Artificial Intelligence 前沿论文解读

Highly Accurate Protein Structure Prediction with AlphaFold

报告人：陈星强

论文作者信息

Affiliations

- **DeepMind, London, UK**
  John Jumper, Richard Evans, Alexander Pritzel, Tim Green,
- **Seoul National University, Seoul, South Korea**
  Martin Steinegger

Correspondence to John Jumper or Demis Hassabis.

Subjects

- Computational biophysics
- Machine learning
- Protein structure predictions
- Structural biology

摘要部分

Abstract

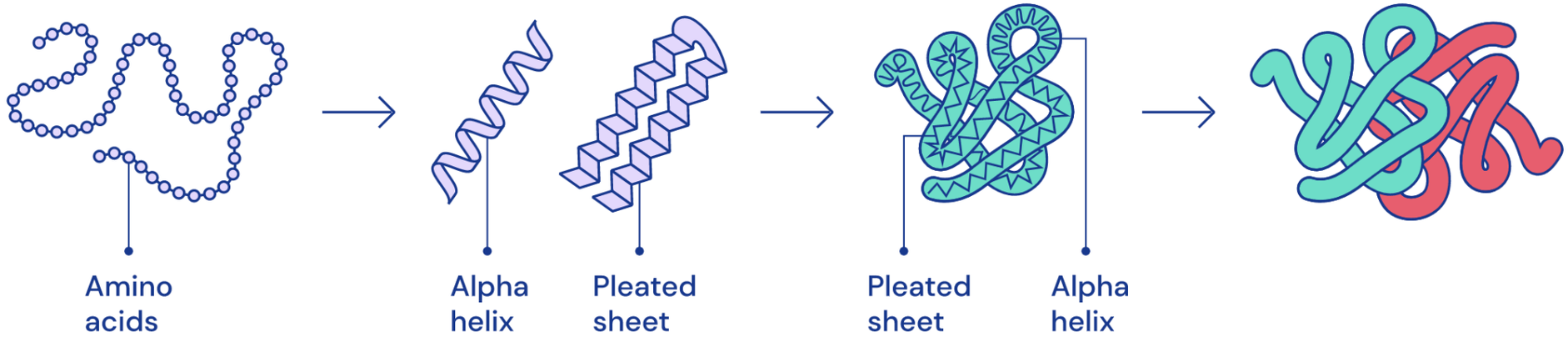
Abstract

Every protein is made up of a sequence of amino acids bonded together

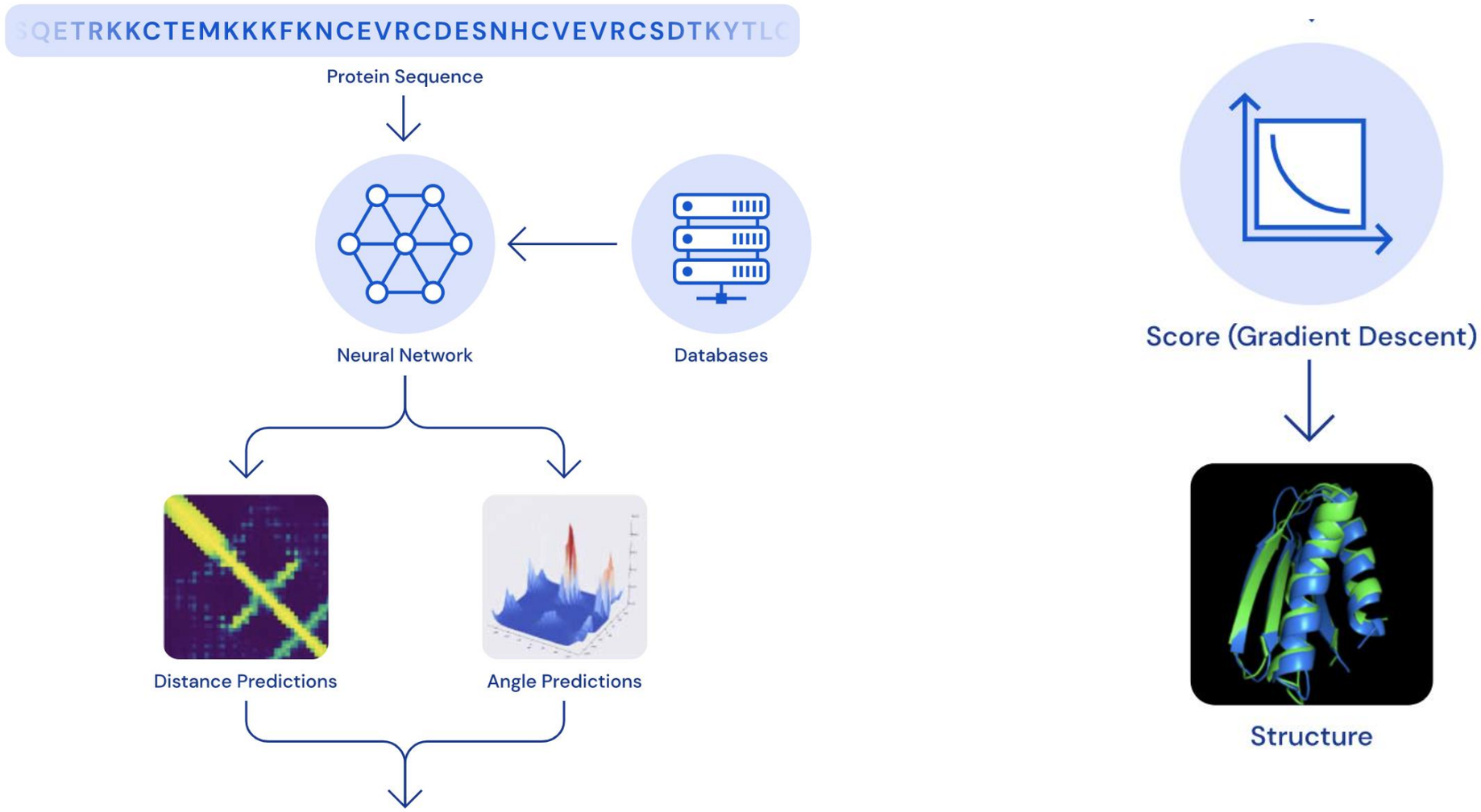
These amino acids interact locally to form shapes like helices and sheets

These shapes fold up on larger scales to form the full three-dimensional protein structure

Proteins can interact with other proteins, performing functions such as signalling and transcribing DNA



Abstract



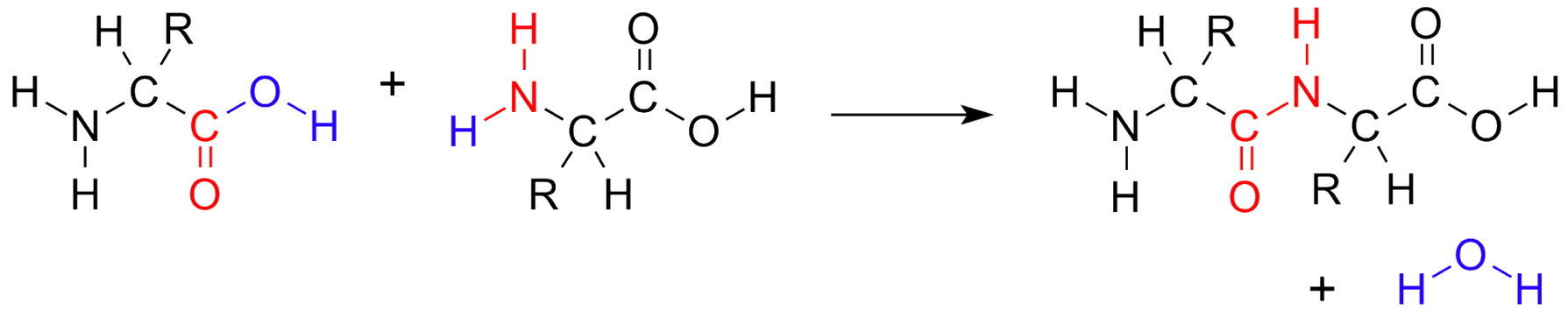
A schematic of the architecture of the AlphaFold system predicting structure from protein sequence.

Notation

基础概念学习

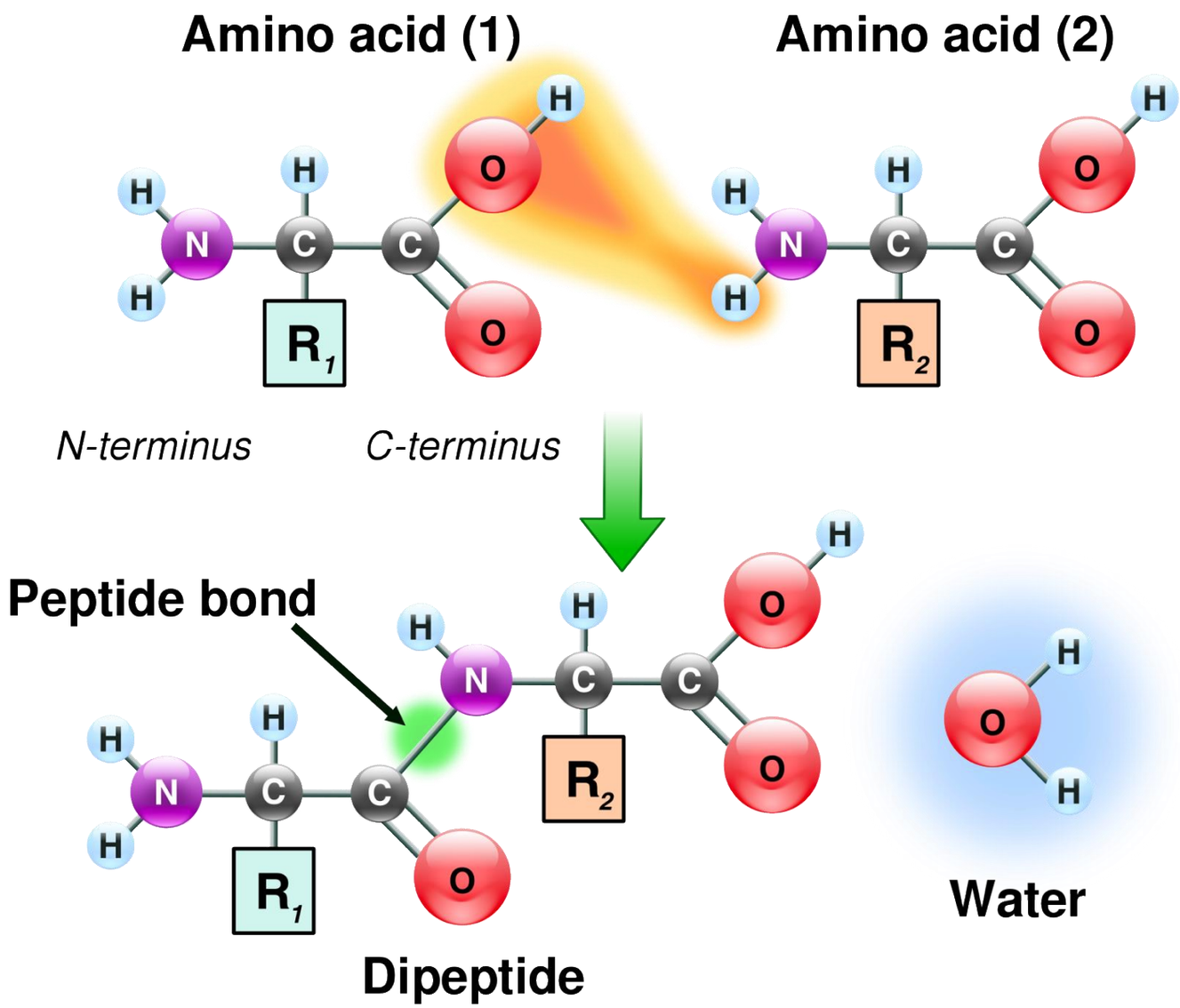
1. Amino acid
2. Pipetide bond
3. Chi Angle
3. lDDT-Cα
4. template modeling score (TM-score)
5. multiple sequence alignments (MSAs)
6. pairwise features
7. homologues
8. IPA
9.Evoformer
10. pLDDT per-residue accuracy of the structure
11. TM-score pTM
12. Pair-wise error prediction
13. Frame-aligned point error FAPE

2. 肽键



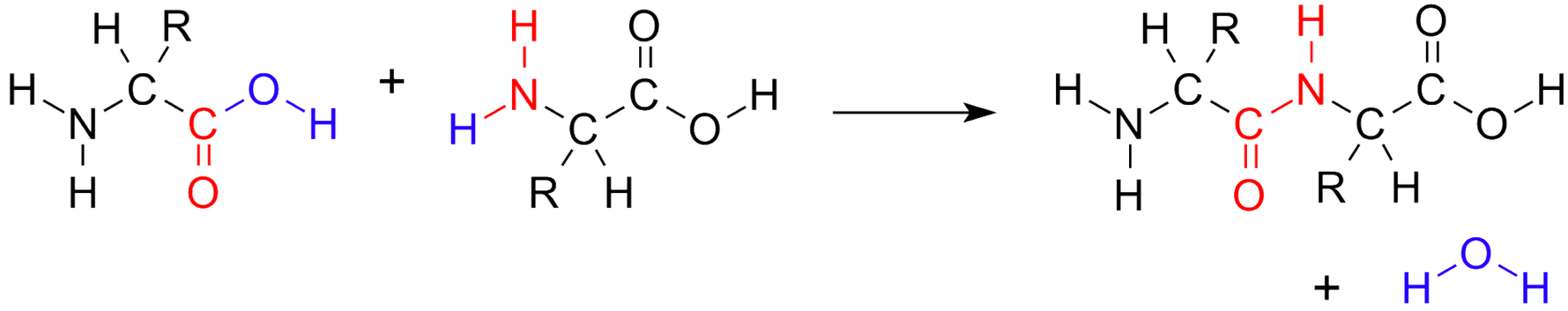
https://en.wikipedia.org/wiki/Amino_acid

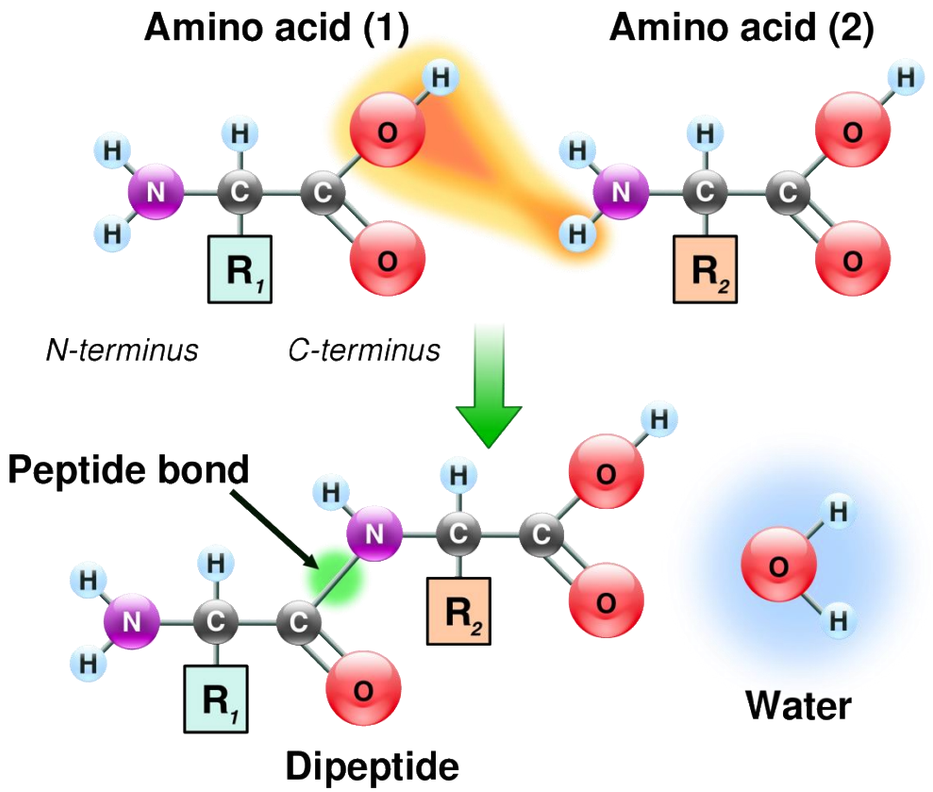
基础信息 **概念解**释 肽键

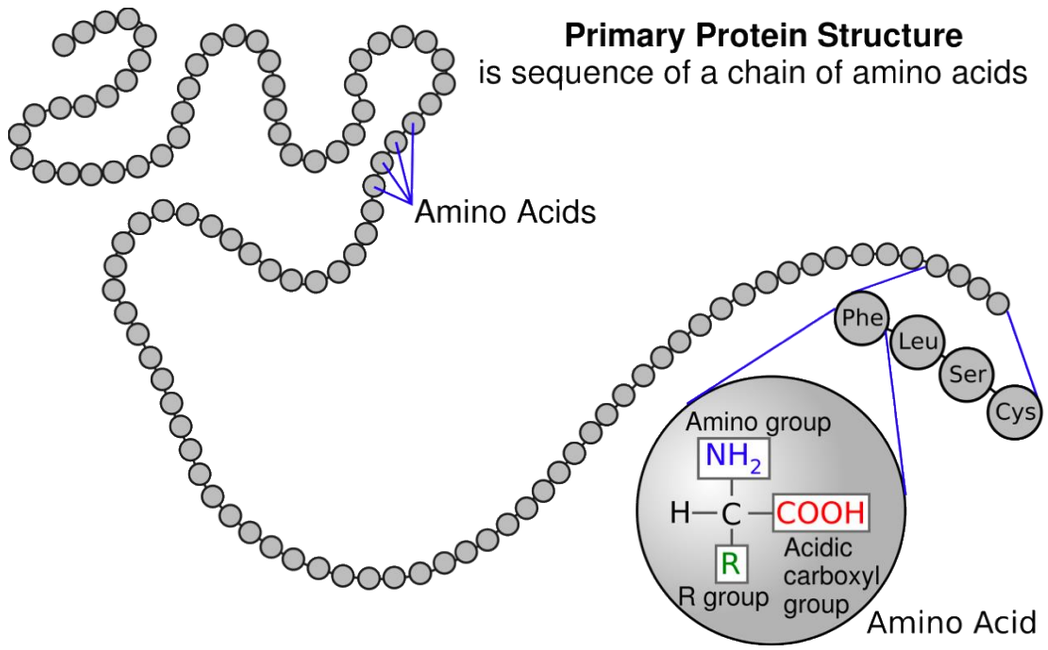


https://en.wikipedia.org/wiki/Peptide_bond

基础信息 **概念解**释 肽键

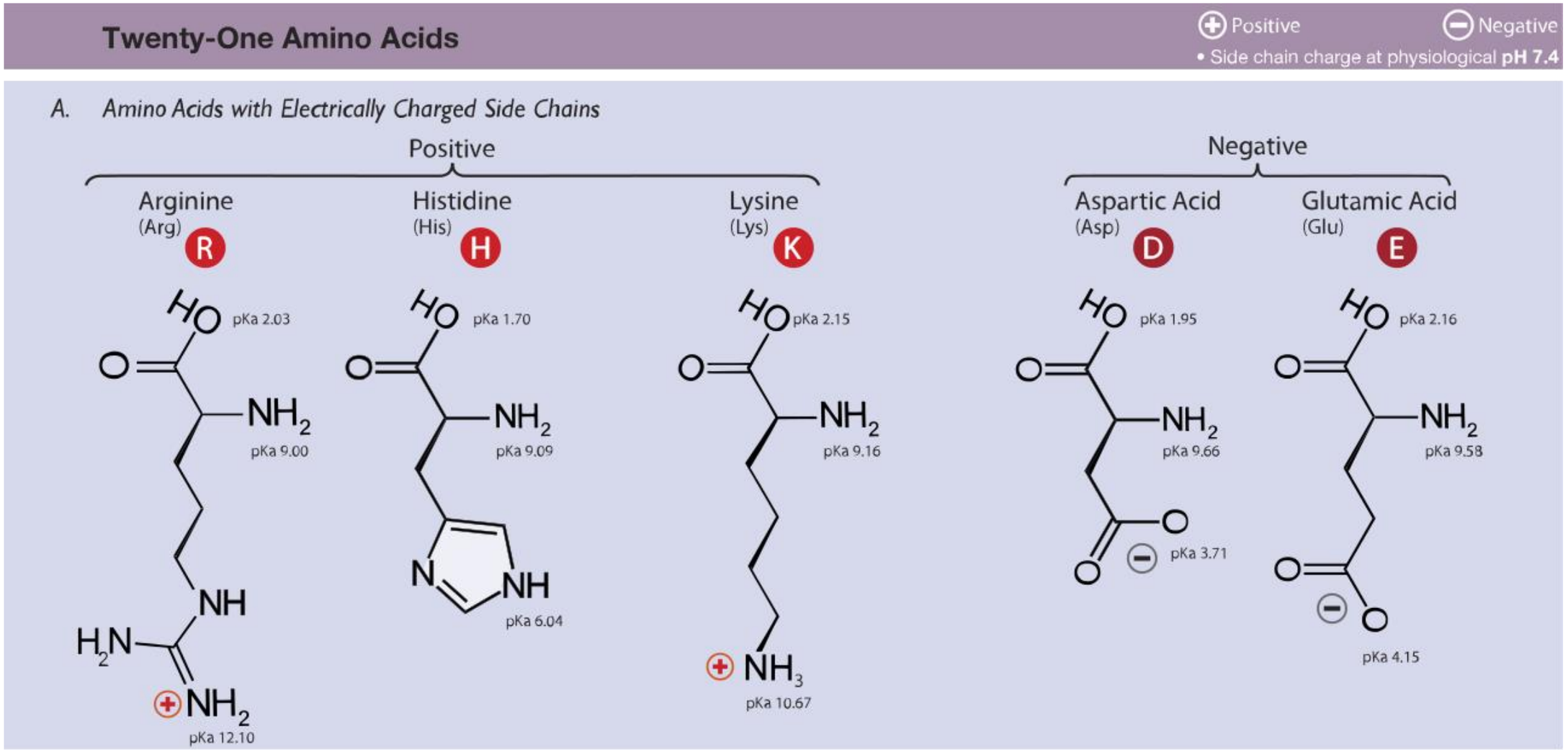






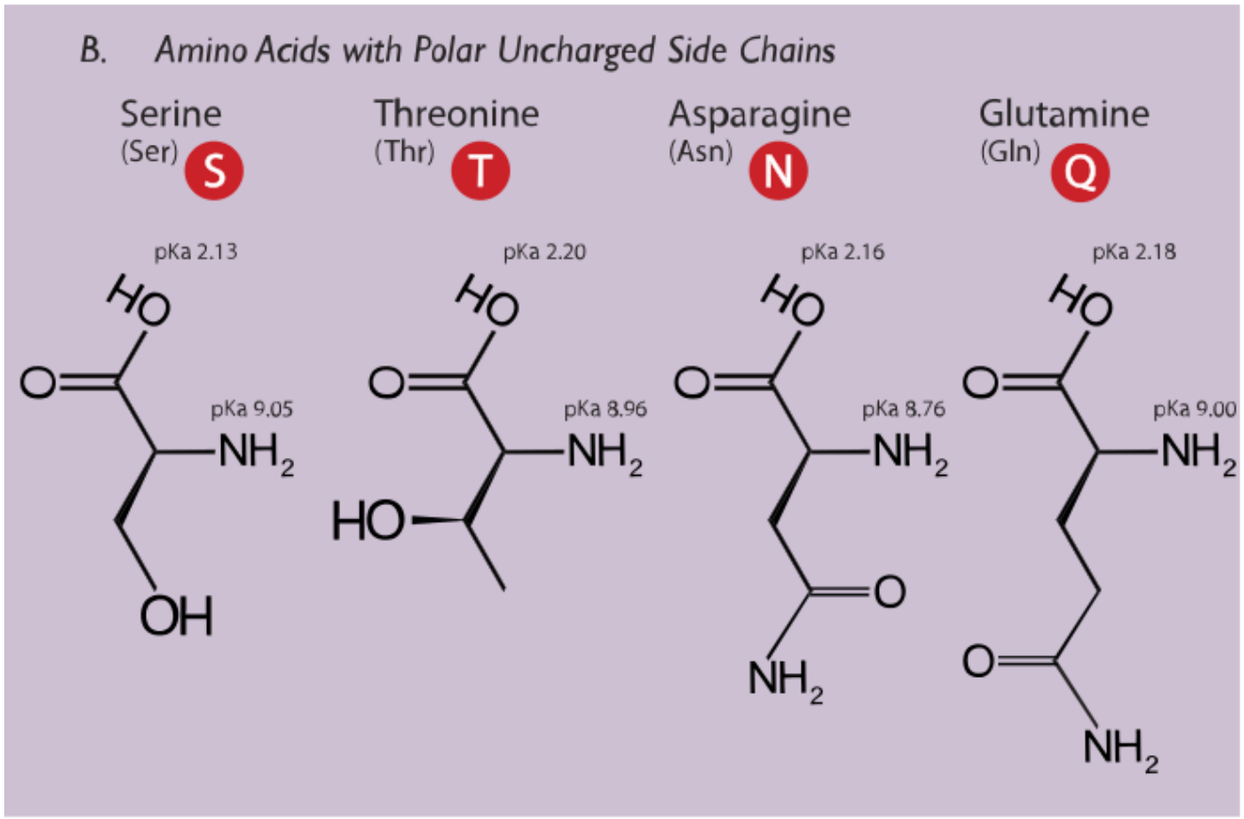
https://en.wikipedia.org/wiki/Amino_acid https://en.wikipedia.org/wiki/Peptide_bond

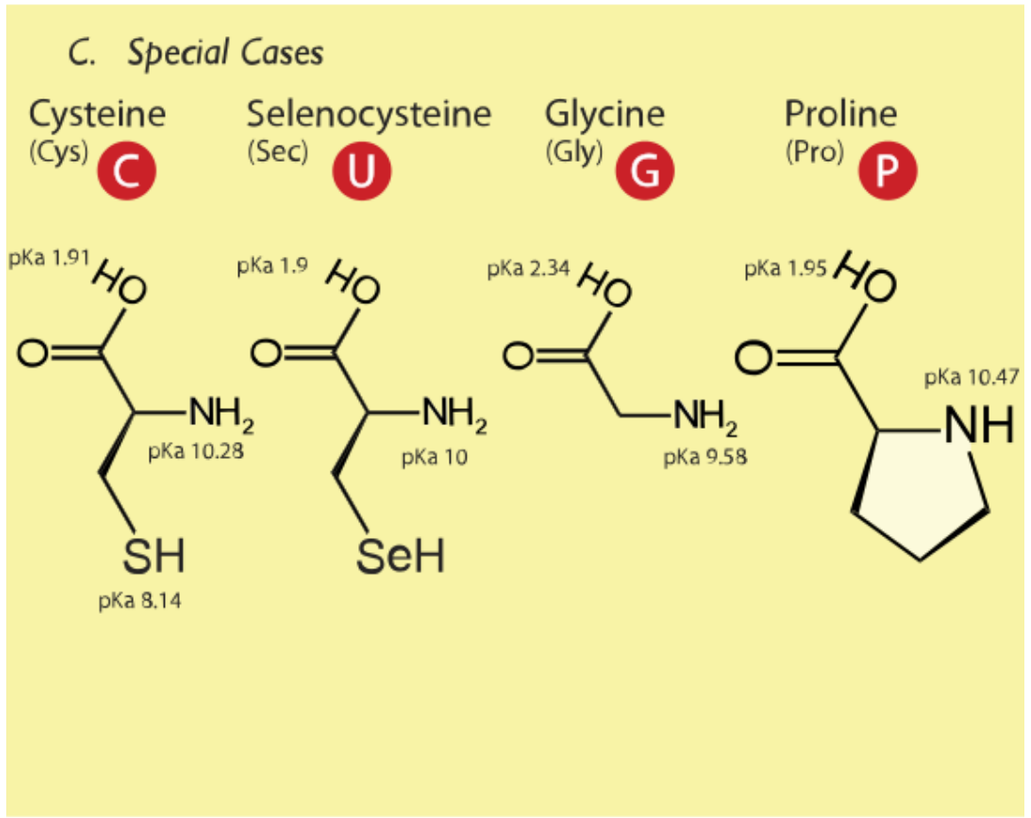
1. 氨基酸



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

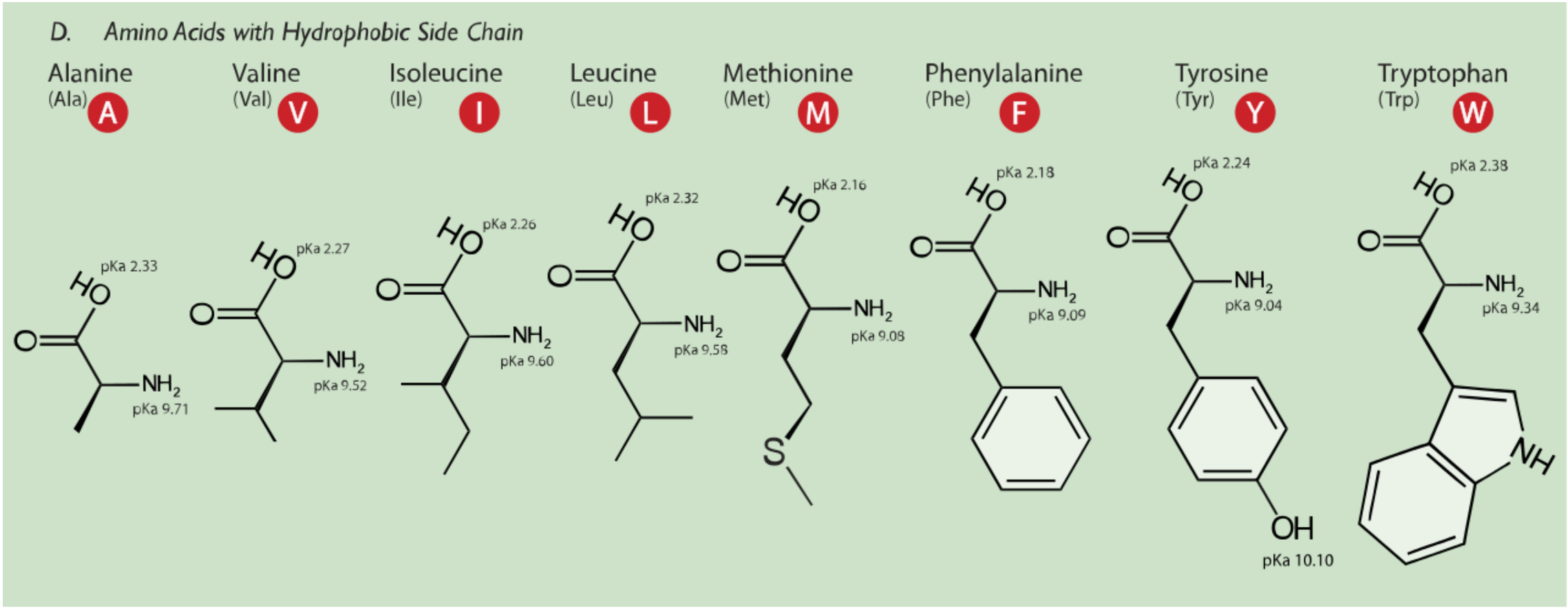
基础信息 概念解释 氨基酸





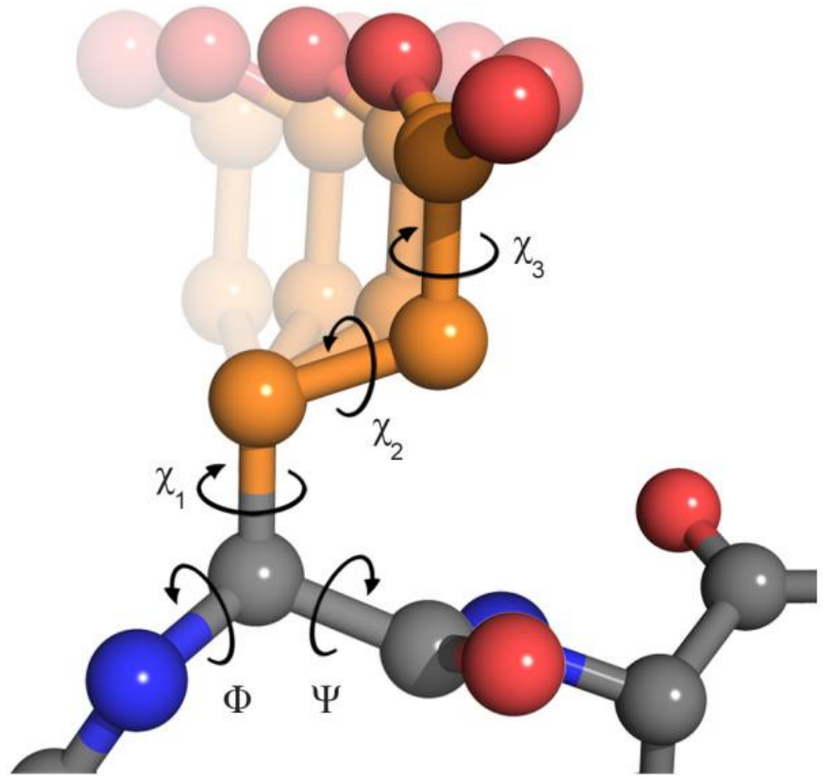
https://en.wikipedia.org/wiki/Amino_acid

基础信息 概念解释 氨基酸



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

3. Chi Angle



Dihedral angles in glutamate

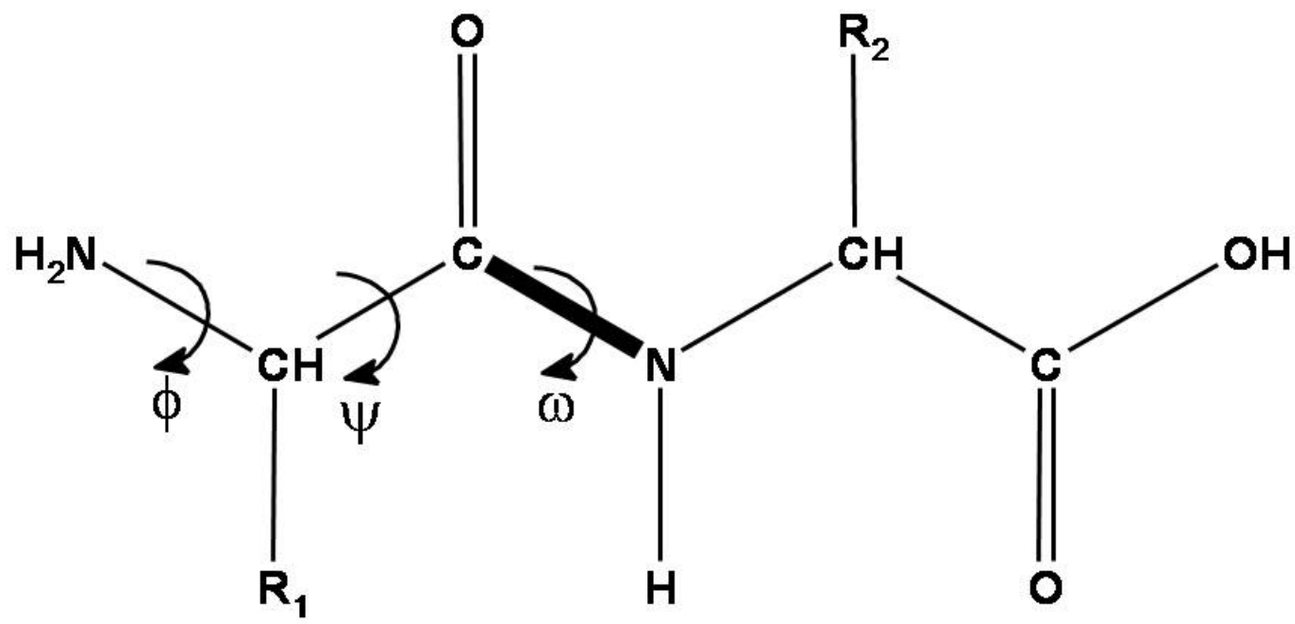
http://www.biomedcentral.com/1471-2105/11/306

https://swissmodel.expasy.org/course/text/chapter3.htm

http://www.mlb.co.jp/linux/science/garlic/doc/commands/dihedrals.html

http://www.mlb.co.jp/linux/science/garlic/doc/commands/dihedrals.html

3. Chi Angle



The angle between two bonds

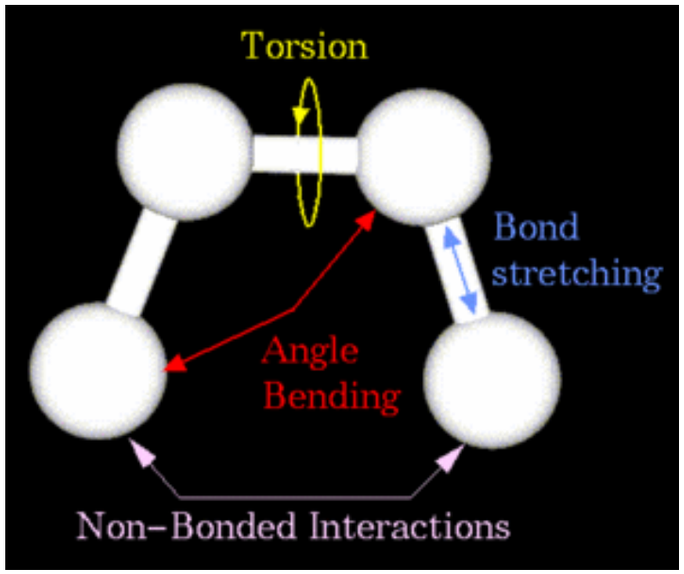
The backbone of a protein has three different torsion angles.

- The phi-angle (φ) - around the N-Cα bond
- The psi-angle (ψ) - around the Cα-C bond
- The omega-angle (ω) - around the peptide bond between C and N.

3. Chi Angle

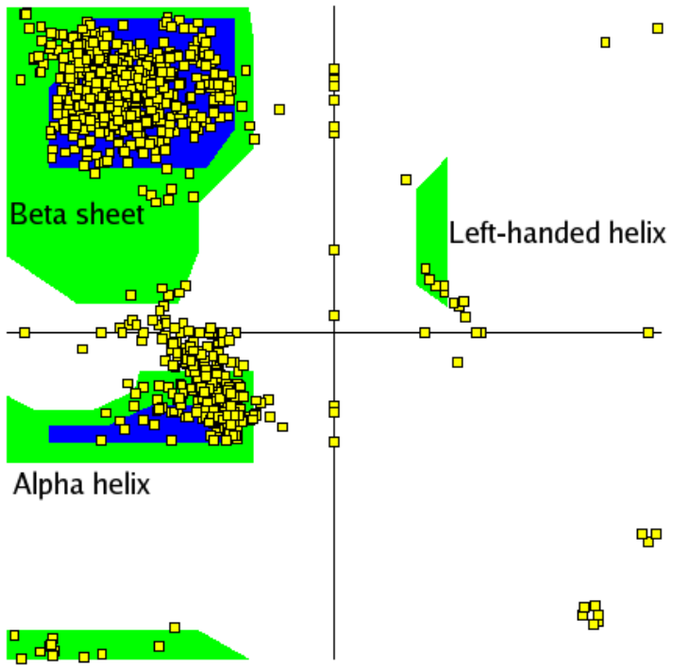
The ω-bond has a slightly double-bond character and is therefore almost always 180 degrees.

The structure of a protein is mainly formed by the φ- and ψ-angles.



Every [secondary structure] element forces the [backbone] into a specific range of torsion angles, this can be visualized in a [ramachandran plot].

The figure shows the location of the φ, ψ and ω-angles.



[Side chains] can also contain torsion angles, they are indicated as [chi-angles].

3. Chi Angle

Table 2 | Rigid groups for constructing all atoms from given torsion angles. Boxes highlight groups that are symmetric under 180° rotations.

aatype	bb	ψ	χ_1	χ_2	χ_3	χ_4
ALA	N, C^{α}, C, C^{β}	O	-	-	-	-
ARG	N, C^{α}, C, C^{β}	O	C^{γ}	C^{δ}	N^{ϵ}	$N^{\eta 1}$, $N^{\eta 2}$, C^{ζ}
ASN	N, C^{α}, C, C^{β}	O	C^{γ}	$N^{\delta 2}$, $O^{\delta 1}$	-	-
ASP	N, C^{α}, C, C^{β}	O	C^{γ}	$\boxed{O^{\delta 1}, O^{\delta 2}}$	-	-
CYS	N, C^{α}, C, C^{β}	O	S^{γ}	-	-	-
GLN	N, C^{α}, C, C^{β}	O	C^{γ}	C^{δ}	$N^{\epsilon 2}$, $O^{\epsilon 1}$	-
GLU	N, C^{α}, C, C^{β}	O	C^{γ}	C^{δ}	$\boxed{O^{\epsilon 1}, O^{\epsilon 2}}$	-
GLY	N, C^{α}, C	O	-	-	-	-
HIS	N, C^{α}, C, C^{β}	O	C^{γ}	$C^{\delta 2}$, $N^{\delta 1}$, $C^{\epsilon 1}$, $N^{\epsilon 2}$	-	-
ILE	N, C^{α}, C, C^{β}	O	$C^{\gamma 1}$, $C^{\gamma 2}$	$C^{\delta 1}$	-	-
LEU	N, C^{α}, C, C^{β}	O	C^{γ}	$C^{\delta 1}$, $C^{\delta 2}$	-	-
LYS	N, C^{α}, C, C^{β}	O	C^{γ}	C^{δ}	C^{ϵ}	N^{ζ}
MET	N, C^{α}, C, C^{β}	O	C^{γ}	S^{δ}	C^{ϵ}	-
PHE	N, C^{α}, C, C^{β}	O	C^{γ}	$\boxed{C^{\delta 1}, C^{\delta 2}, C^{\epsilon 1}, C^{\epsilon 2}, C^{\zeta}}$	-	-
PRO	N, C^{α}, C, C^{β}	O	C^{γ}	C^{δ}	-	-
SER	N, C^{α}, C, C^{β}	O	O^{γ}	-	-	-
THR	N, C^{α}, C, C^{β}	O	$C^{\gamma 2}$, $O^{\gamma 1}$	-	-	-
TRP	N, C^{α}, C, C^{β}	O	C^{γ}	$C^{\delta 1}$, $C^{\delta 2}$, $C^{\epsilon 2}$, $C^{\epsilon 3}$, $N^{\epsilon 1}$, $C^{\eta 2}$, $C^{\zeta 2}$, $C^{\zeta 3}$	-	-
TYR	N, C^{α}, C, C^{β}	O	C^{γ}	$\boxed{C^{\delta 1}, C^{\delta 2}, C^{\epsilon 1}, C^{\epsilon 2}, O^{\eta}, C^{\zeta}}$	-	-
VAL	N, C^{α}, C, C^{β}	O	$C^{\gamma 1}$, $C^{\gamma 2}$	-	-	-

基础概念学习

N_{res} : the number of residues in the input primary sequence (cropped during training).

N_{templ} : the number of templates used in the model.

N_{all_seq} : the number of all available MSA sequences.

N_{clust} : the number of clusters after MSA clustering.

N_{seq} : the number of sequences processed in the MSA stack (where $N_{seq} = N_{clust} + N_{templ}$).

N_{extra_seq} : the number of unclustered MSA sequences (after sub-sampling).

N_{block} : the number of blocks in Evoformer-like stacks.

$N_{ensemble}$: the number of ensembling iterations.

N_{cycle} : the number of recycling iterations.

基础概念学习

Table 4 | AlphaFold training protocol. We train each stage until convergence with the approximate timings and number of samples provided.

Model	Initial training	Fine-tuning
Number of templates N_{templ}	4	4
Sequence crop size N_{res}	256	384
Number of sequences N_{seq}	128	512
Number of extra sequences $N_{\text{extra\_seq}}$	1024	5120
Parameters initialized from	Random	Initial training
Initial learning rate	10^{-3}	$5 \cdot 10^{-4}$
Learning rate linear warm-up samples	128000	0
Structural violation loss weight	0.0	1.0
Training samples ($\cdot 10^6$)	≈ 10	≈ 1.5
Training time	≈ 7 days	≈ 4 days

算法列表

List of Algorithms

List of Algorithms

	Function Name	Instruction
1	MSABlockDeletion	**MSA block deletion**
2	Inference	**AlphaFold Model Inference**
3	InputEmbedder	**Embeddings for initial representations**
4	relpos	**Relative position encoding**
5	one_hot	**One-hot encoding with nearest bin**
6	EvoformerStack	**Evoformer stack**
7	MSARowAttentionWithPairBias	**MSA row-wise gated self-attention with pair bias**
8	MSAColumnAttention	**MSA column-wise gated self-attention**
9	MSATransition	**Transition layer in the MSA stack**
10	OuterProductMean	**Outer product mean**
11	TriangleMultiplicationOutgoing	**Triangular multiplicative update using “outgoing” edges**
12	TriangleMultiplicationIncoming	**Triangular multiplicative update using “incoming” edges**
13	TriangleAttentionStartingNode	**Triangular gated self-attention around starting node**
14	TriangleAttentionEndingNode	**Triangular gated self-attention around ending node**
15	PairTransition	**Transition layer in the pair stack**

AlphaFold Model Inference

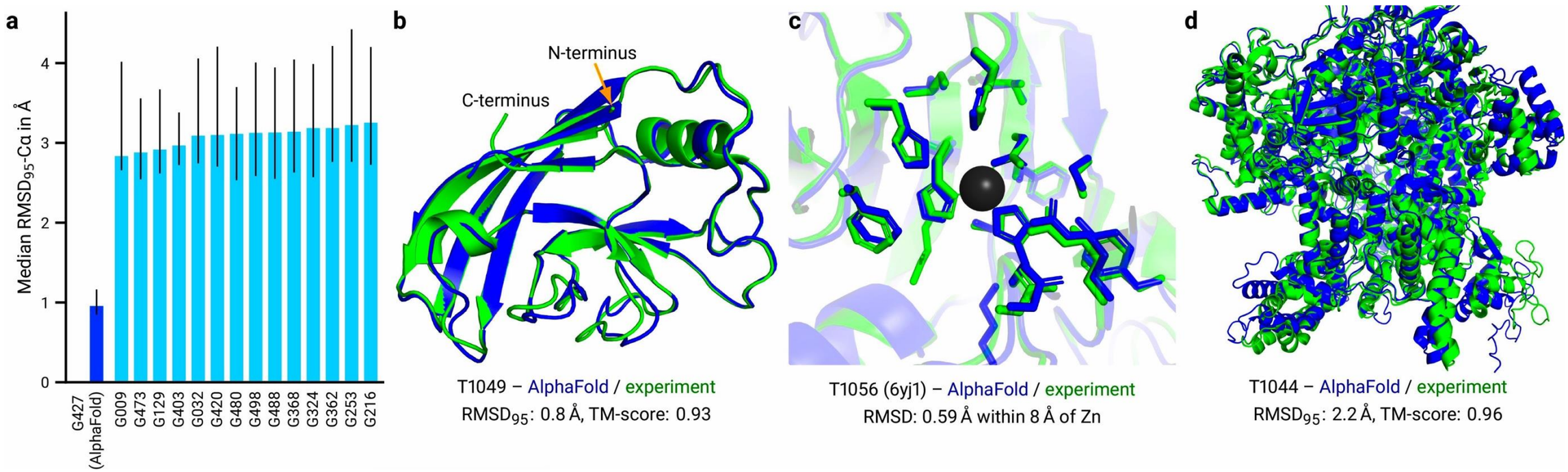
List of Algorithms

	Function Name	Instruction
16	TemplatePairStack	**Template pair stack**
17	TemplatePointwiseAttention	**Template pointwise attention**
18	ExtraMsaStack	**Extra MSA stack**
19	MSAColumnGlobalAttention	**MSA global column-wise gated self-attention**
20	StructureModule	**Structure module**
21	rigidFrom3Points	**Rigid from 3 points using the Gram–Schmidt process**
22	InvariantPointAttention	**Invariant point attention(IPA)**
23	BackboneUpdate	**Backbone update**
24	computeAllAtomCoordinates	**Compute all atom coordinates**
25	makeRotX	**Make a transformation that rotates around the x-axis**
26	renameSymmetricGroundTruthAtoms	**Rename symmetric ground truth atoms**
27	torsionAngleLoss	**Side chain and backbone torsion angle loss**
28	computeFAPE	**Compute the Frame aligned point error**
29	predictPerResidueLDDT	**Predict model confidence pLDDT**
30	RecyclingInference	**Generic recycling inference procedure**
31	RecyclingTraining	**Generic recycling training procedure**
32	RecyclingEmbedder	**Embedding of Evoformer and Structure module outputs for recycling**

论文背景介绍

Introduction

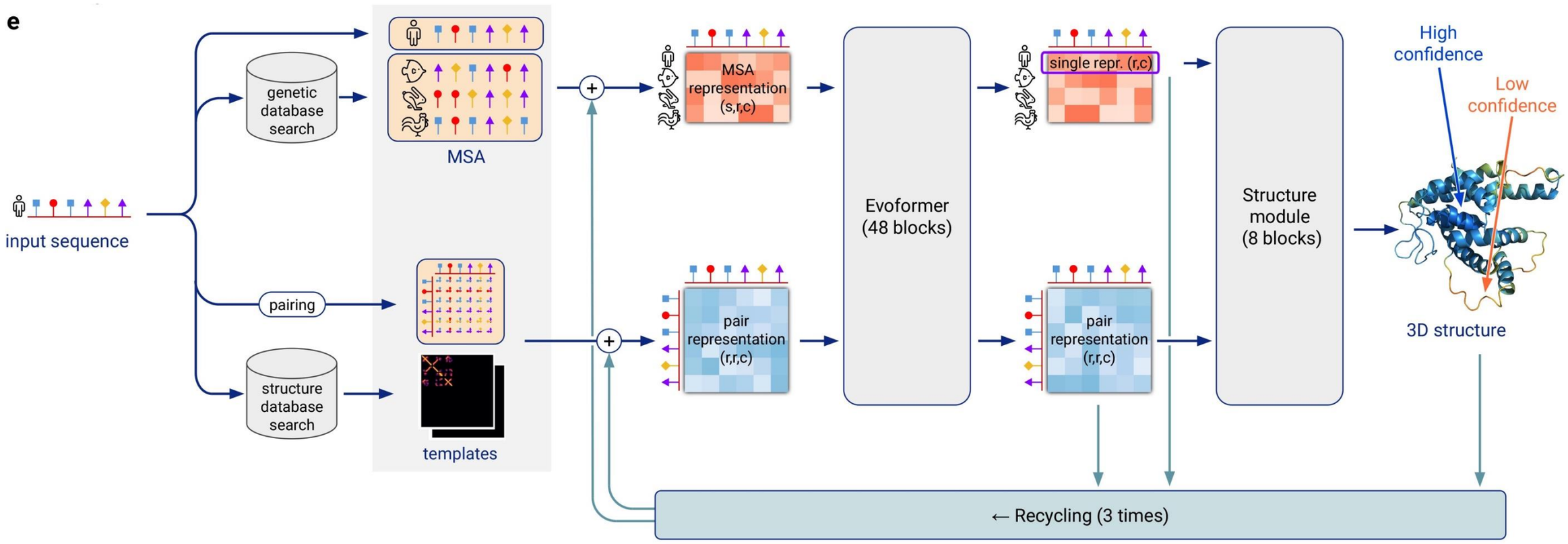
基础信息 模型效果


a
b
c
d
Median RMSD95-Cα in Å
0
1
2
3
4
G427 (AlphaFold)
G009
G473
G129
G403
G032
G420
G480
G498
G488
G368
G324
G362
G253
G216
N-terminus
C-terminus
T1049 – AlphaFold / experiment
RMSD95: 0.8 Å, TM-score: 0.93
T1056 (6yj1) – AlphaFold / experiment
RMSD: 0.59 Å within 8 Å of Zn
T1044 – AlphaFold / experiment
RMSD95: 2.2 Å, TM-score: 0.96

论文方法介绍

AlphaFlod network

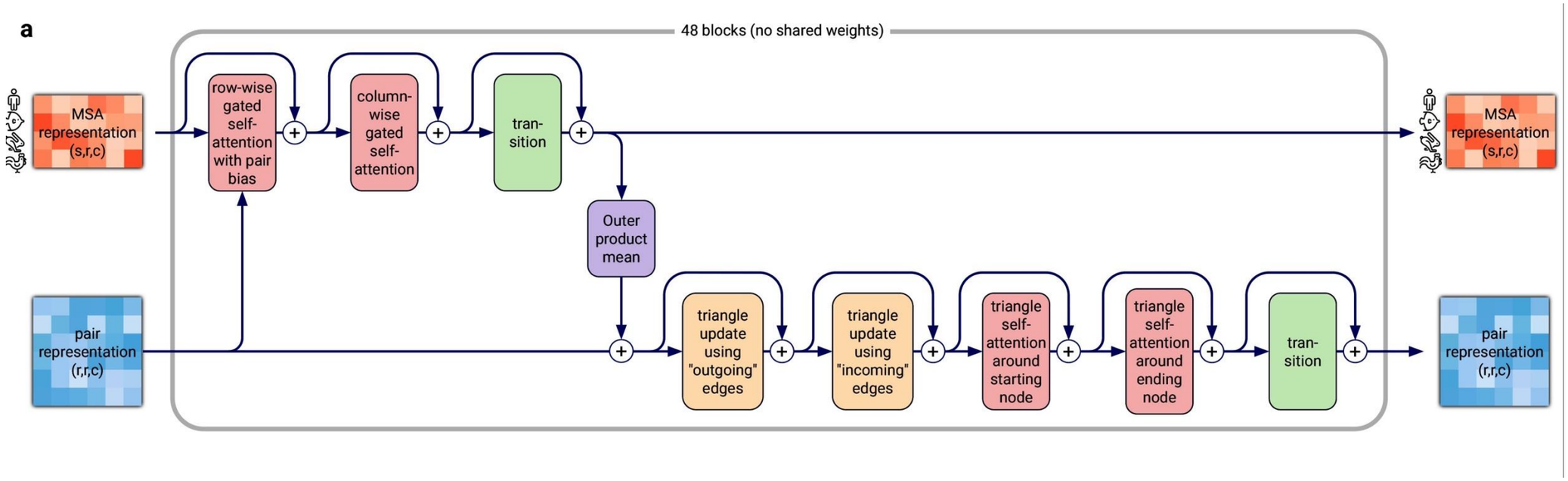
基础信息 模型框架


e
genetic database search
MSA
input sequence
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 (3 times)

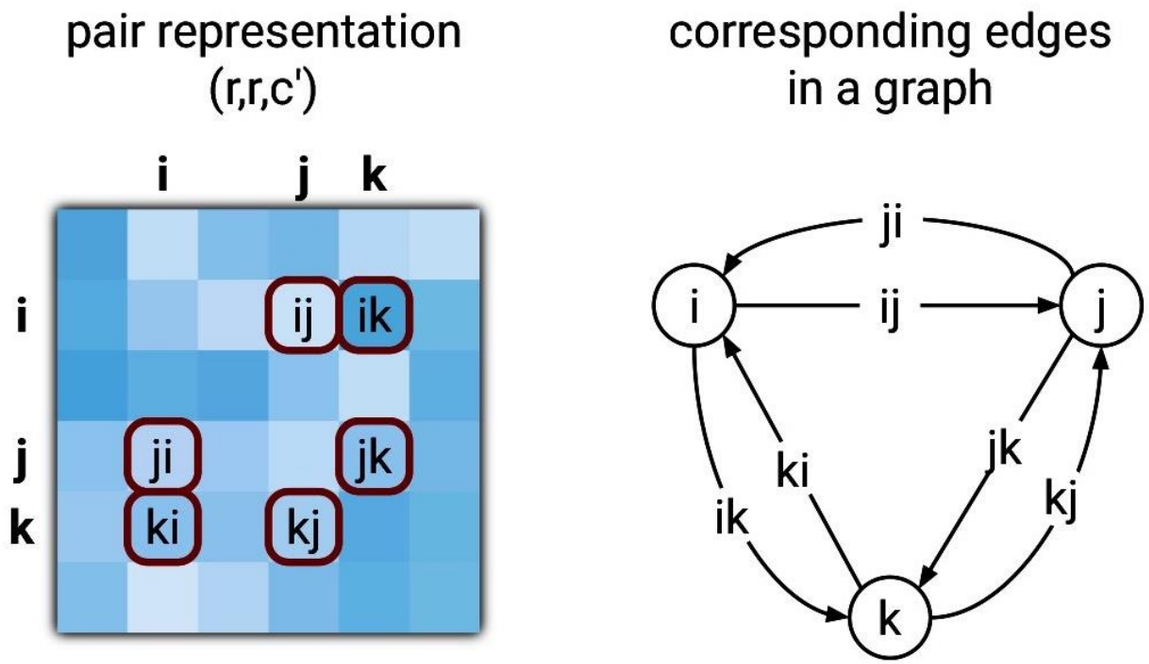
核心算法模块 Evoformer

Evoformer

模型模块 Evoformer



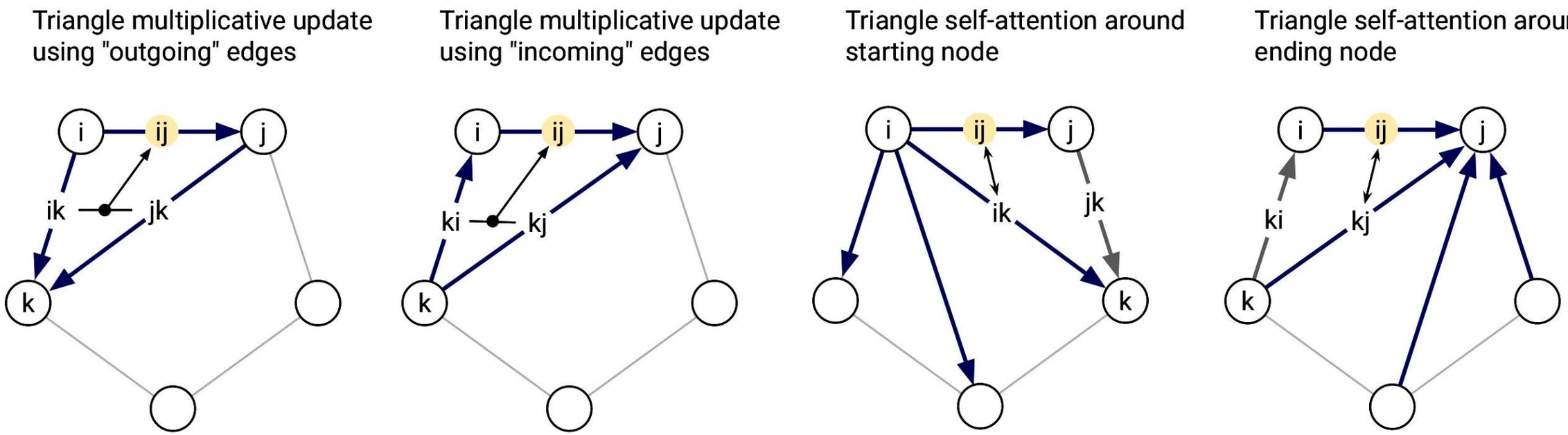
模型模块 Evoformer



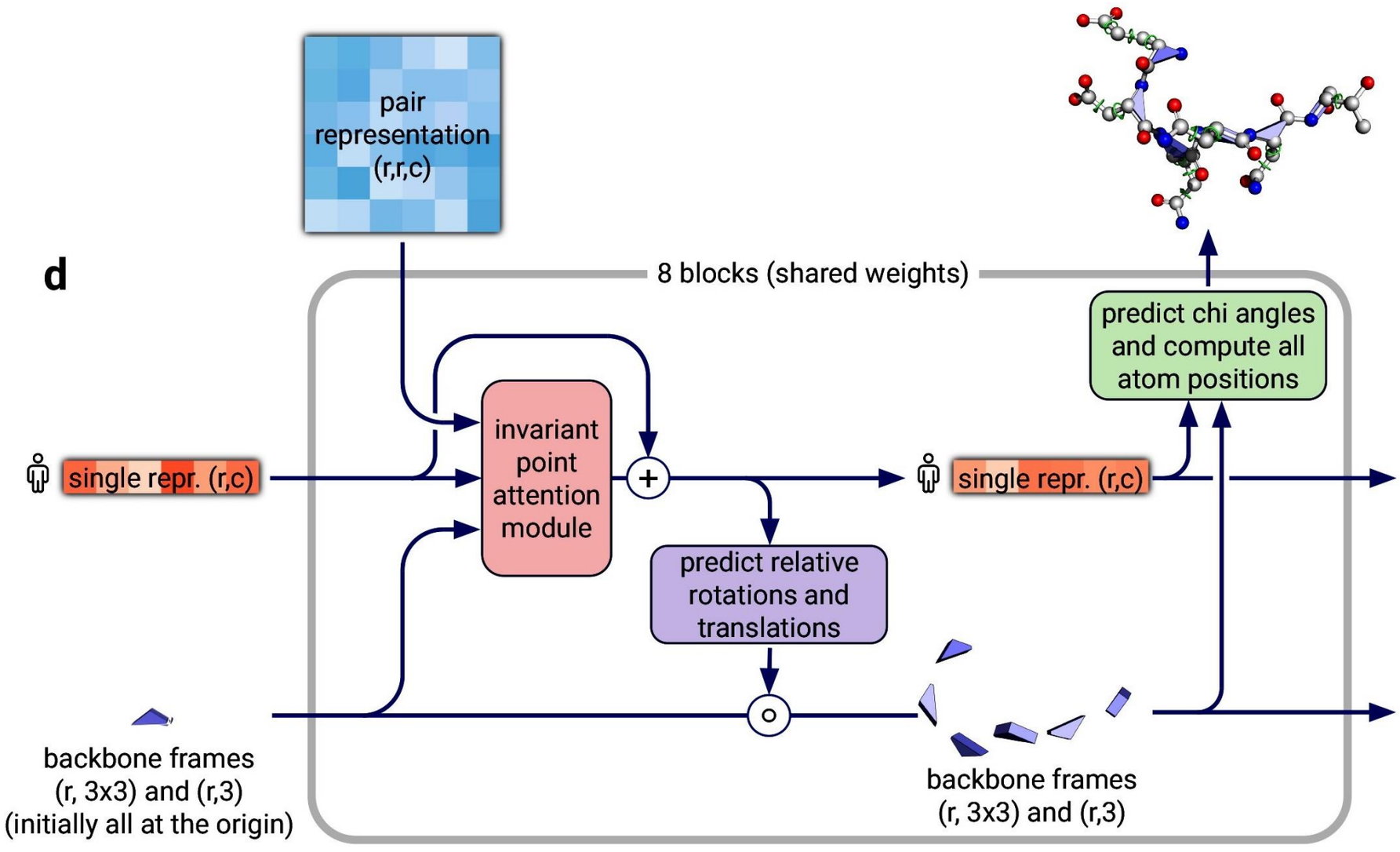
The pair representation interpreted as directed edges in a graph.

模型模块 Evoformer

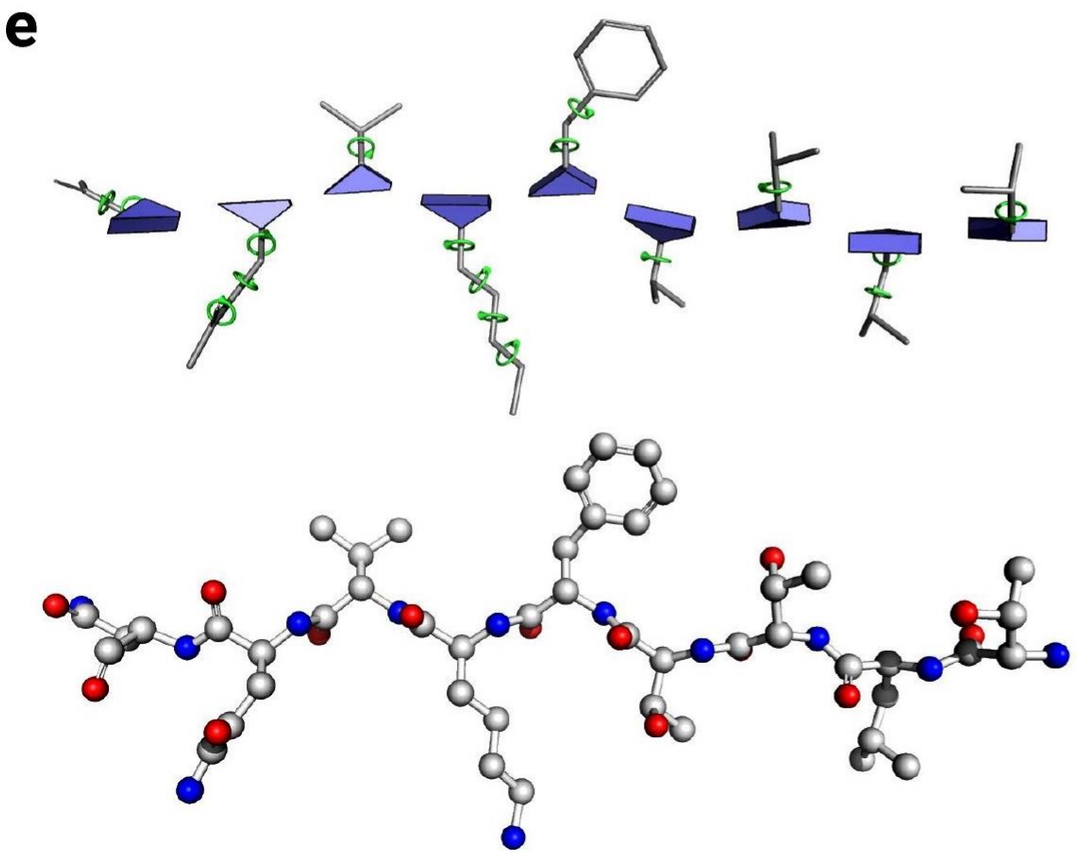
c

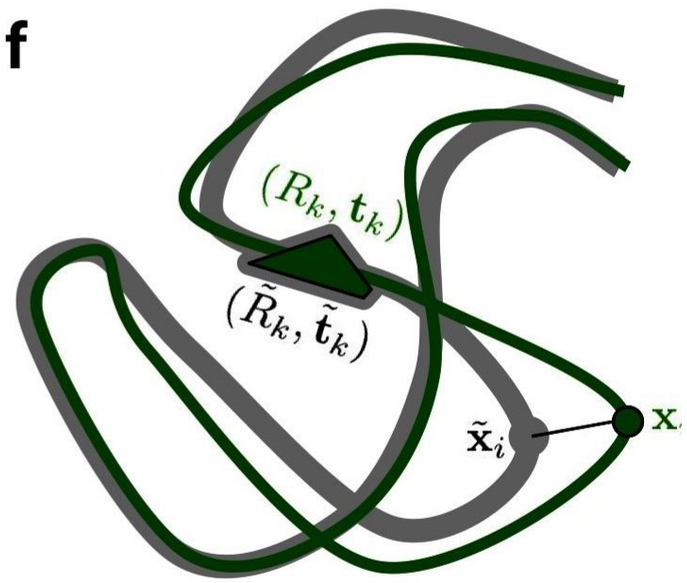


模型模块 Evoformer



模型模块 Evoformer

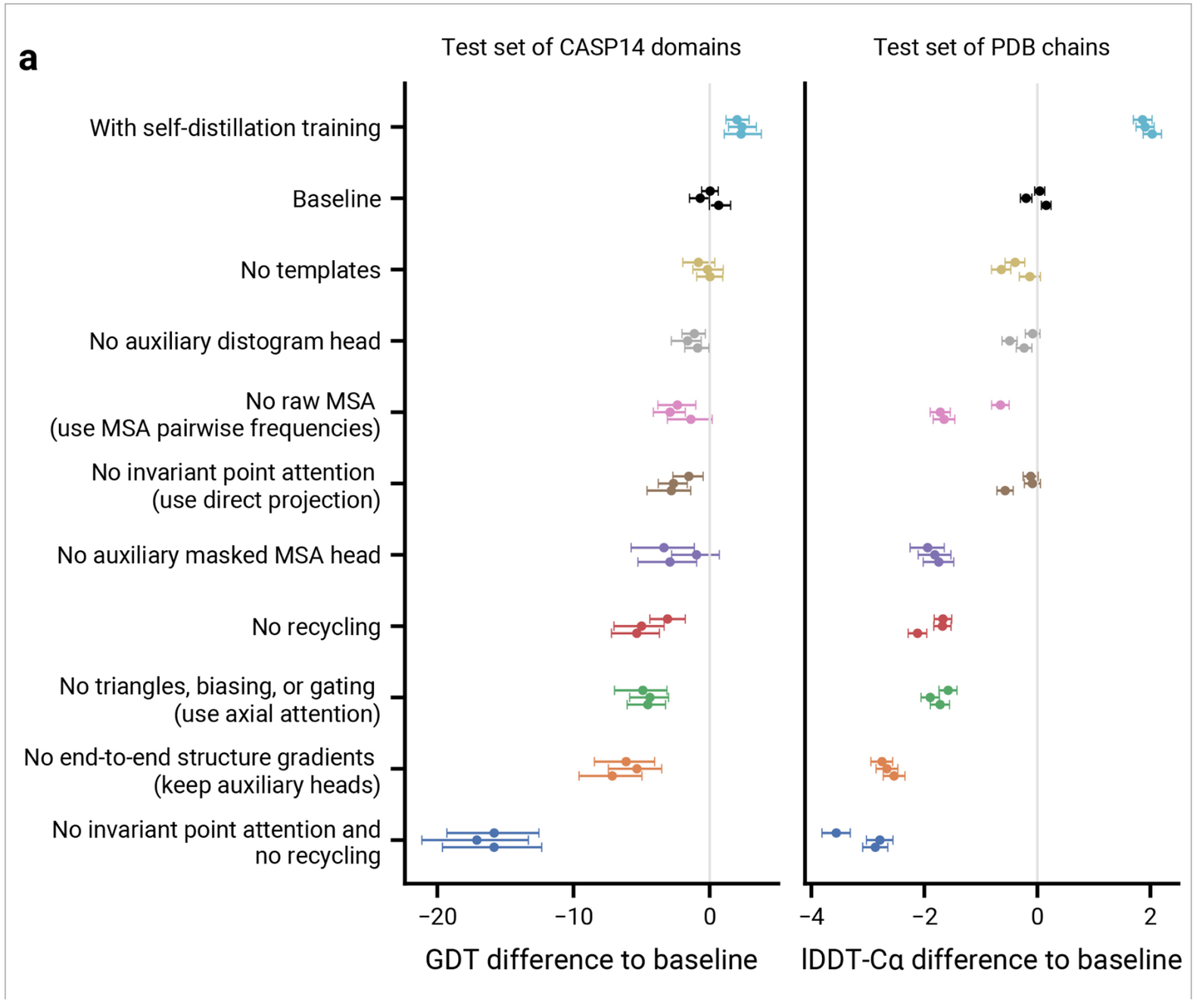




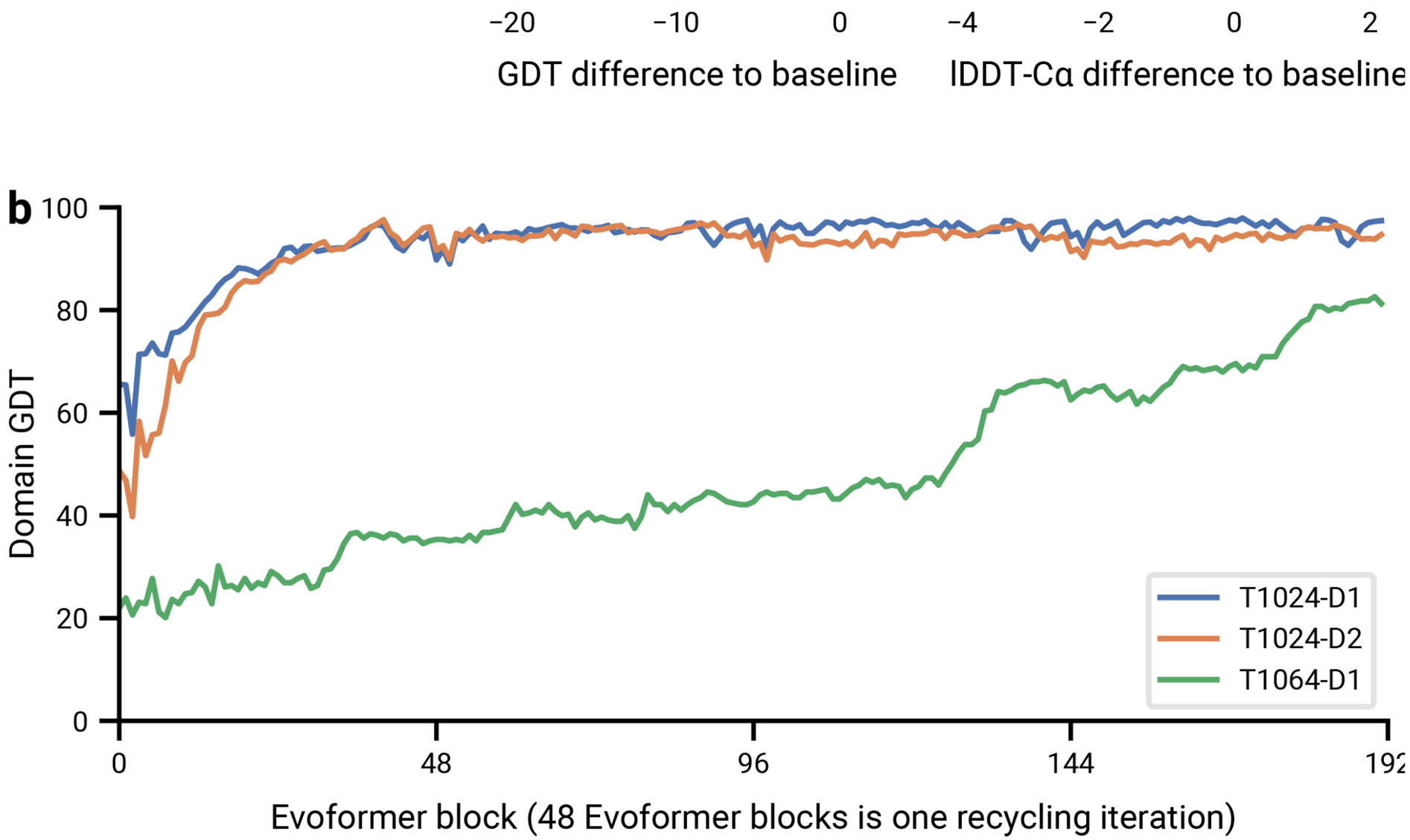
核心算法框架

End-to-end structure prediction

核心算法框架


a
Test set of CASP14 domains
Test set of PDB chains
With self-distillation training
Baseline
No templates
No auxiliary distogram head
No raw MSA
(use MSA pairwise frequencies)
No invariant point attention
(use direct projection)
No auxiliary masked MSA head
No recycling
No triangles, biasing, or gating
(use axial attention)
No end-to-end structure gradients
(keep auxiliary heads)
No invariant point attention and
no recycling
−20
−10
0
−4
−2
0
2
GDT difference to baseline
lDDT-Cα difference to baseline

核心算法框架


−20
−10
0
GDT difference to baseline
−4
−2
0
2
lDDT-Cα difference to baseline
b
100
80
60
40
20
0
Domain GDT
0
48
96
144
192
Evoformer block (48 Evoformer blocks is one recycling iteration)
T1024-D1
T1024-D2
T1064-D1

Training with labelled and Unlabelled data

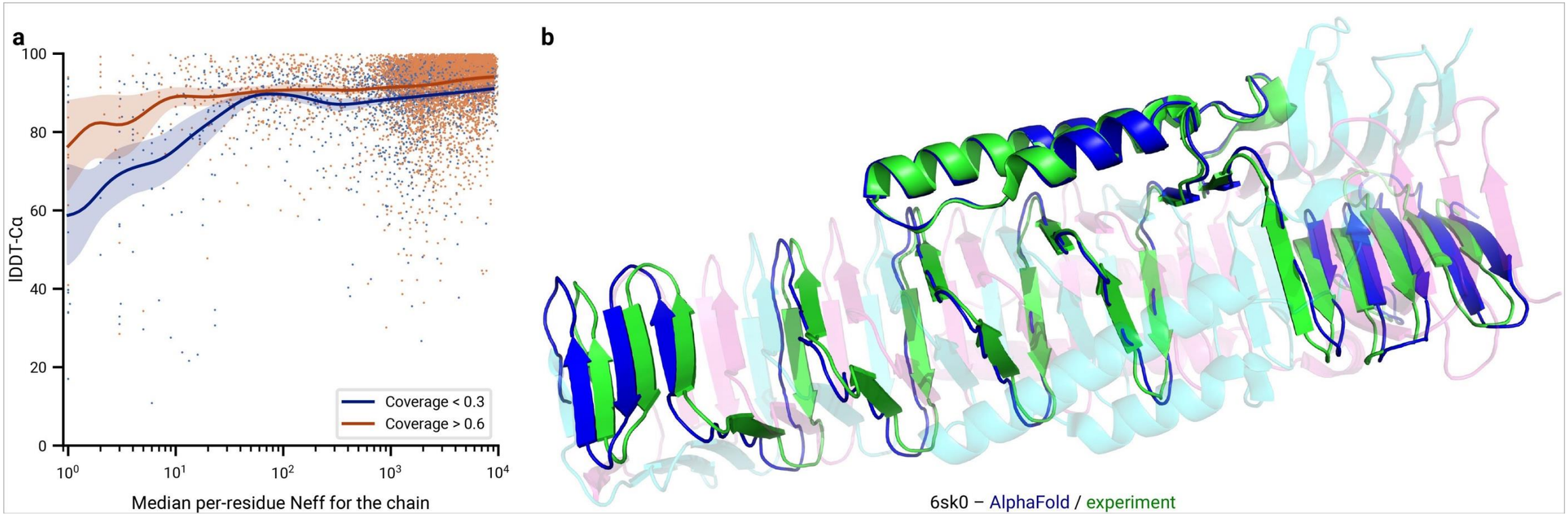
Training with labelled and Unlabelled data

Training with labelled and Unlabelled data

MSA 处理方法

MSA depth and cross-chain contacts

MSA 处理方法


a
100
80
60
40
20
0
IDDT-Cα
Coverage < 0.3
Coverage > 0.6
10^0
10^1
10^2
10^3
10^4
Median per-residue Neff for the chain
b
6sk0 – AlphaFold / experiment

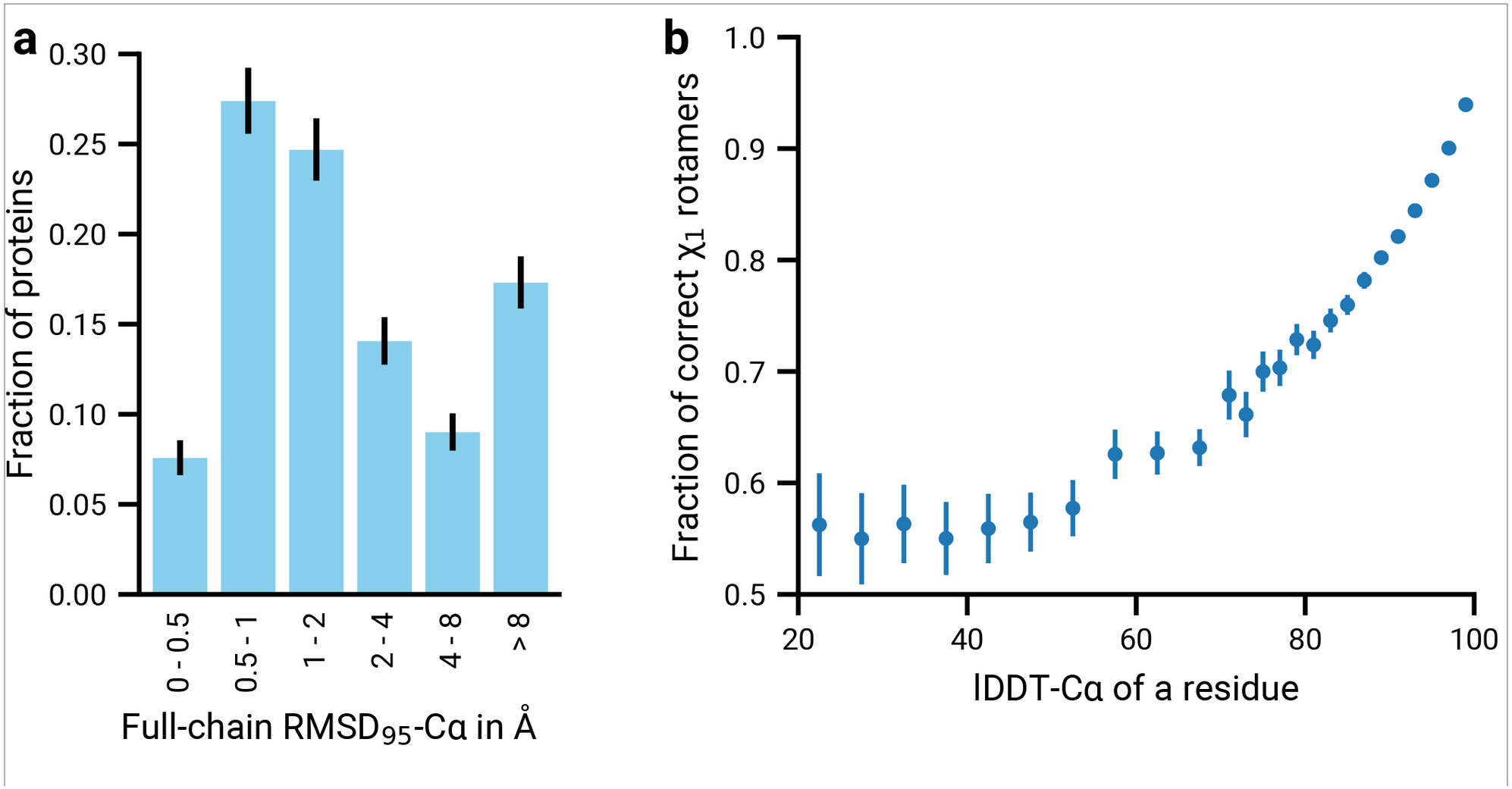
MSA 处理方法

MSA depth and cross-chain contacts

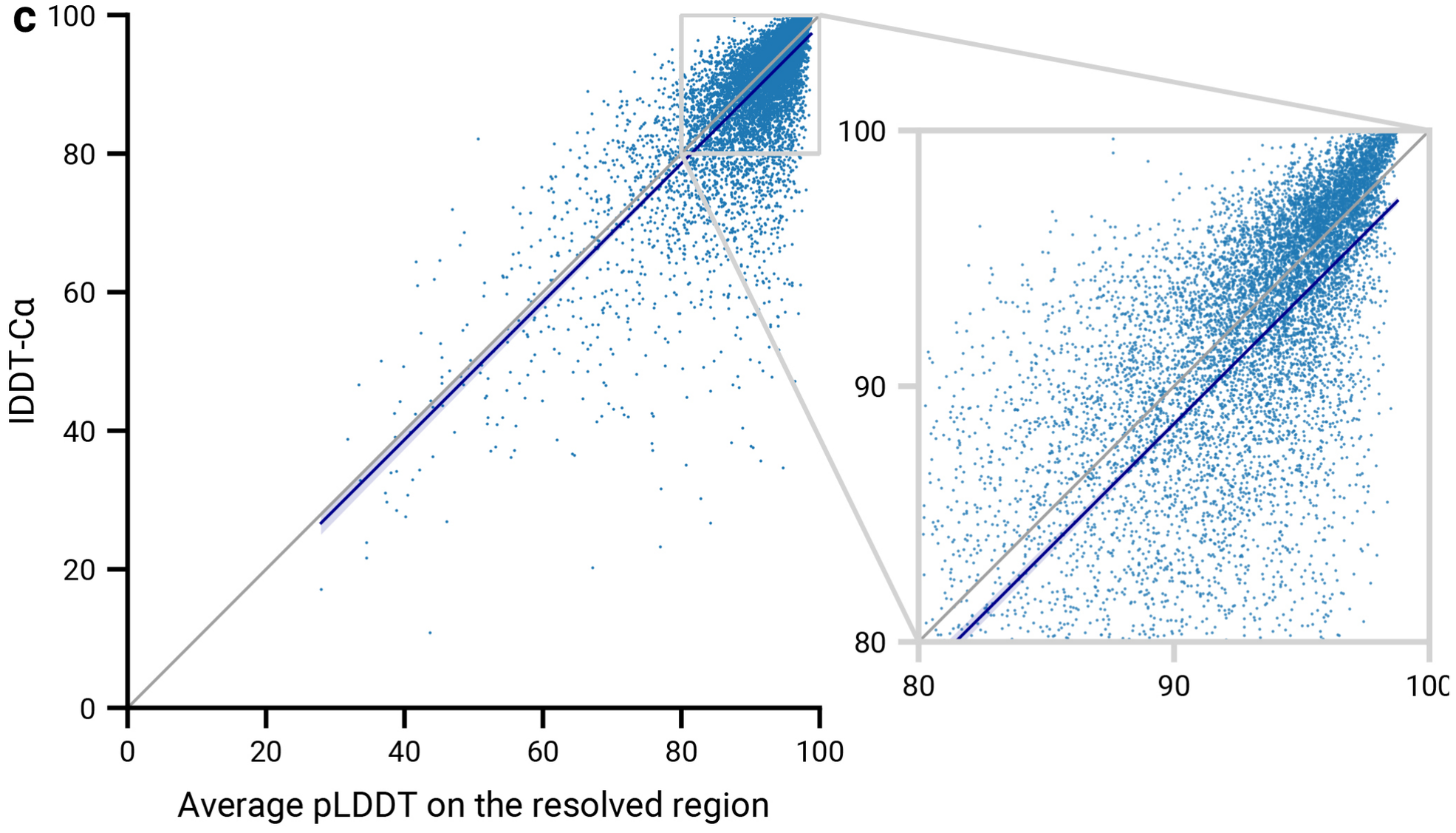
模型效果

Model Result

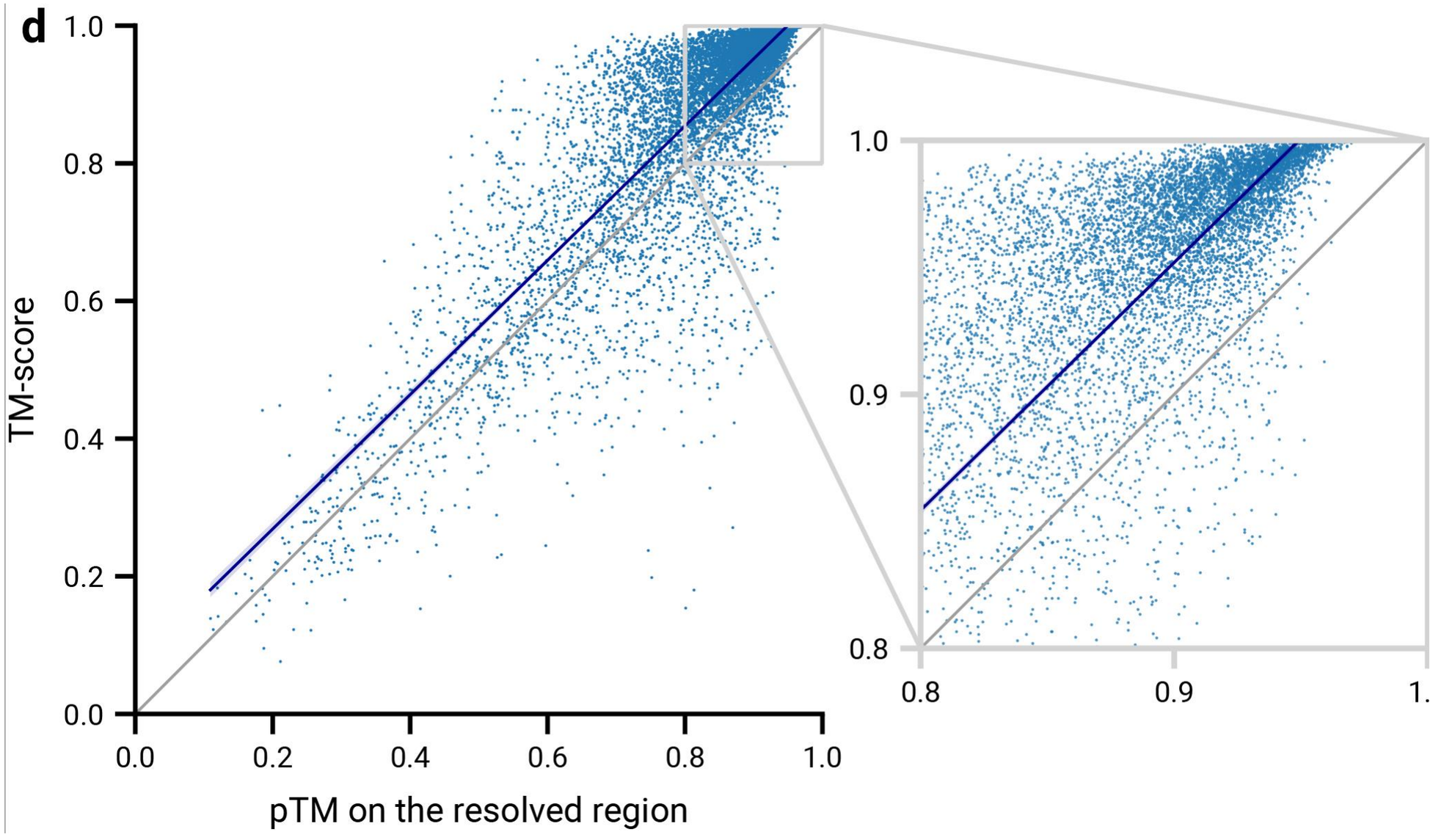
模型效果


a
0.30
0.25
0.20
0.15
0.10
0.05
0.00
Fraction of proteins
0 - 0.5
0.5 - 1
1 - 2
2 - 4
4 - 8
> 8
Full-chain RMSD95-Cα in Å
b
1.0
0.9
0.8
0.7
0.6
0.5
Fraction of correct χ1 rotamers
20
40
60
80
100
lDDT-Cα of a residue

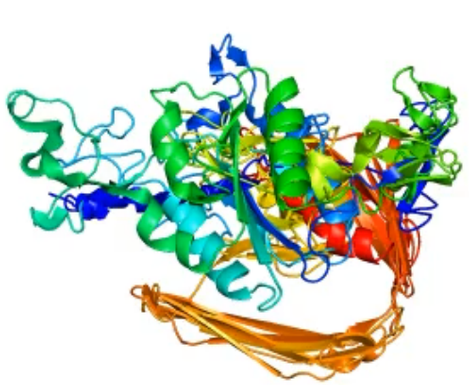
模型效果


c
100
80
60
40
20
0
IDDT-Cα
0
20
40
60
80
100
Average pLDDT on the resolved region
100
90
80
80
90
100

模型效果

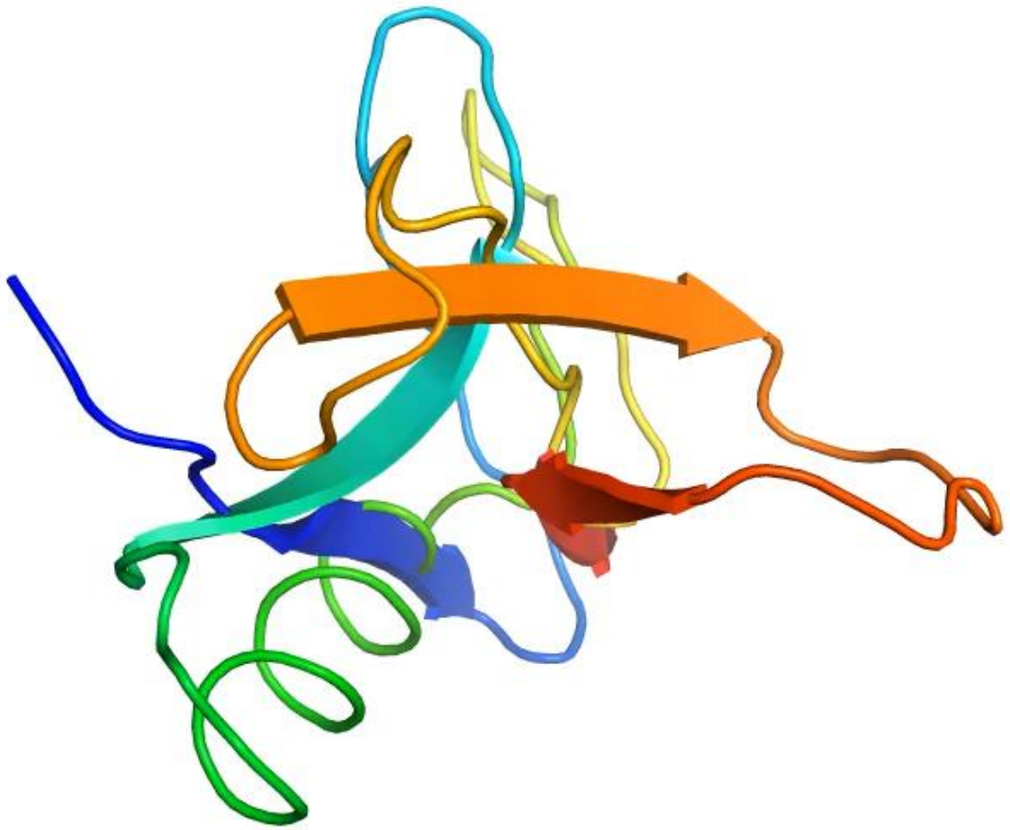


模型效果



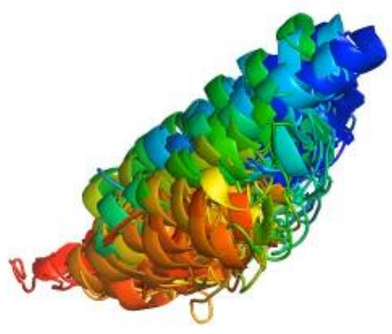
Recycling iteration 0, block 01
Secondary structure assigned from the final prediction

模型效果



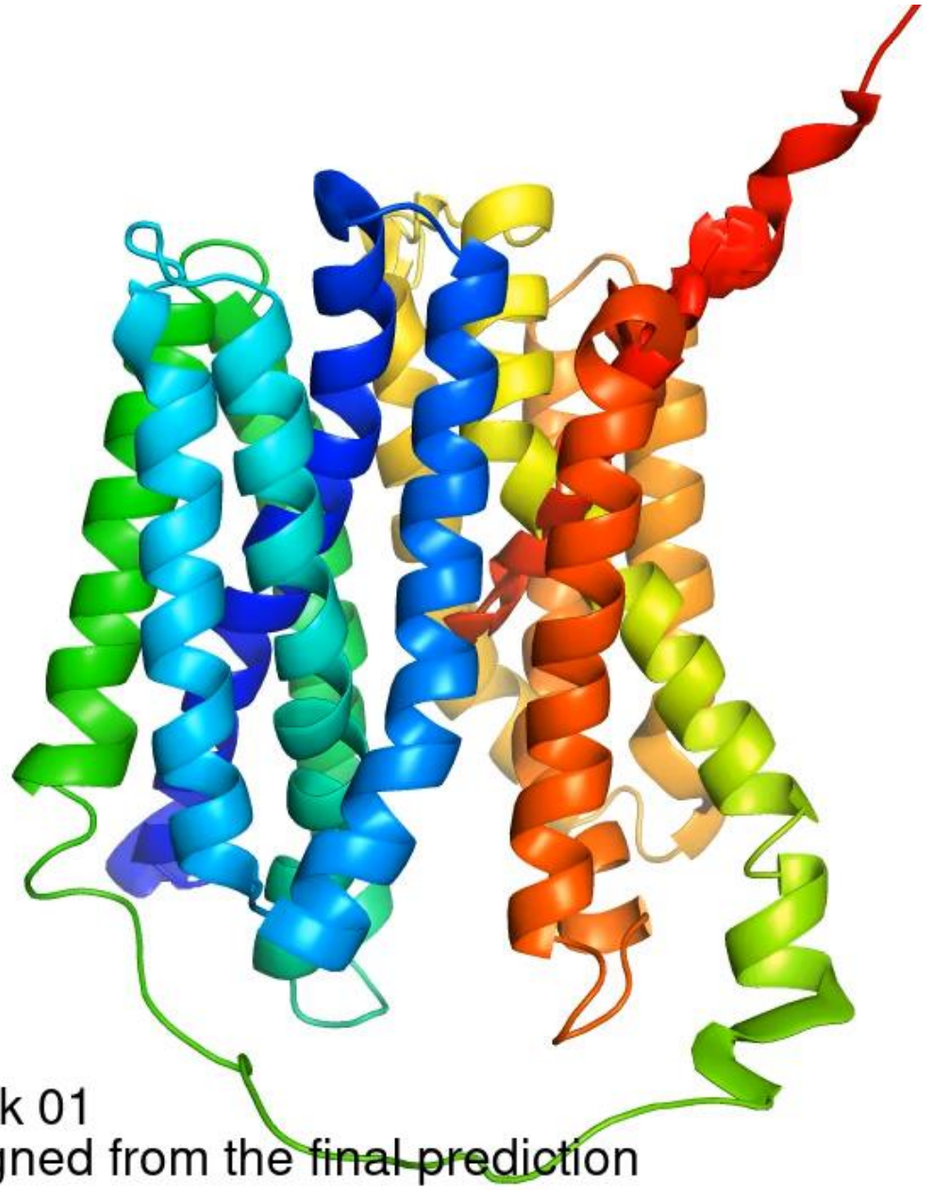
Recycling iteration 0, block 01
Secondary structure assigned from the final prediction

模型效果



Recycling iteration 0, block 01
Secondary structure assigned from the final prediction

模型效果



Recycling iteration 0, block 01
Secondary structure assigned from the final prediction

相关工作

Related work

相关工作

https://github.com/chenxingqiang/ref-Alphafold-Code

讨论

Discusion

讨论

The methodology we have taken in designing AlphaFold is a combination of the bioinformatic and physical approaches: we use a physical and geometric inductive bias to build components that learn from PDB data with minimal imposition of handcrafted features (e.g. AlphaFold builds hydrogen bonds effectively without a hydrogen bond score function). This results in a network that learns far more efficiently from the limited data in the PDB but is able to cope with the complexity and variety of structural data.

讨论

In particular, AlphaFold is able to handle missing physical context and produce accurate models in challenging cases like intertwined homomers or proteins that only fold in the presence of an unknown heme group. The ability to handle underspecified structural conditions is essential to learning from PDB structures as the PDB represents the full range of conditions in which structures have been solved. In general, AlphaFold is trained to produce the protein structure most likely to appear as part of a PDB structure. In cases where a particular stoichiometry or ligand/ion is predictable from the sequence alone, AlphaFold is likely to produce a structure that respects those constraints implicitly

讨论

AlphaFold has already demonstrated its utility to the experimental community, both for molecular replacement , and interpreting cryogenic electron microscopy (cryo-EM) maps. Moreover, because AlphaFold outputs protein coordinates directly, AlphaFold produces predictions in graphics processing unit (GPU)-minutes to GPU-hours depending on the length of the protein sequence (e.g. around one GPU-minute per model for 384 residues, see Methods for details). This opens up the exciting possibility of predicting structures at the proteome-scale and beyond.

讨论

The explosion in available genomic sequencing techniques and data has revolutionized bioinformatics but the intrinsic challenge of experimental structure determination has prevented a similar expansion in our structural knowledge. By developing an accurate protein structure prediction algorithm, coupled with existing large and well-curated structure and sequence databases assembled by the experimental community, we hope to accelerate the advancement of structural bioinformatics that can keep pace with the genomics revolution. We hope that AlphaFold, and computational approaches that apply its techniques for other biophysical problems, will become essential tools of modern biology.

文献/资源列表

https://github.com/chenxingqiang/alphafold2-codecs