build ***w*** orders (individual chains)
- For each base position in sequence to assess, combine probabilities from each chain
- Weight by how much evidence was witnessed for that state for that particular chain

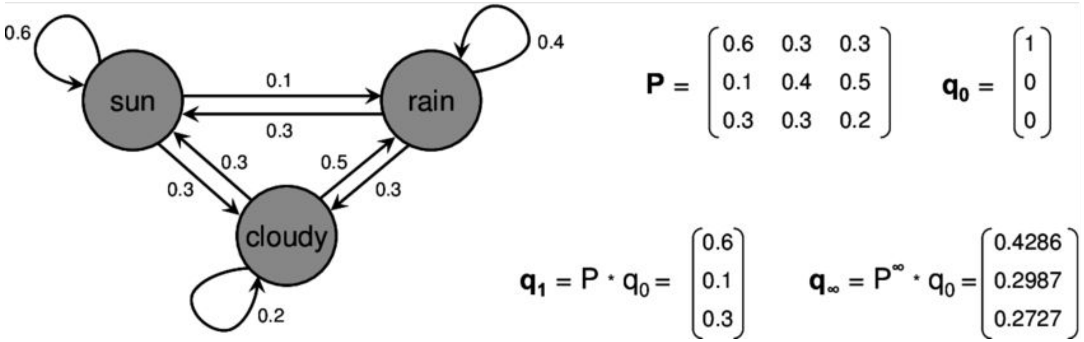
Hidden Markov Model (HMM)

Markov Chain incorporating hidden ‘states’ which affect witnessed output

Hidden Markov Model (HMM)

Markov Chain incorporating hidden ‘states’ which affect witnessed output

Markov Model



$$\mathbf{P} = \begin{pmatrix} 0.6 & 0.3 & 0.3 \\ 0.1 & 0.4 & 0.5 \\ 0.3 & 0.3 & 0.2 \end{pmatrix} \qquad \mathbf{q_0} = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$$

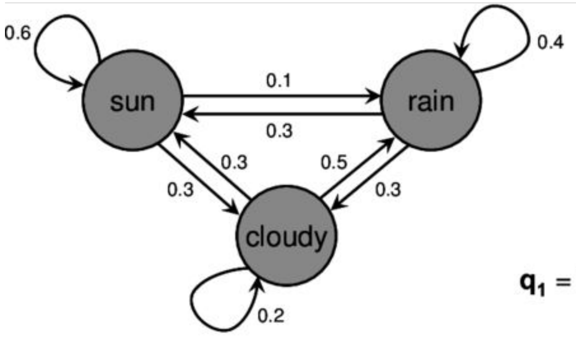
$$\mathbf{q_1} = P \cdot q_0 = \begin{pmatrix} 0.6 \\ 0.1 \\ 0.3 \end{pmatrix} \qquad \mathbf{q_\infty} = P^{\infty} \cdot q_0 = \begin{pmatrix} 0.4286 \\ 0.2987 \\ 0.2727 \end{pmatrix}$$

Arnold, Ruedi. (2023). Interactive Learning Environments for Mathematical Topics.

Hidden Markov Model (HMM)

Markov Chain incorporating hidden ‘states’ which affect witnessed output

Markov Model

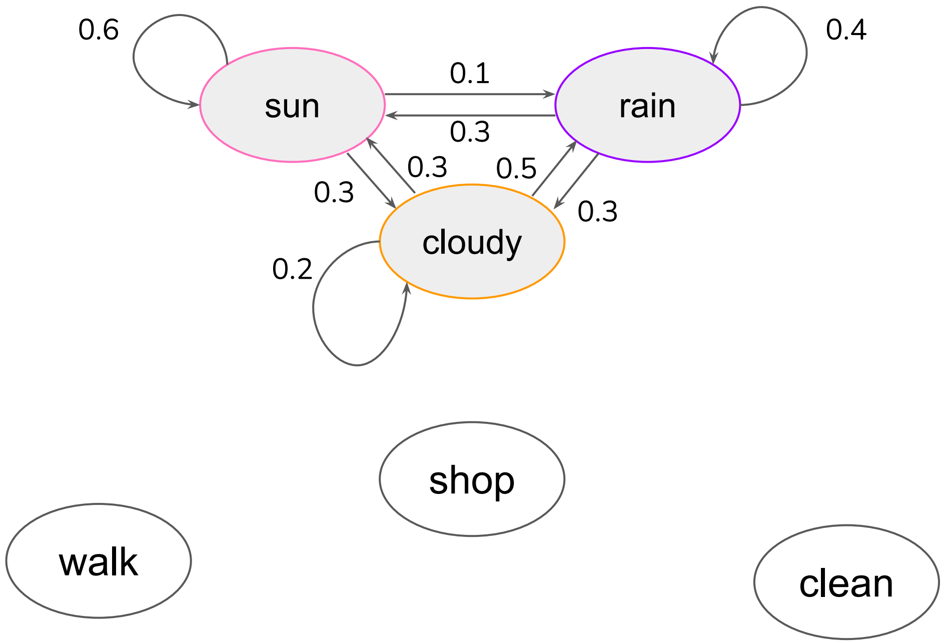


$$\mathbf{P} = \begin{bmatrix} 0.6 & 0.3 & 0.3 \\ 0.1 & 0.4 & 0.5 \\ 0.3 & 0.3 & 0.2 \end{bmatrix} \qquad \mathbf{q_0} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$$

$$\mathbf{q_1} = \mathrm{P} \cdot \mathrm{q}_0 = \begin{bmatrix} 0.6 \\ 0.1 \\ 0.3 \end{bmatrix} \qquad \mathbf{q_\infty} = \mathrm{P}^{\infty} \cdot \mathrm{q}_0 = \begin{bmatrix} 0.4286 \\ 0.2987 \\ 0.2727 \end{bmatrix}$$

Arnold, Ruedi. (2023). Interactive Learning Environments for Mathematical Topics.

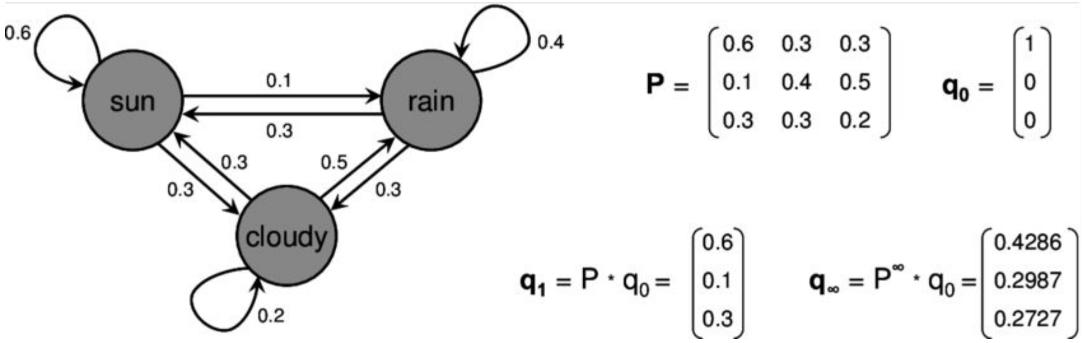
Hidden Markov Model



Hidden Markov Model (HMM)

Markov Chain incorporating hidden ‘states’ which affect witnessed output

Markov Model

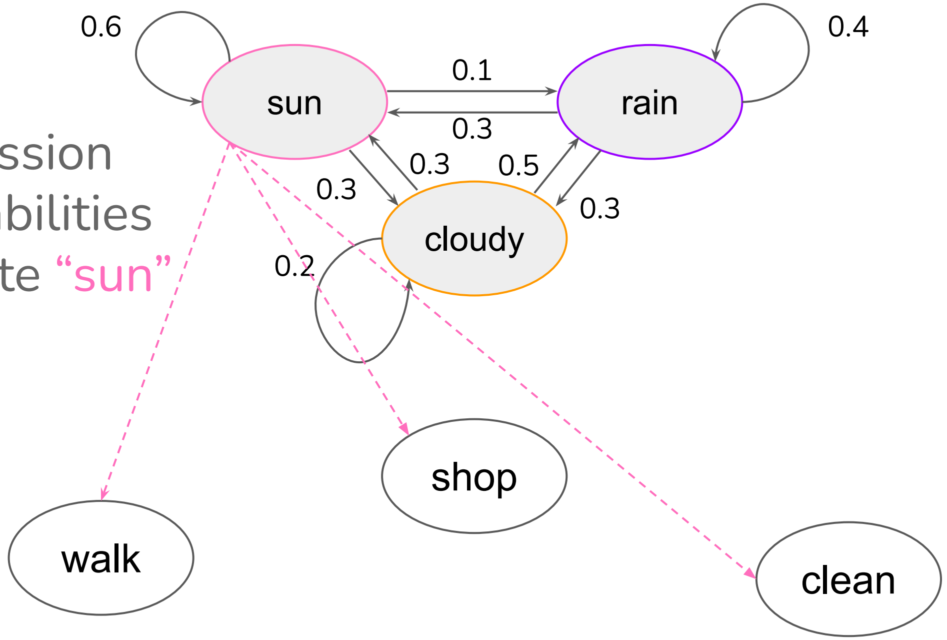


$$\mathbf{P} = \begin{bmatrix} 0.6 & 0.3 & 0.3 \\ 0.1 & 0.4 & 0.5 \\ 0.3 & 0.3 & 0.2 \end{bmatrix} \quad \mathbf{q_0} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$$

$$\mathbf{q_1} = P \cdot q_0 = \begin{bmatrix} 0.6 \\ 0.1 \\ 0.3 \end{bmatrix} \quad \mathbf{q_\infty} = P^{\infty} \cdot q_0 = \begin{bmatrix} 0.4286 \\ 0.2987 \\ 0.2727 \end{bmatrix}$$

Arnold, Ruedi. (2023). Interactive Learning Environments for Mathematical Topics.

Hidden Markov Model

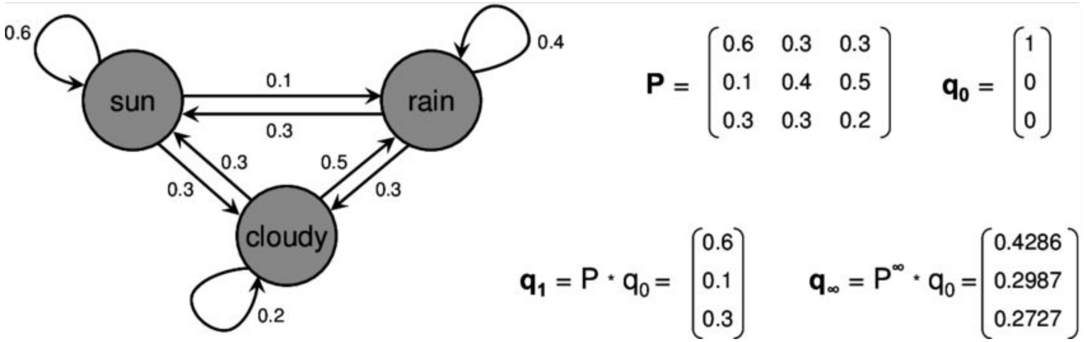


Emission probabilities for state “sun”

Hidden Markov Model (HMM)

Markov Chain incorporating hidden ‘states’ which affect witnessed output

Markov Model

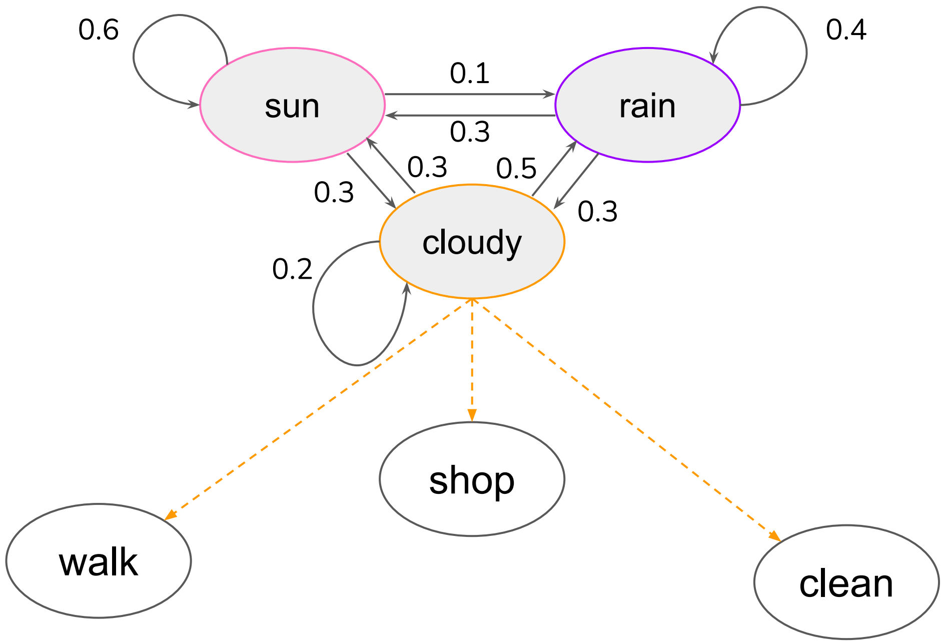


$$\mathbf{P} = \begin{bmatrix} 0.6 & 0.3 & 0.3 \\ 0.1 & 0.4 & 0.5 \\ 0.3 & 0.3 & 0.2 \end{bmatrix} \qquad \mathbf{q_0} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$$

$$\mathbf{q_1} = \mathrm{P} \cdot \mathrm{q}_0 = \begin{bmatrix} 0.6 \\ 0.1 \\ 0.3 \end{bmatrix} \qquad \mathbf{q_\infty} = \mathrm{P}^{\infty} \cdot \mathrm{q}_0 = \begin{bmatrix} 0.4286 \\ 0.2987 \\ 0.2727 \end{bmatrix}$$

Arnold, Ruedi. (2023). Interactive Learning Environments for Mathematical Topics.

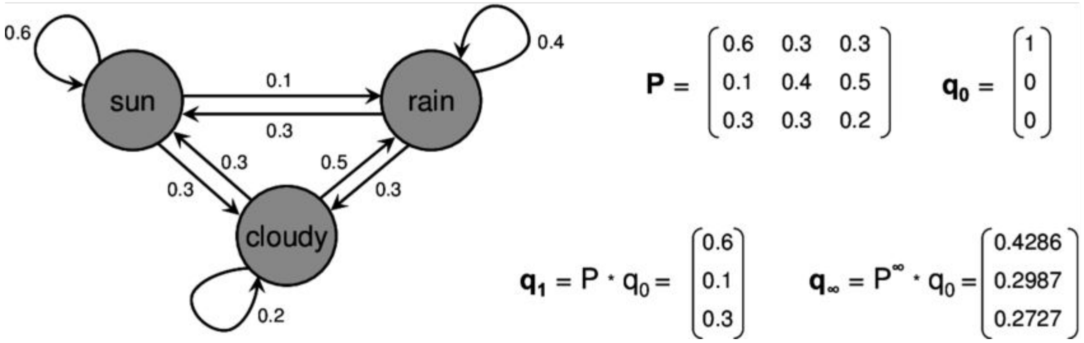
Hidden Markov Model



Hidden Markov Model (HMM)

Markov Chain incorporating hidden ‘states’ which affect witnessed output

Markov Model

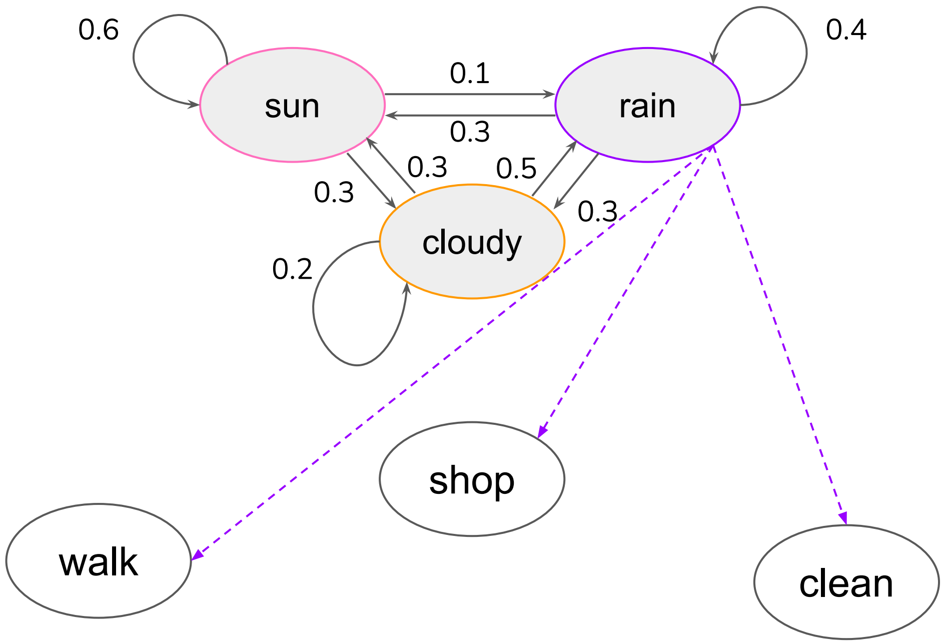


$$\mathbf{P} = \begin{bmatrix} 0.6 & 0.3 & 0.3 \\ 0.1 & 0.4 & 0.5 \\ 0.3 & 0.3 & 0.2 \end{bmatrix} \qquad \mathbf{q_0} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$$

$$\mathbf{q_1} = \mathrm{P} \cdot \mathrm{q}_0 = \begin{bmatrix} 0.6 \\ 0.1 \\ 0.3 \end{bmatrix} \qquad \mathbf{q_\infty} = \mathrm{P}^\infty \cdot \mathrm{q}_0 = \begin{bmatrix} 0.4286 \\ 0.2987 \\ 0.2727 \end{bmatrix}$$

Arnold, Ruedi. (2023). Interactive Learning Environments for Mathematical Topics.

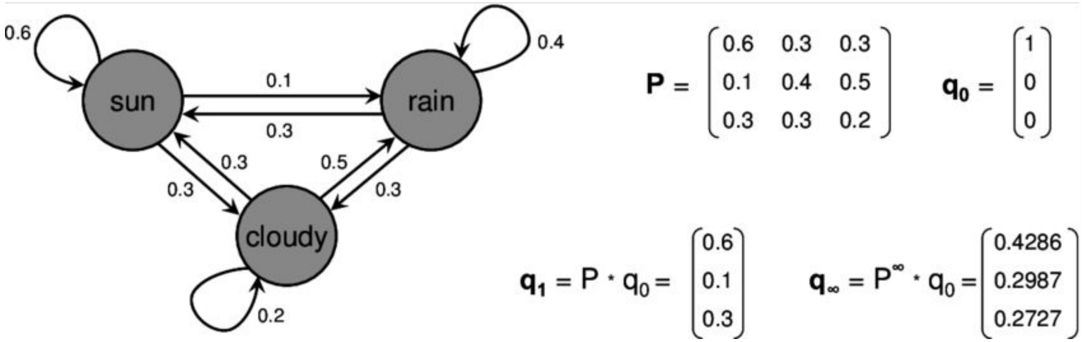
Hidden Markov Model



Hidden Markov Model (HMM)

Markov Chain incorporating hidden ‘states’ which affect witnessed output

Markov Model

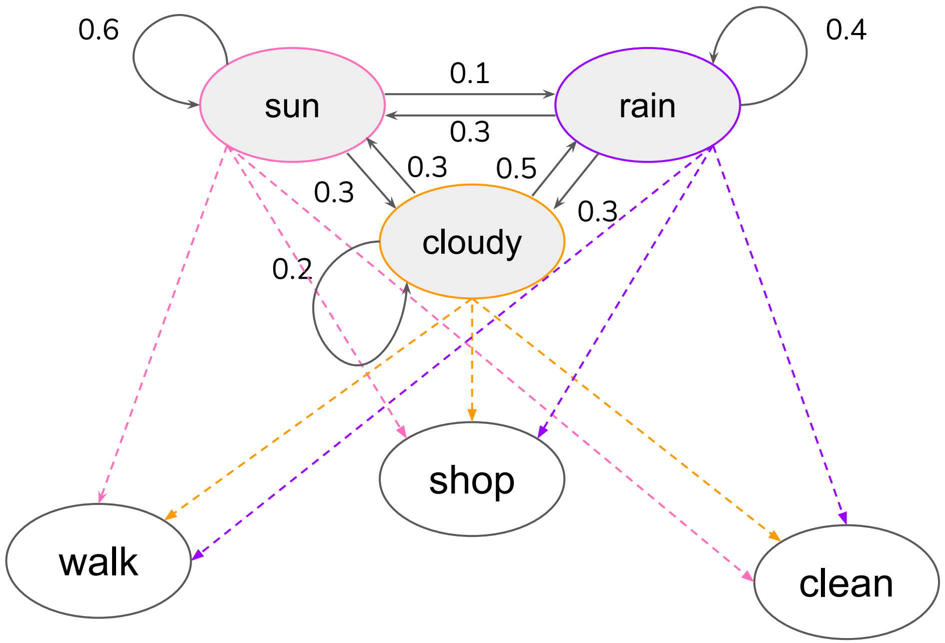


$$\mathbf{P} = \begin{bmatrix} 0.6 & 0.3 & 0.3 \\ 0.1 & 0.4 & 0.5 \\ 0.3 & 0.3 & 0.2 \end{bmatrix} \qquad \mathbf{q_0} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$$

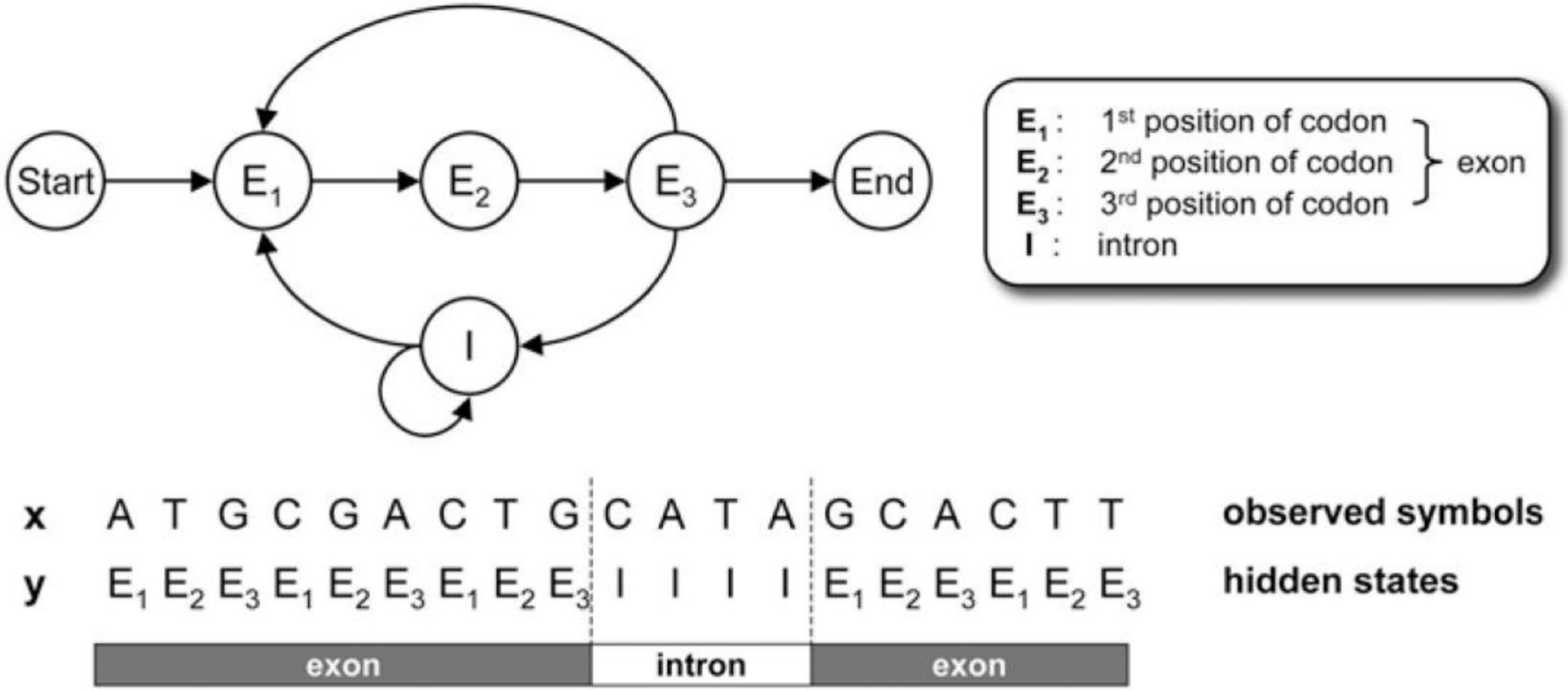
$$\mathbf{q_1} = \mathrm{P} \cdot \mathrm{q_0} = \begin{bmatrix} 0.6 \\ 0.1 \\ 0.3 \end{bmatrix} \qquad \mathbf{q_\infty} = \mathrm{P}^{\infty} \cdot \mathrm{q_0} = \begin{bmatrix} 0.4286 \\ 0.2987 \\ 0.2727 \end{bmatrix}$$

Arnold, Ruedi. (2023). Interactive Learning Environments for Mathematical Topics.

Hidden Markov Model



Hidden Markov Model (HMM)

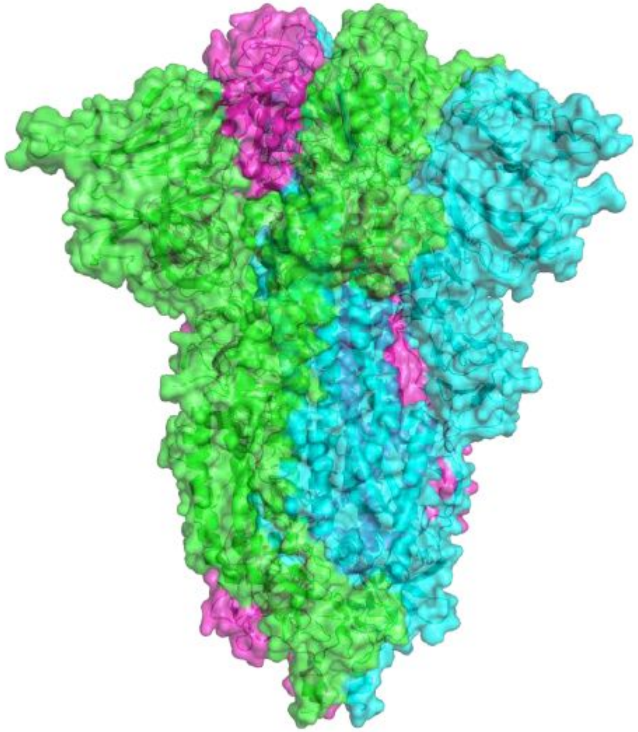

Start
E1
E2
E3
End
I
E1 : 1st position of codon
E2 : 2nd position of codon
E3 : 3rd position of codon
exon
I : intron
x A T G C G A C T G C A T A G C A C T T
observed symbols
y E1 E2 E3 E1 E2 E3 E1 E2 E3 I I I I E1 E2 E3 E1 E2 E3
hidden states
exon
intron
exon

Yoon, BJ 2009. doi: 10.2174/138920209789177575

Deep Learning: AlphaFold 2.0

AlphaFold 2.0

Given a sequence of amino acids, predicts 3D fold.



"Spike" protein (7CAB)

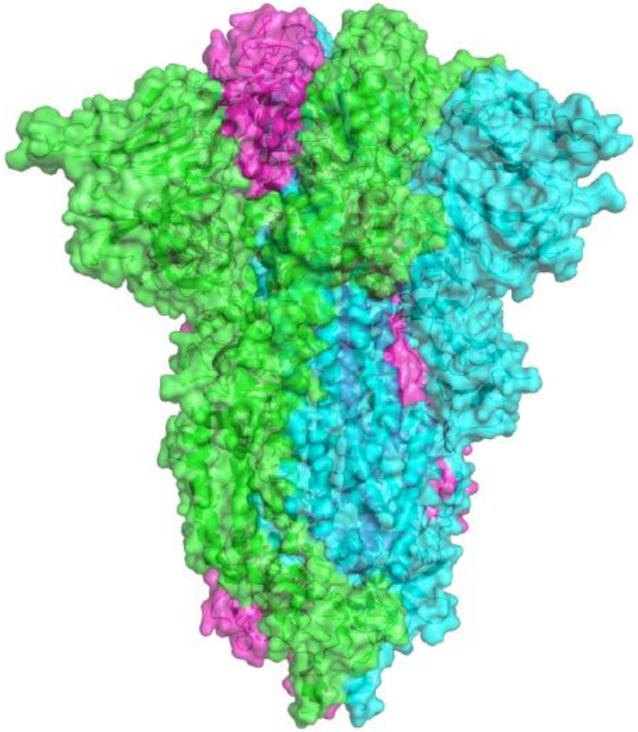
Mediates viral entry into the host cell

AlphaFold 2.0

Given a sequence of amino acids, predicts 3D fold.

Can be used for gene annotation:

- For situations where DNA / AA identity to known proteins devolves below recognisable threshold
- Conservation: DNA (least) -> AA -> Fold (most)



"Spike" protein (7CAB)

Mediates viral entry into the host cell

AlphaFold 2.0

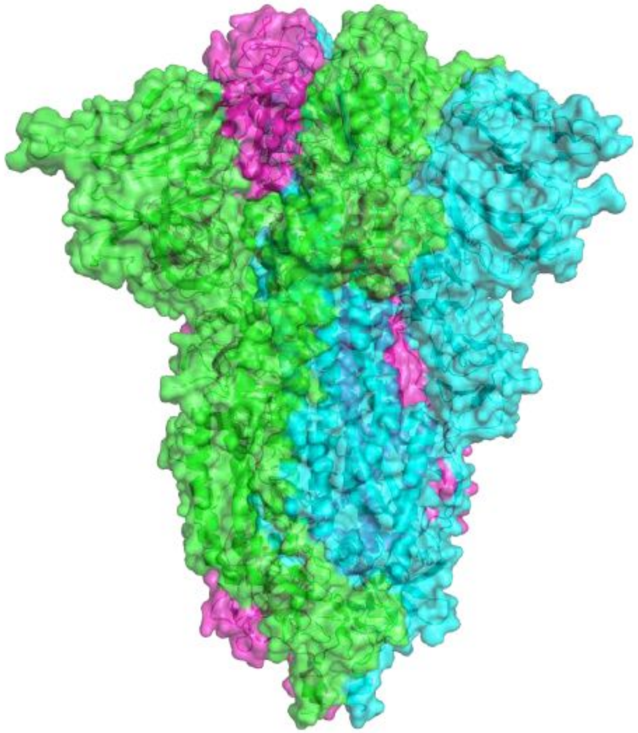
Given a sequence of amino acids, predicts 3D fold.

Can be used for gene annotation:

For situations where DNA / AA identity to known proteins devolves below recognisable threshold

Conservation: DNA (least) -> AA -> Fold (most)

1. Identify possible genes (maybe HMM / IMM)
2. Fold each candidate
3. Compare each fold to structures on PDB



"Spike" protein (7CAB)

Mediates viral entry into the host cell

AlphaFold 2.0

Given a sequence of amino acids, predicts 3D fold.

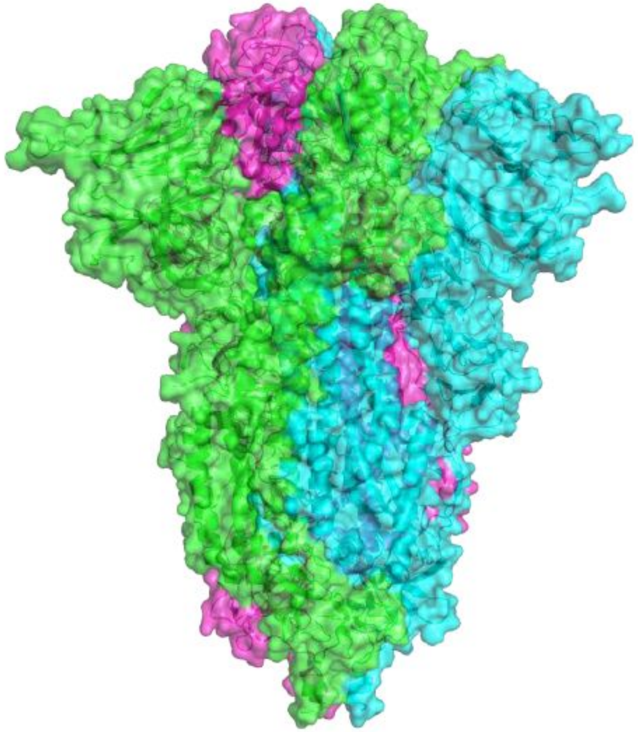
Can be used for gene annotation:

For situations where DNA / AA identity to known proteins devolves below recognisable threshold

Conservation: DNA (least) -> AA -> Fold (most)

1. Identify possible genes (maybe HMM / IMM)
2. Fold each candidate
3. Compare each fold to structures on PDB

Machine learning (training & inference), but then uses homology (RMSD) to known proteins



"Spike" protein (7CAB)

Mediates viral entry into the host cell

Protein Folding - Levinthal's Paradox (1969)

- The number of possible conformation available to a protein is astronomically large
  - 100 Amino acid protein
  - 3 conformations per amino acid

3^{100} conformations → 10^{12} conformations/second

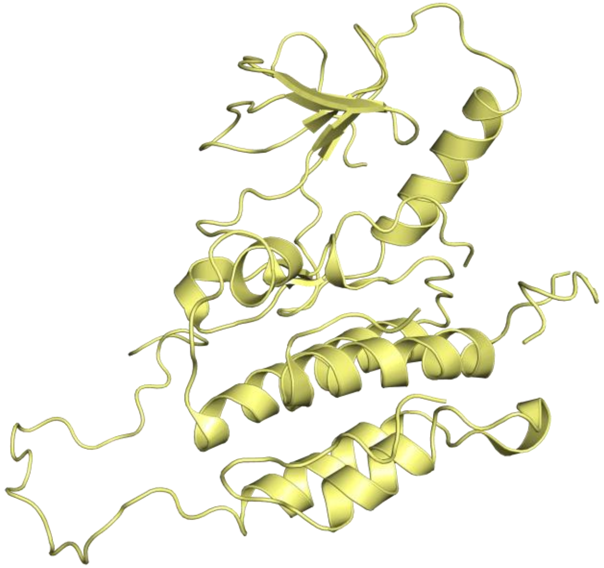
↓

2 x 10^{28} years

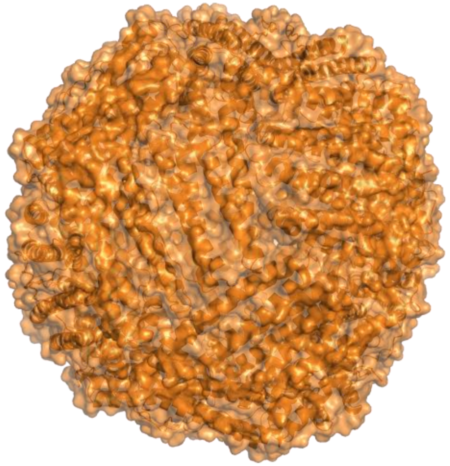
Protein folding is not random and must have a specific pathway

```
>P15056|BRAF_HUMAN
MAALSGGGGGGAEPGQALFNGDMEPEAGAGAGAAASSAADPAIPEEVWNIKQMIKLTQEH
IEALLDKFGGEHNPPSIYLEAYEEYTSKLDALQQREQQLLESLGNGTDFSVSSSASMDTV
TSSSSSSLSVLPSSLSVFQNPTDVARSNPKSPQKPIVRVFLPNKQRTVVPARCGVTVRDS
LKKALMMRGLIPECCAVYRIQDGEKKPIGWDTDISWLTGEELHVEVLENVPLTTHNFVRK
TFFTLAFCDFCRKLLFQGFRCQTCGYKFHQRCSTEVPLMCVNYDQLDLLFVSKFFEHHPI
PQEEASLAETALTSGSSPSAPASDSIGPQILTSPSPSKSIPIPQPFRPADEDHRNQFGQR
DRSSSAPNVHINTIEPVNIDDLIRDQGFRGDGGSTTGLSATPPASLPGSLTNVKALQKSP
GPQRERKSSSSSEDRNRMKTLGRRDSSDDWEIPDGQITVGQRIGSGSFGTVYKGKWHGDV
AVKMLNVTAPTPQQLQAFKNEVGVLRKTRHVNILLFMGYSTKPQLAIVTQWCEGSSLYHH
LHIIETKFEMIKLIDIARQTAQGMDYLHAKSIIHRDLKSNNIFLHEDLTVKIGDFGLATV
KSRWSGSHQFEQLSGSILWMAPEVIRMQDKNPYSFQSDVYAFGIVLYELMTGQLPYSNIN
NRDQIIFMVGRGYLSPDLSKVRSNCPKAMKRLMAECLKKKRDERPLFPQILASIELLARS
LPKIHRSASEPSLNRAGFQTEDFSLYACASPKTPIQAGGYGAFPVH
```

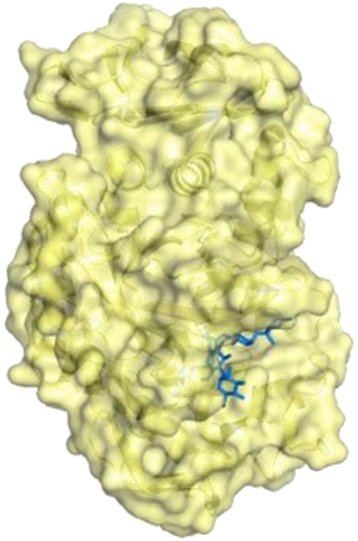
→



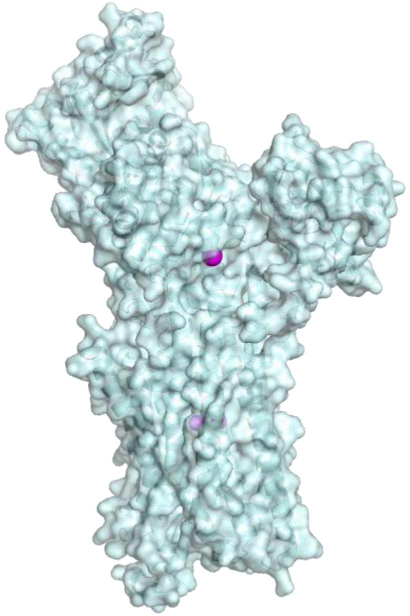
Proteins: Structure and Function



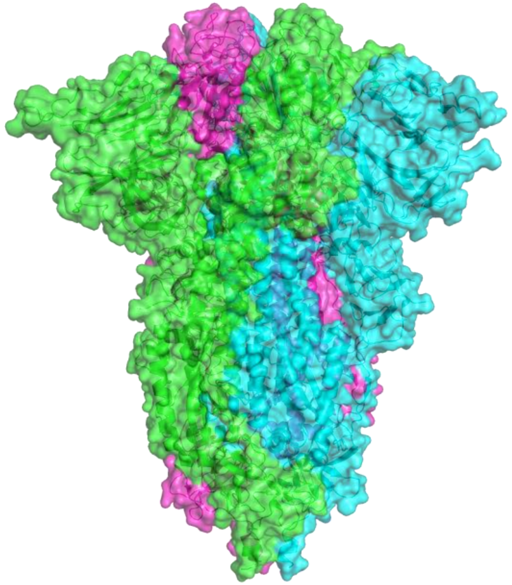
Ferritin (1FHA)
Forms a hollow shell that stores iron from our food



Alpha-amylase (1PPI)
An enzyme with a catalytic site that begins the breakdown of carbohydrates in our saliva



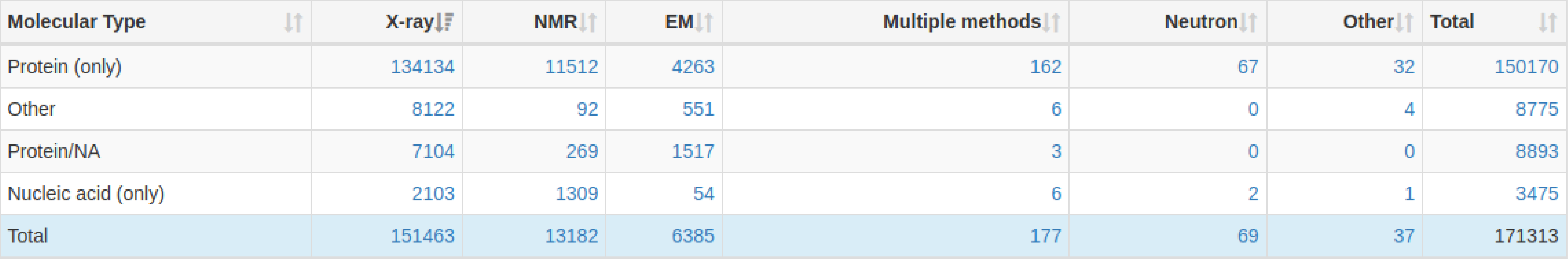
Calcium Pump (1SU4)
Moves ions across cell membrane



"Spike" protein (7CAB)
Mediate viral entry in the host cell

Protein Data Bank

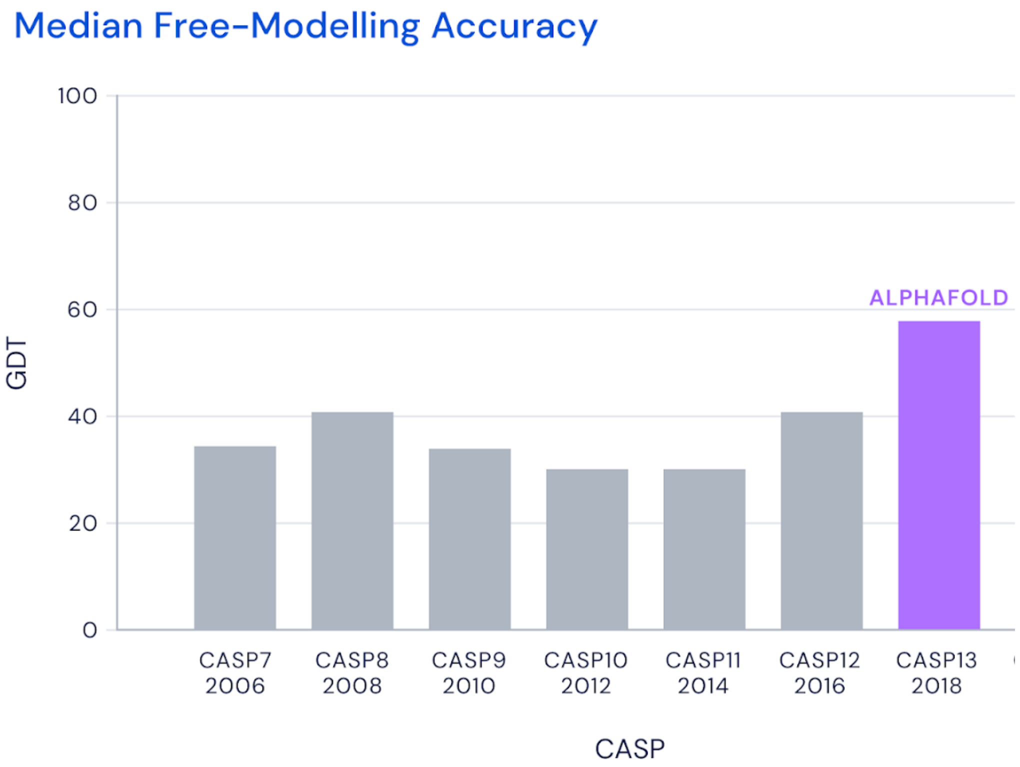
- 3D data for biological macromomolecules
  - Proteins
  - Nucleic Acids (DNA – RNA)
  - Organelles
  - Viruses

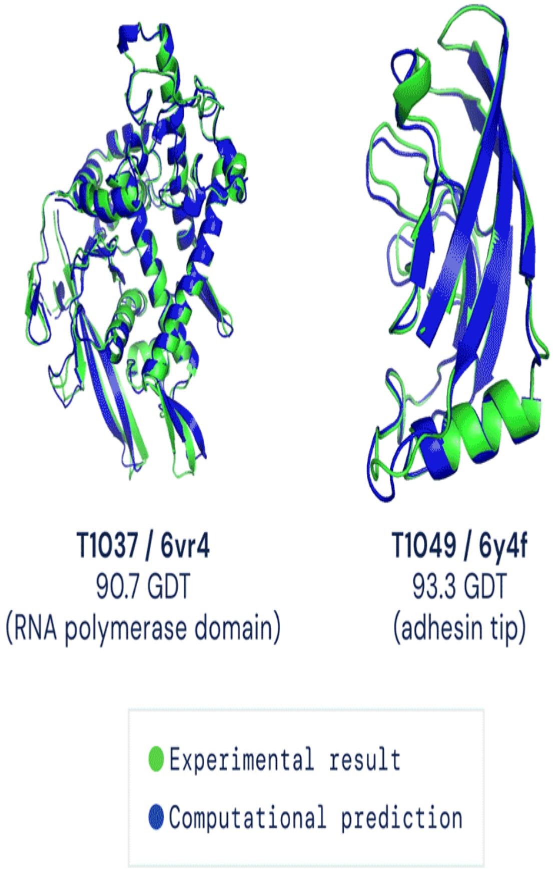


Molecular Type	X-ray	NMR	EM	Multiple methods	Neutron	Other	Total
Protein (only)	134134	11512	4263	162	67	32	150170
Other	8122	92	551	6	0	4	8775
Protein/NA	7104	269	1517	3	0	0	8893
Nucleic acid (only)	2103	1309	54	6	2	1	3475
Total	151463	13182	6385	177	69	37	171313

https://www.rcsb.org/stats/summary

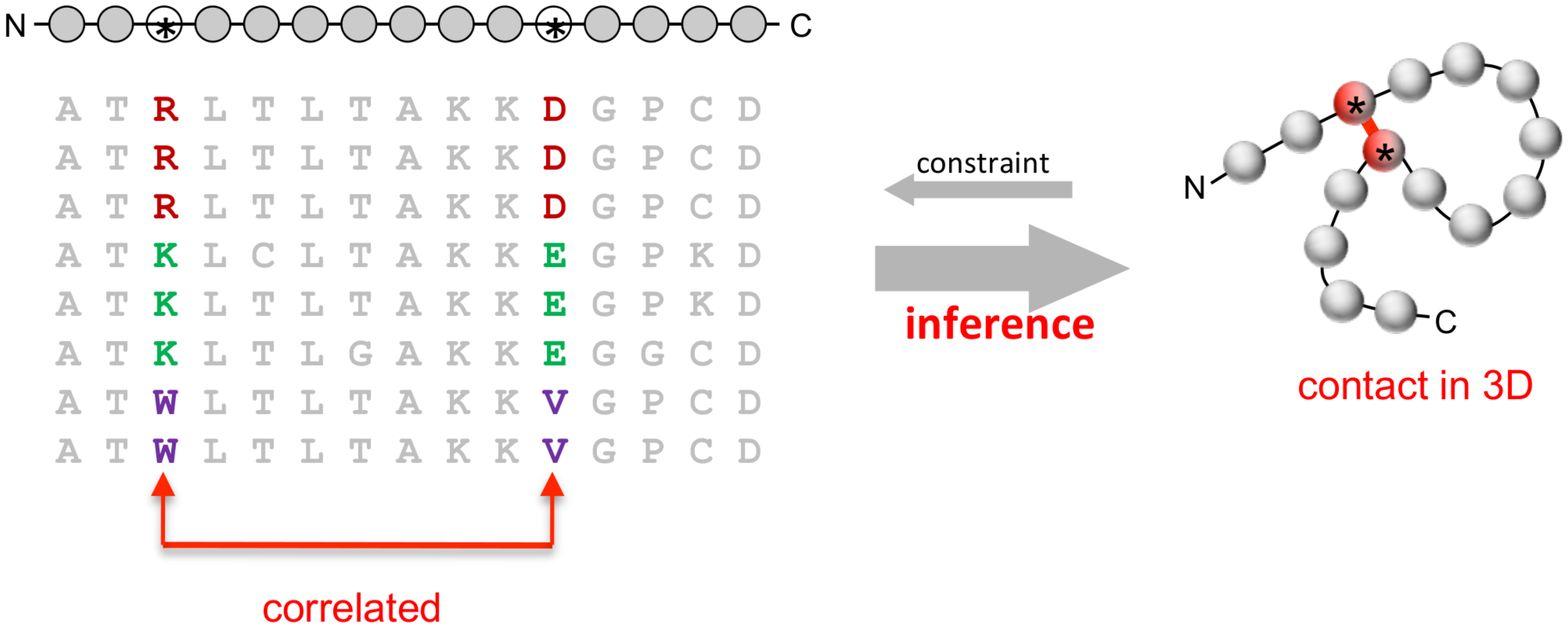
Critical Assessment of Structure Prediction 2018 (CASP13)





Senior, Andrew W., et al. "Improved protein structure prediction using potentials from deep learning." Nature 577.7792 (2020): 706-710.

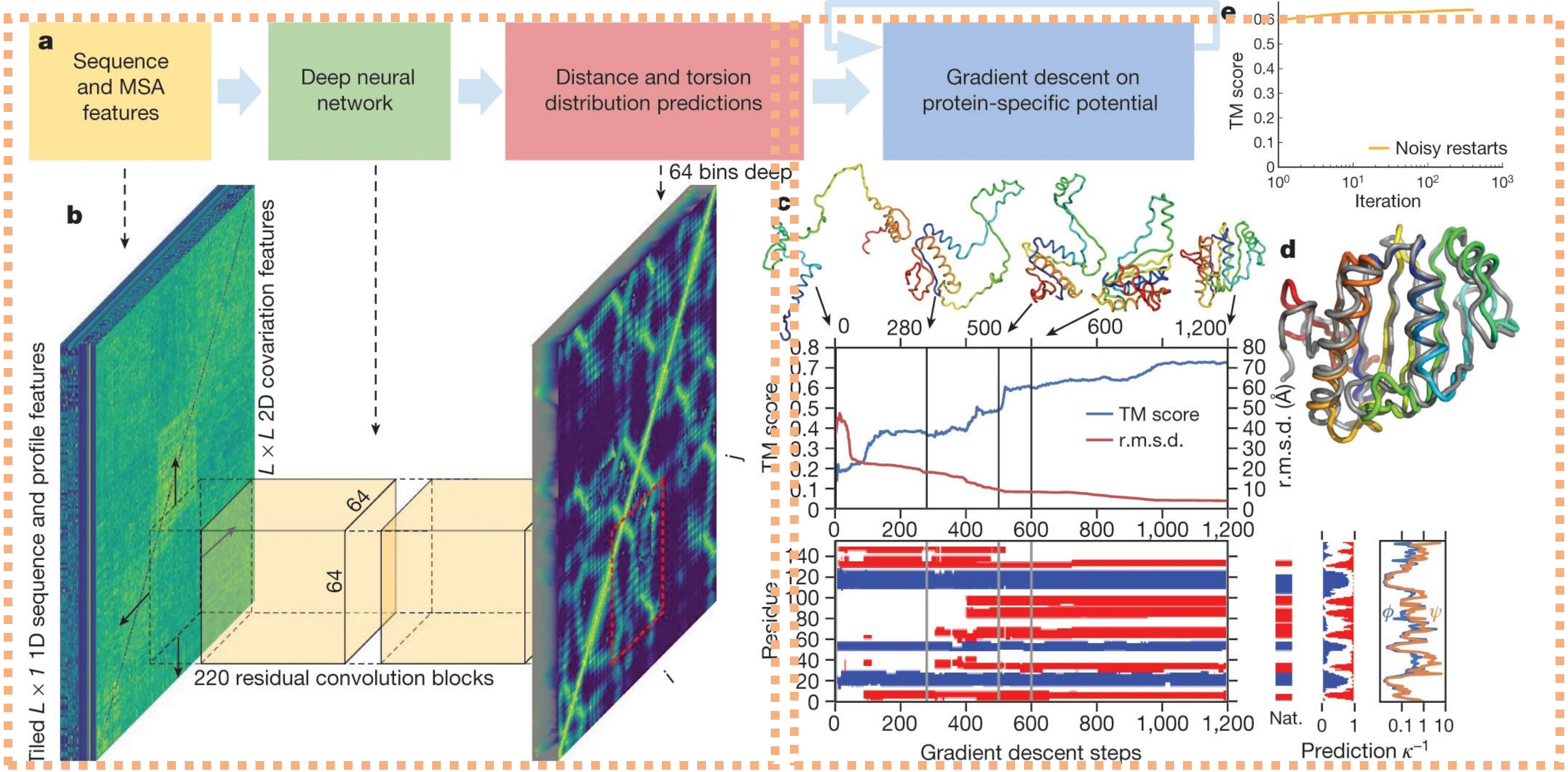
Coevolution from Multiple Sequence Alignments (MSA)



```
A T R L T L T A K K D G P C D
A T R L T L T A K K D G P C D
A T R L T L T A K K D G P C D
A T K L C L T A K K E G P K D
A T K L T L T A K K E G P K D
A T K L T L G A K K E G G C D
A T W L T L T A K K V G P C D
A T W L T L T A K K V G P C D
```

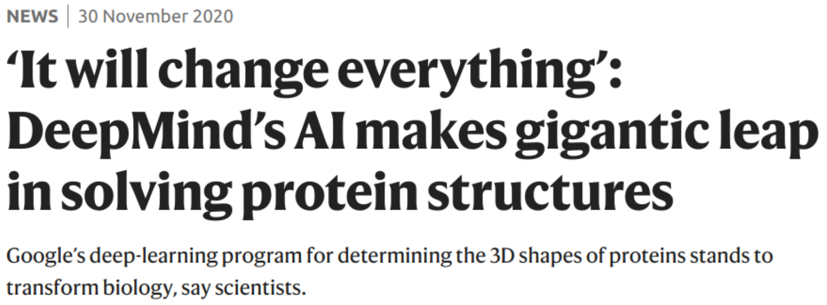
Marks, Debora S., et al. "Protein 3D structure computed from evolutionary sequence variation." *PloS one* 6.12 (2011): e28766.

Critical Assessment of Structure Prediction 2018 (CASP13) - AlphaFold1



Senior, Andrew W., et al. "Improved protein structure prediction using potentials from deep learning." Nature 577.7792 (2020): 706-710.

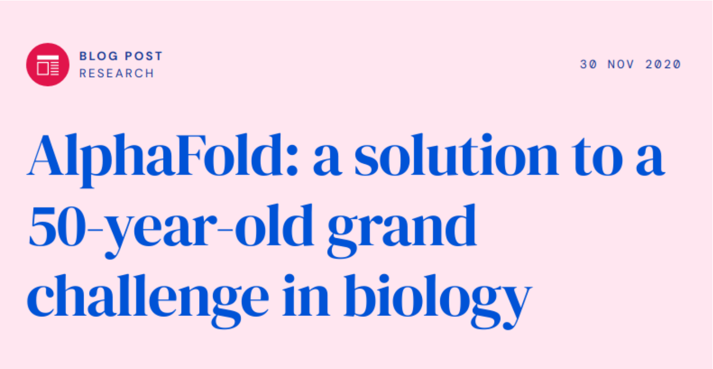
AlphaFold2



NEWS | 30 November 2020

‘It will change everything’: DeepMind’s AI makes gigantic leap in solving protein structures

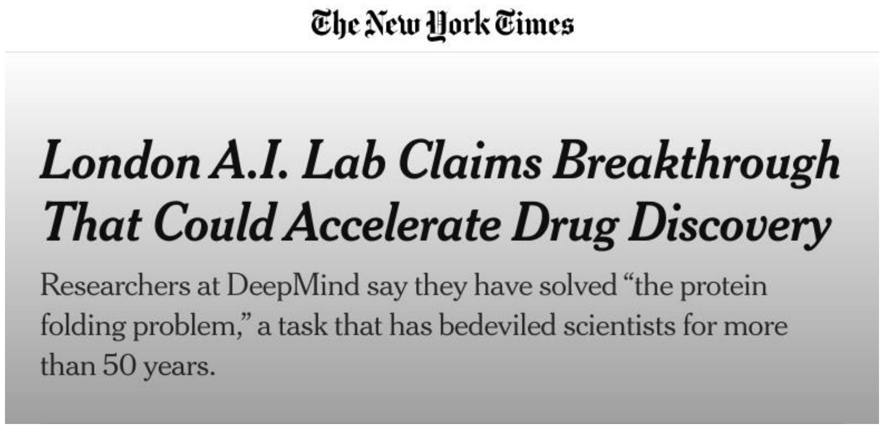
Google’s deep-learning program for determining the 3D shapes of proteins stands to transform biology, say scientists.



BLOG POST
RESEARCH

30 NOV 2020

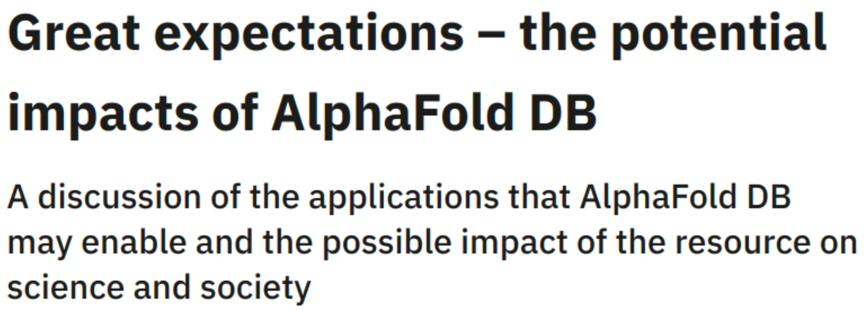
AlphaFold: a solution to a 50-year-old grand challenge in biology



The New York Times

London A.I. Lab Claims Breakthrough That Could Accelerate Drug Discovery

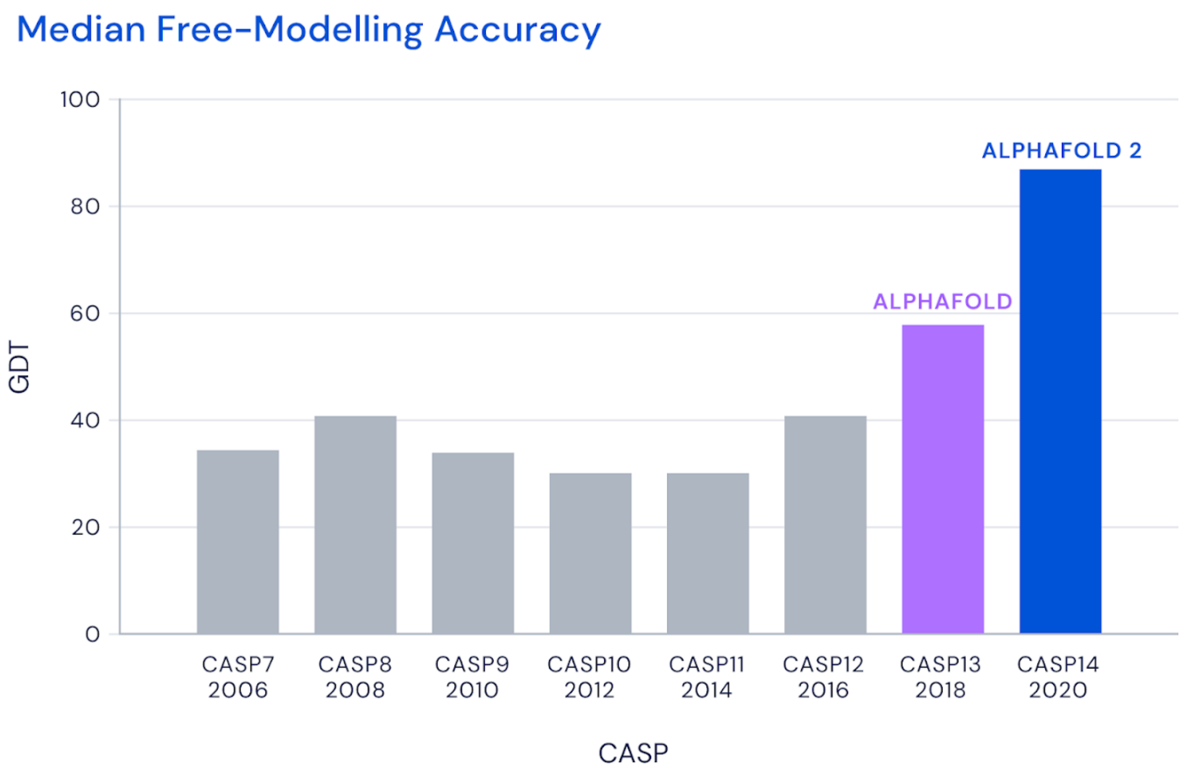
Researchers at DeepMind say they have solved “the protein folding problem,” a task that has bedeviled scientists for more than 50 years.

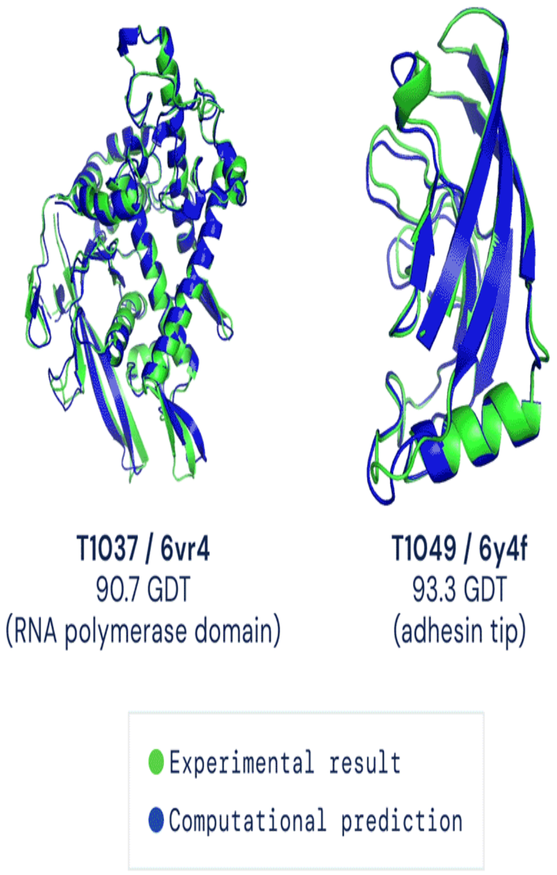


Great expectations – the potential impacts of AlphaFold DB

A discussion of the applications that AlphaFold DB may enable and the possible impact of the resource on science and society

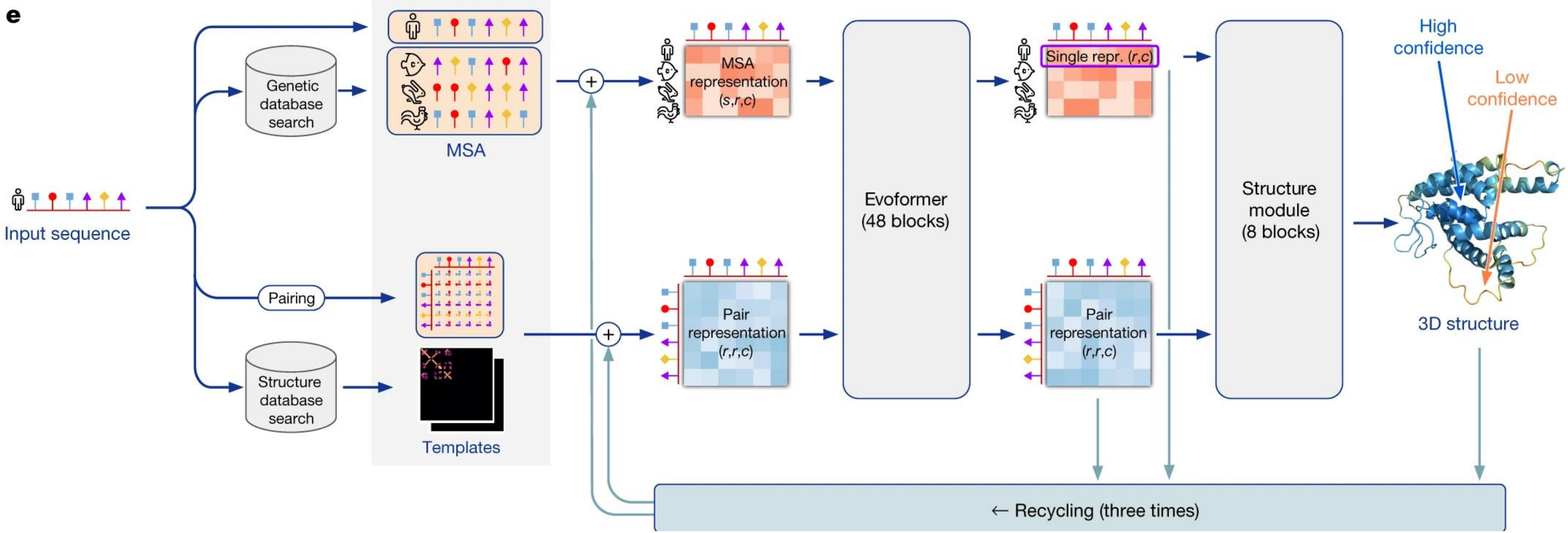
Critical Assessment of Structure Prediction 2020 (CASP14) - AlphaFold2




T1037 / 6vr4
90.7 GDT
(RNA polymerase domain)

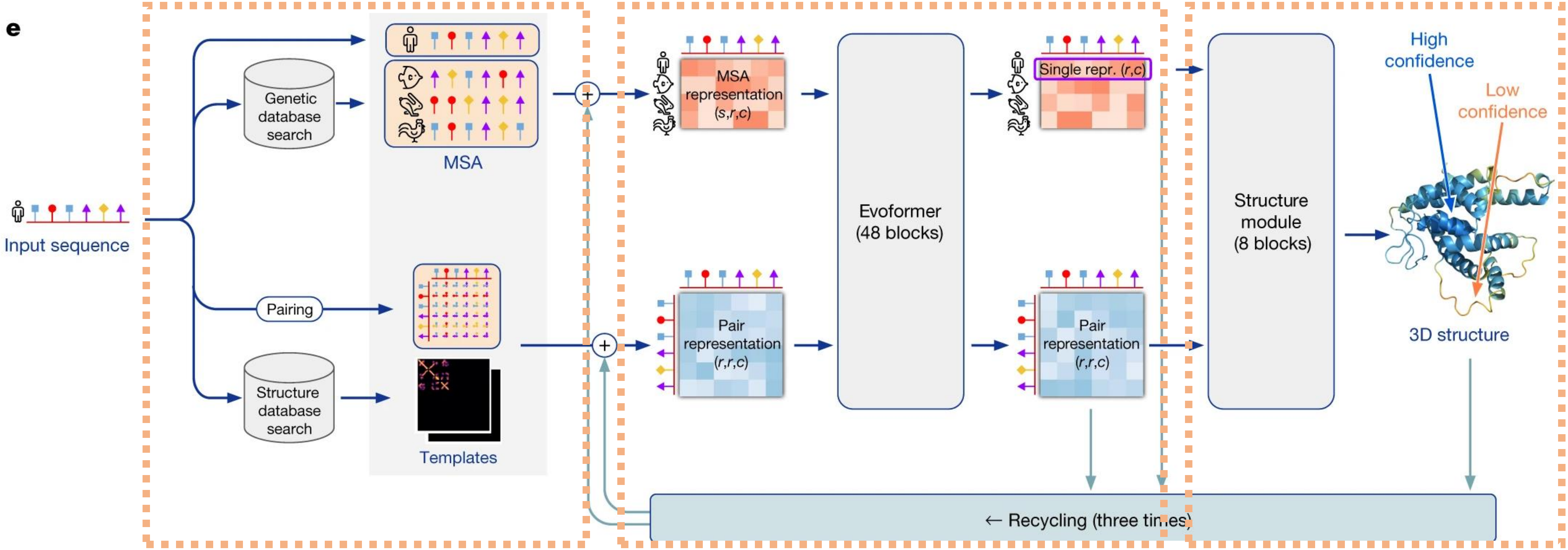
T1049 / 6y4f
93.3 GDT
(adhesin tip)

Critical Assessment of Structure Prediction 2020 (CASP14) - AlphaFold2


e
Input sequence
Genetic database search
Pairing
Structure database search
MSA
Templates
MSA representation (s,r,c)
Pair representation (r,r,c)
Evoformer (48 blocks)
Single repr. (r,c)
Pair representation (r,r,c)
Structure module (8 blocks)
High confidence
Low confidence
3D structure
← Recycling (three times)

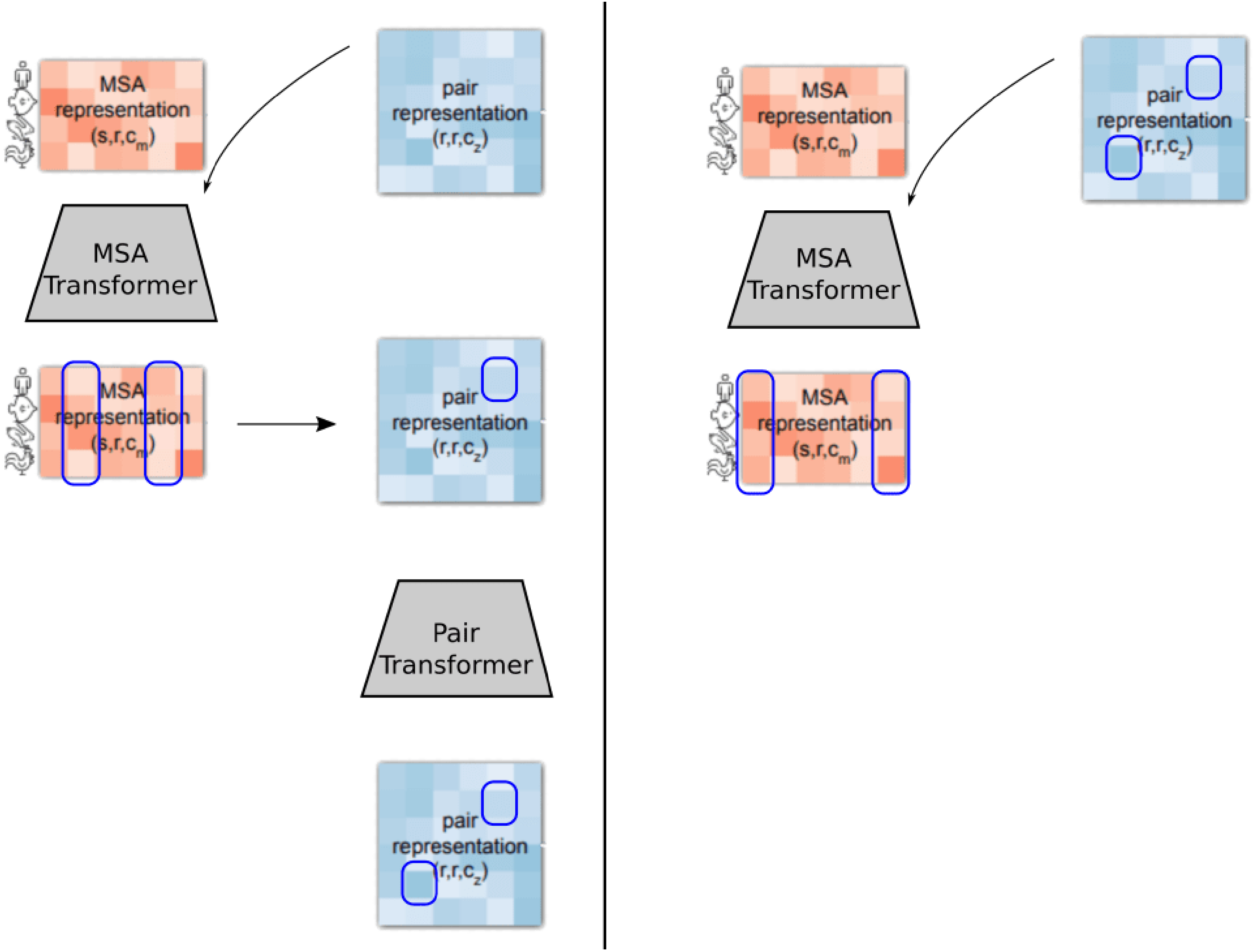
Jumper, John, et al. "Highly accurate protein structure prediction with AlphaFold." Nature 596.7873 (2021): 583-589.

Critical Assessment of Structure Prediction 2020 (CASP14) - AlphaFold2


e
Input sequence
Genetic database search
MSA
Pairing
Structure database search
Templates
MSA representation (s,r,c)
Pair representation (r,r,c)
Evoformer (48 blocks)
Single repr. (r,c)
Pair representation (r,r,c)
Structure module (8 blocks)
High confidence
Low confidence
3D structure
← Recycling (three times)

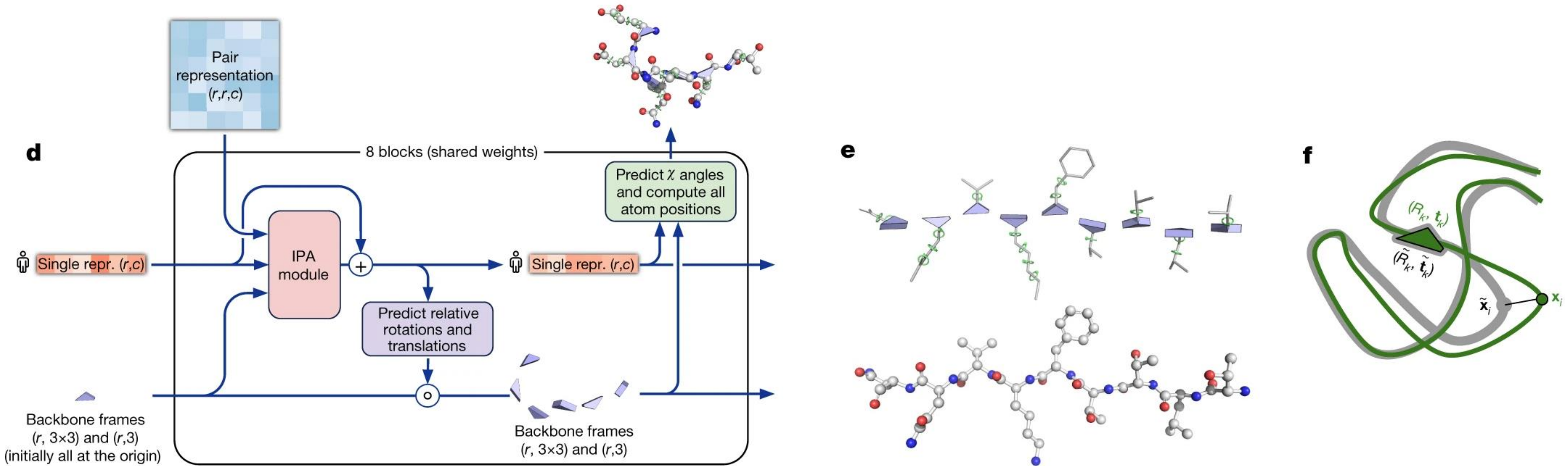
Jumper, John, et al. "Highly accurate protein structure prediction with AlphaFold." Nature 596.7873 (2021): 583-589.

AF2 - "Evolutionary Transformer" (Evoformer)


MSA representation (s,r,c_m)
pair representation (r,r,c_z)
MSA Transformer
MSA representation (s,r,c_m)
pair representation (r,r,c_z)
Pair Transformer
pair representation (r,r,c_z)
MSA representation (s,r,c_m)
pair representation (r,r,c_z)
MSA Transformer
MSA representation (s,r,c_m)

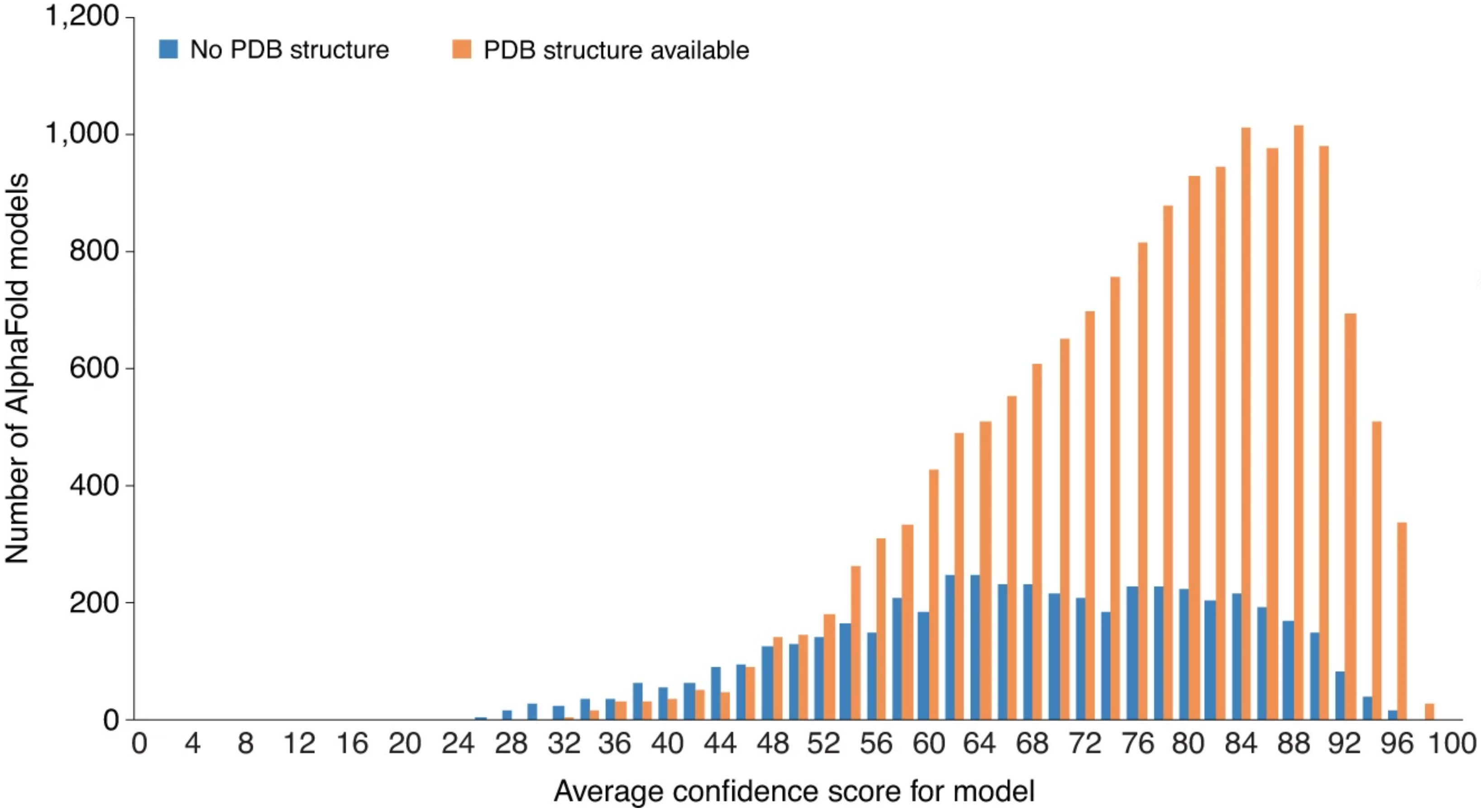
Jumper, John, et al. "Highly accurate protein structure prediction with AlphaFold." Nature 596.7873 (2021): 583-589.

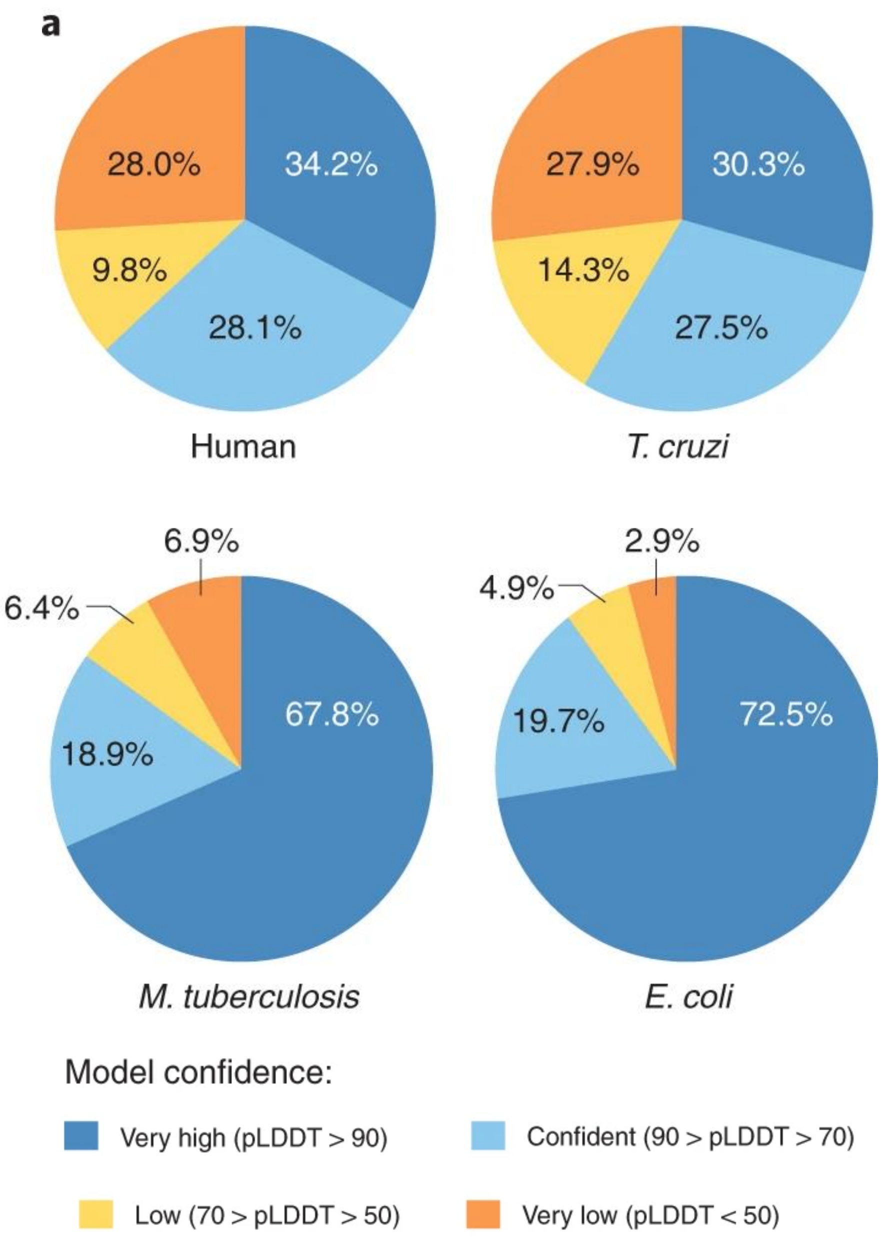
AF2 - Structure Module


d
Pair representation (r,r,c)
8 blocks (shared weights)
Single repr. (r,c)
IPA module
+
Predict relative rotations and translations
∘
Single repr. (r,c)
Predict χ angles and compute all atom positions
Backbone frames (r, 3×3) and (r,3) (initially all at the origin)
Backbone frames (r, 3×3) and (r,3)
e
f
$(R_k, \mathbf{t}_k)$
$(\tilde{R}_k, \tilde{\mathbf{t}}_k)$
$\tilde{\mathbf{x}}_i$
$\mathbf{x}_i$

Jumper, John, et al. "Highly accurate protein structure prediction with AlphaFold." Nature 596.7873 (2021): 583-589.

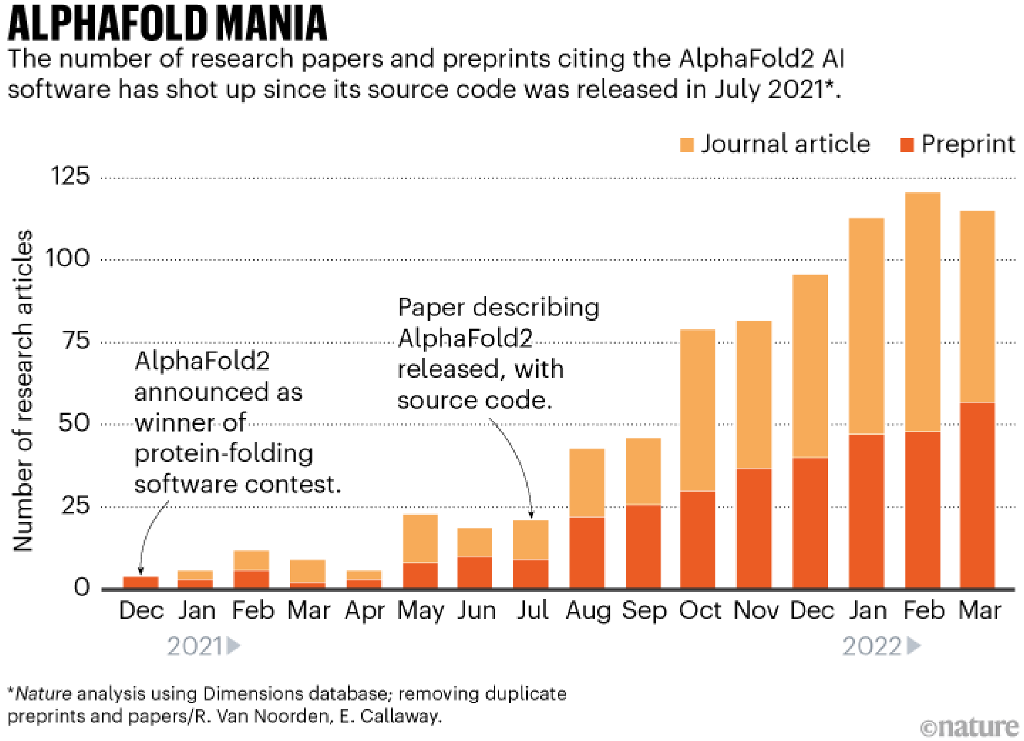
AF2 – Coverage and Quality of predictions





Thornton, J. M., Laskowski, R. A. & Borkakoti, N. AlphaFold heralds a data-driven revolution in biology and medicine. Nat Med 27, 1666–1669 (2021).

AF2 – Most recent updates



ALPHAFOLD MANIA

The number of research papers and preprints citing the AlphaFold2 AI software has shot up since its source code was released in July 2021*.

*Nature analysis using Dimensions database; removing duplicate preprints and papers/R. Van Noorden, E. Callaway.

©nature

- AF2-Multimer
  - Prediction of protein-protein interactions
- Release of over structures for over 200k different proteins
  - Estimated human proteome ~25k
- Limitations
  - Min of 16 and max of 2,700 amino acids
    - Model smaller overlapping fragments
    - Alignment of predicted structures
    - "Glue" fragments on overlapping regions

Thank you!

Adam Taranto this Thursday! (31st August)

Ryan Wick next Tuesday! (5th September)

Assignment 2 released tomorrow! (Wednesday 30th @ midnight)

Today: Genomic Features & Regions

Next time: Genomic Intervals (actual algorithms)