BIOINFORMATICS

DATABASES OF PROTEIN SEQUENCES

UniProtKB

- SWISS-PROT: high-quality manual annotation
- TrEMBL: automatic annotation (TrEMBL → SWISS-PROT)
- PIR: USA

DATABASES OF DNA SEQUENCES

- EMBL-Bank : Europe (EMBL-EBI), access from ENA (European Nucleotide Archive)
- GenBank: USA, retrieved by ENTREZ
- DDBJ Japan, retrieved by ARSA, DBGet

STRUCTURE DATABASES

- PDB
- PDBsum: summaries and analyses
- EDS (Uppsala): electron density maps
- EMDataBank: 3D maps from alectron microscopy
- SCOP: fold–superfamily–family
- CATH: class–architecture–topology–homology

PAIRWISE ALIGNMENT

```
D A G T K V S A E Q I L
| | | | |
D A G T K E C H Q I L
```

score=5, gap=0

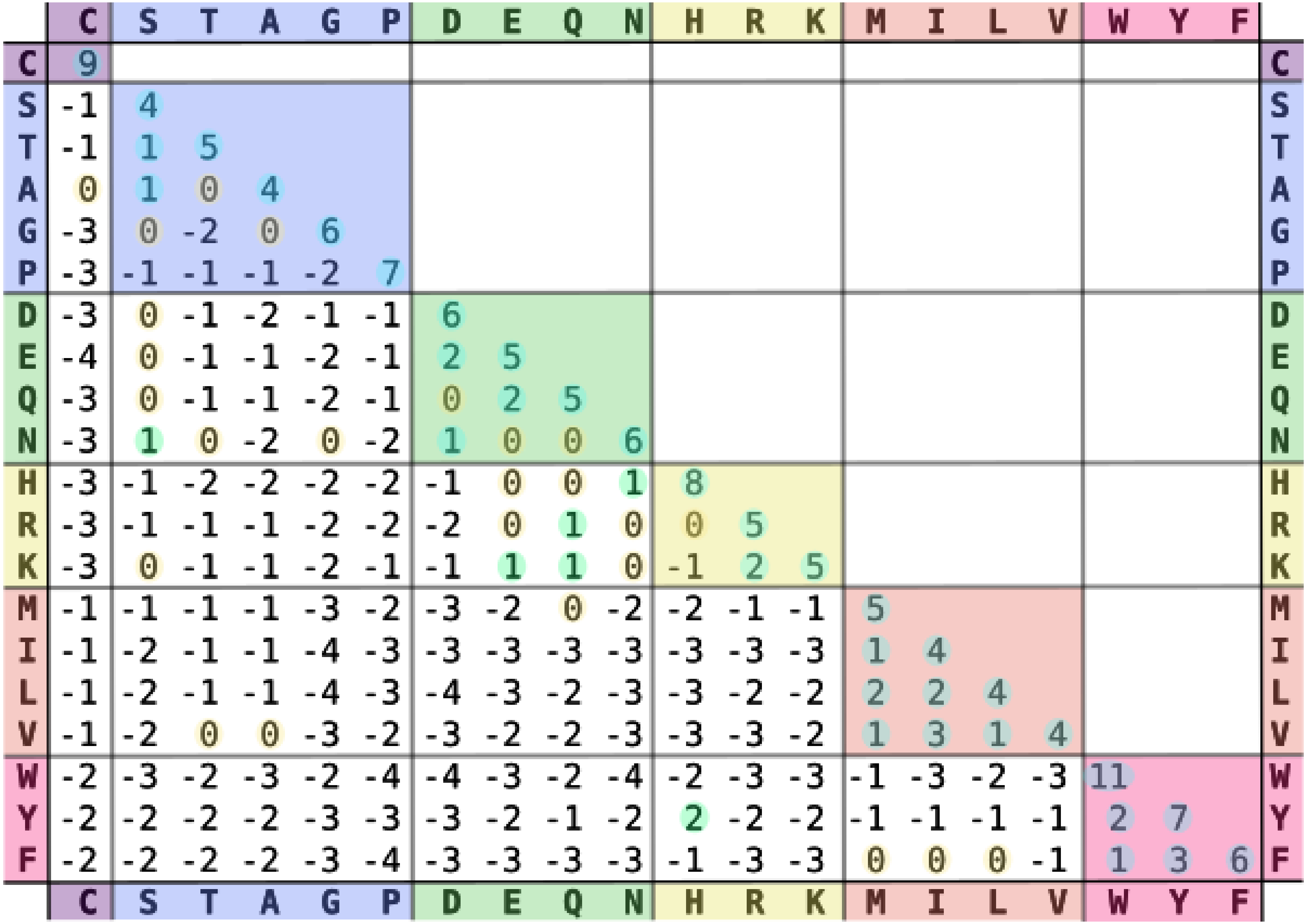
```
D A G T K V S A E Q I L
| | | | |         | | |
D A G T K E C H - Q I L
```

score=8, gap=1

```
D A G T K V S A E - - Q I L
| | | | |       |     | | |
D A G T K - - - E C H Q I L
```

score=9, gap=5

BLOSUM62



	C	S	T	A	G	P	D	E	Q	N	H	R	K	M	I	L	V	W	Y	F	
C	9																				**C**
S	-1	4																			**S**
T	-1	1	5																		**T**
A	0	1	0	4																	**A**
G	-3	0	-2	0	6																**G**
P	-3	-1	-1	-1	-2	7															**P**
D	-3	0	-1	-2	-1	-1	6														**D**
E	-4	0	-1	-1	-2	-1	2	5													**E**
Q	-3	0	-1	-1	-2	-1	0	2	5												**Q**
N	-3	1	0	-2	0	-2	1	0	0	6											**N**
H	-3	-1	-2	-2	-2	-2	-1	0	0	1	8										**H**
R	-3	-1	-1	-1	-2	-2	-2	0	1	0	0	5									**R**
K	-3	0	-1	-1	-2	-1	-1	1	1	0	-1	2	5								**K**
M	-1	-1	-1	-1	-3	-2	-3	-2	0	-2	-2	-1	-1	5							**M**
I	-1	-2	-1	-1	-4	-3	-3	-3	-3	-3	-3	-3	-3	1	4						**I**
L	-1	-2	-1	-1	-4	-3	-4	-3	-2	-3	-3	-2	-2	2	2	4					**L**
V	-1	-2	0	0	-3	-2	-3	-2	-2	-3	-3	-3	-2	1	3	1	4				**V**
W	-2	-3	-2	-3	-2	-4	-4	-3	-2	-4	-2	-3	-3	-1	-3	-2	-3	11			**W**
Y	-2	-2	-2	-2	-3	-3	-3	-2	-1	-2	2	-2	-2	-1	-1	-1	-1	2	7		**Y**
F	-2	-2	-2	-2	-3	-4	-3	-3	-3	-3	-1	-3	-3	0	0	0	-1	1	3	6	**F**
	C	**S**	**T**	**A**	**G**	**P**	**D**	**E**	**Q**	**N**	**H**	**R**	**K**	**M**	**I**	**L**	**V**	**W**	**Y**	**F**	

PAIRWISE DATABASE SEARCH

Fast local similarity algorithms

- FastA
- BLAST

MULTIPLE ALIGNMENT

Progressive algorithms

- CLUSTAL: evolutionary tree + pairwise alignment
- PSI-BLAST: hybrid (pairwise + multiple), iterative, sensitive

Databases: Pfam, PRINTS

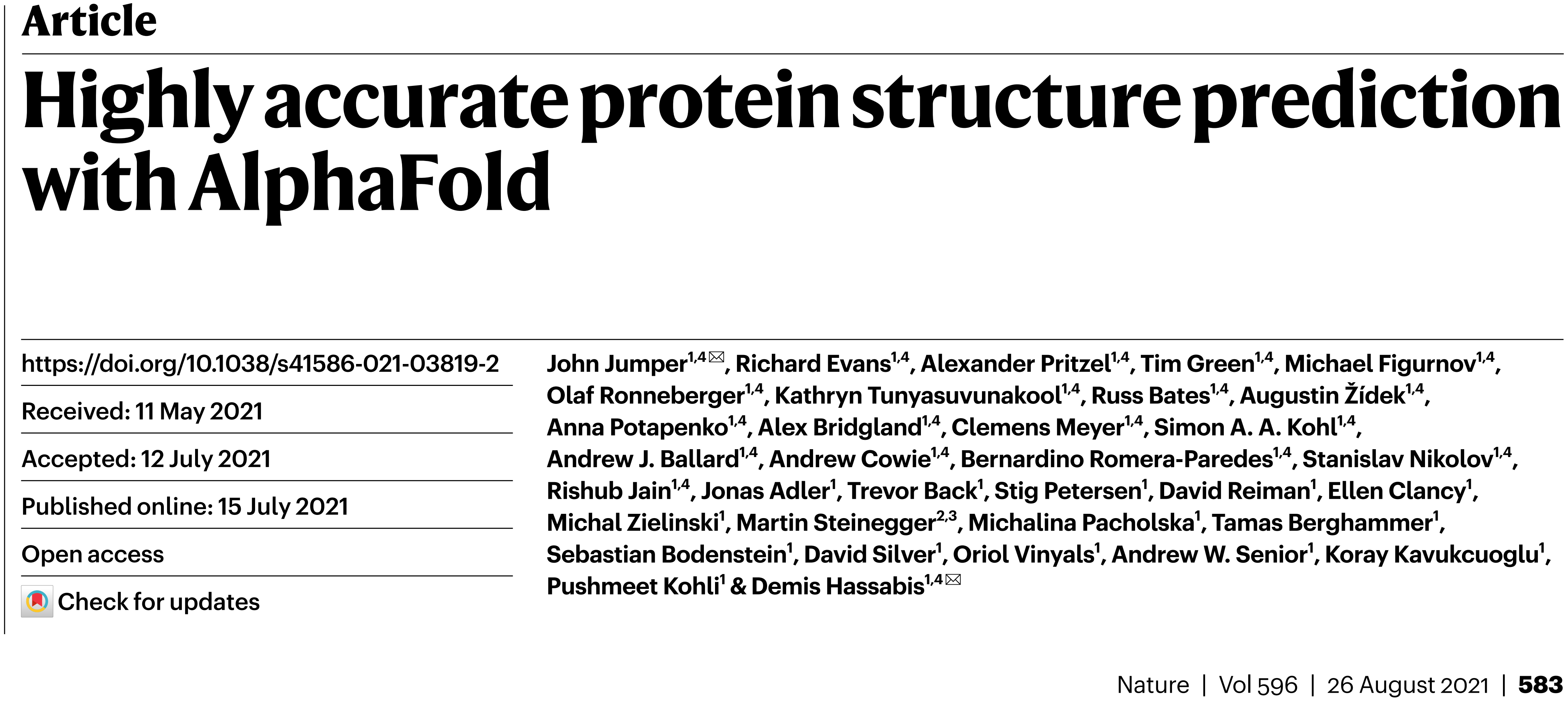
STRUCTURE PREDICTION

- Secondary structure: PSI-PRED
- Fold: threading
- Tertiary structure from homologous structure: homology modelling
- Tertiary structure from multiple sequence alignment: AlphaFold.2

test sequence:

`PIAQIHILEGRSDEQKETLIREVSEAISRSLDAPLTSVRVIITEMAKGHFGIGGELASK`

ALPHAFOLD2



Article

Highly accurate protein structure prediction with AlphaFold

https://doi.org/10.1038/s41586-021-03819-2

Received: 11 May 2021

Accepted: 12 July 2021

Published online: 15 July 2021

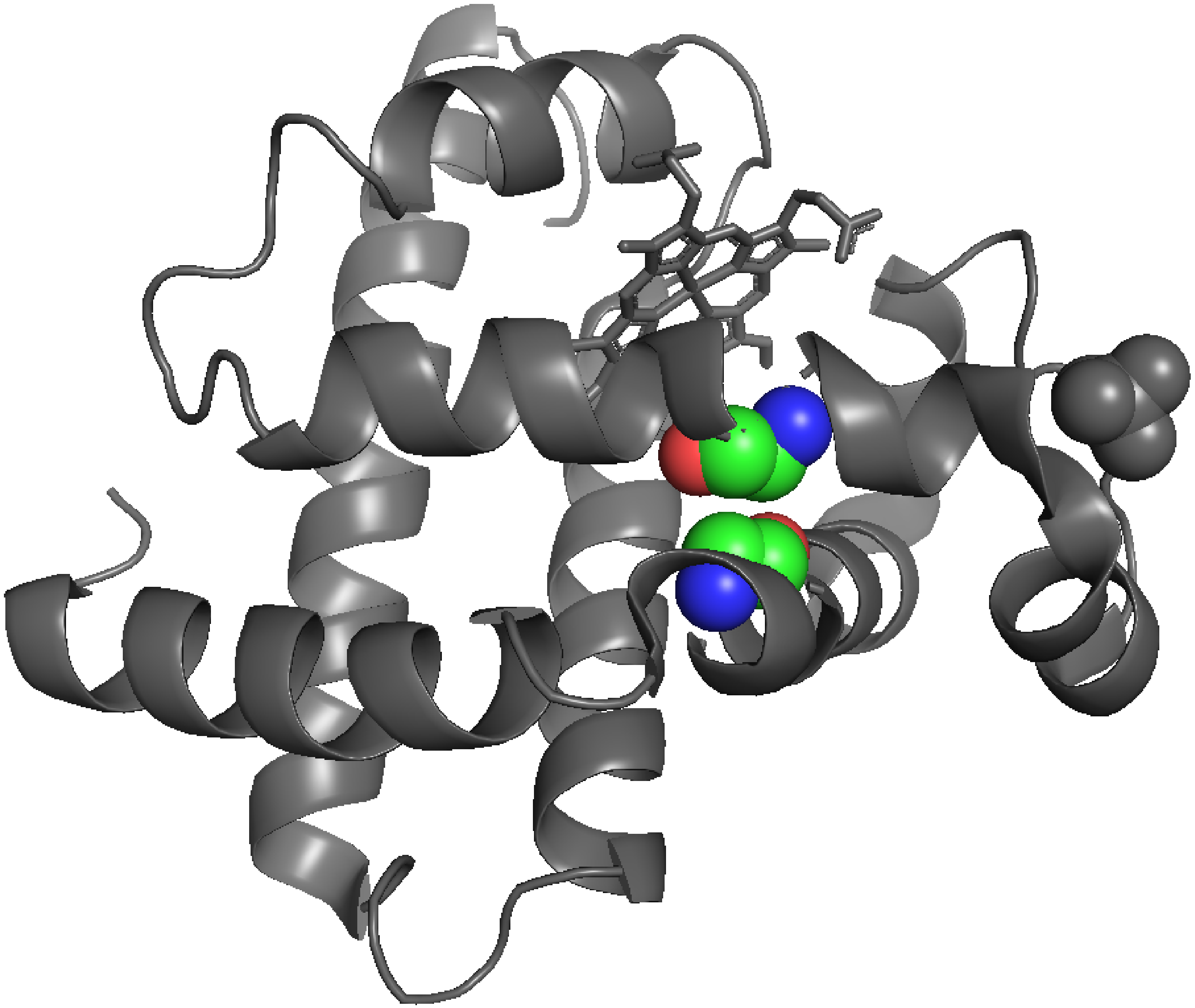
Open access

Check for updates

John Jumper[1,4]✉, Richard Evans[1,4], Alexander Pritzel[1,4], Tim Green[1,4], Michael Figurnov[1,4], Olaf Ronneberger[1,4], Kathryn Tunyasuvunakool[1,4], Russ Bates[1,4], Augustin Žídek[1,4], Anna Potapenko[1,4], Alex Bridgland[1,4], Clemens Meyer[1,4], Simon A. A. Kohl[1,4], Andrew J. Ballard[1,4], Andrew Cowie[1,4], Bernardino Romera-Paredes[1,4], Stanislav Nikolov[1,4], Rishub Jain[1,4], Jonas Adler[1], Trevor Back[1], Stig Petersen[1], David Reiman[1], Ellen Clancy[1], Michal Zielinski[1], Martin Steinegger[2,3], Michalina Pacholska[1], Tamas Berghammer[1], Sebastian Bodenstein[1], David Silver[1], Oriol Vinyals[1], Andrew W. Senior[1], Koray Kavukcuoglu[1], Pushmeet Kohli[1] & Demis Hassabis[1,4]✉

		24	64
Physeter macrocephalus	sperm whale	HGQ	HGVTV
Balaena mysticetus	bowhead	HGQ	HGNTV
Sus scrofa	pig	HGQ	HGNTI
Orycteropus afer afer	aardvark	HGQ	HGTTV
Equus caballus	horse	HGQ	HGTVV
Homo sapiens	man	HGQ	HGATV

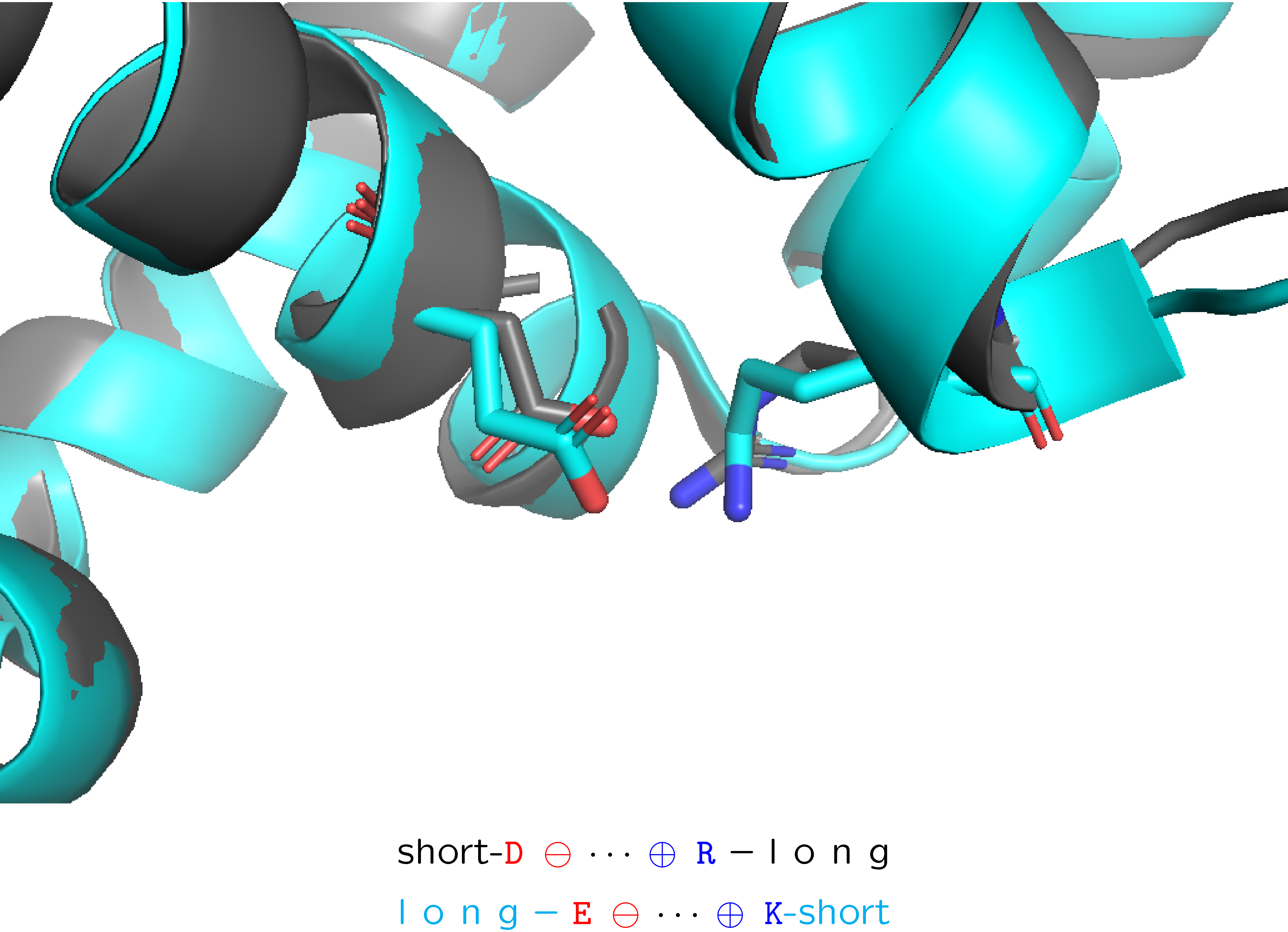
G small

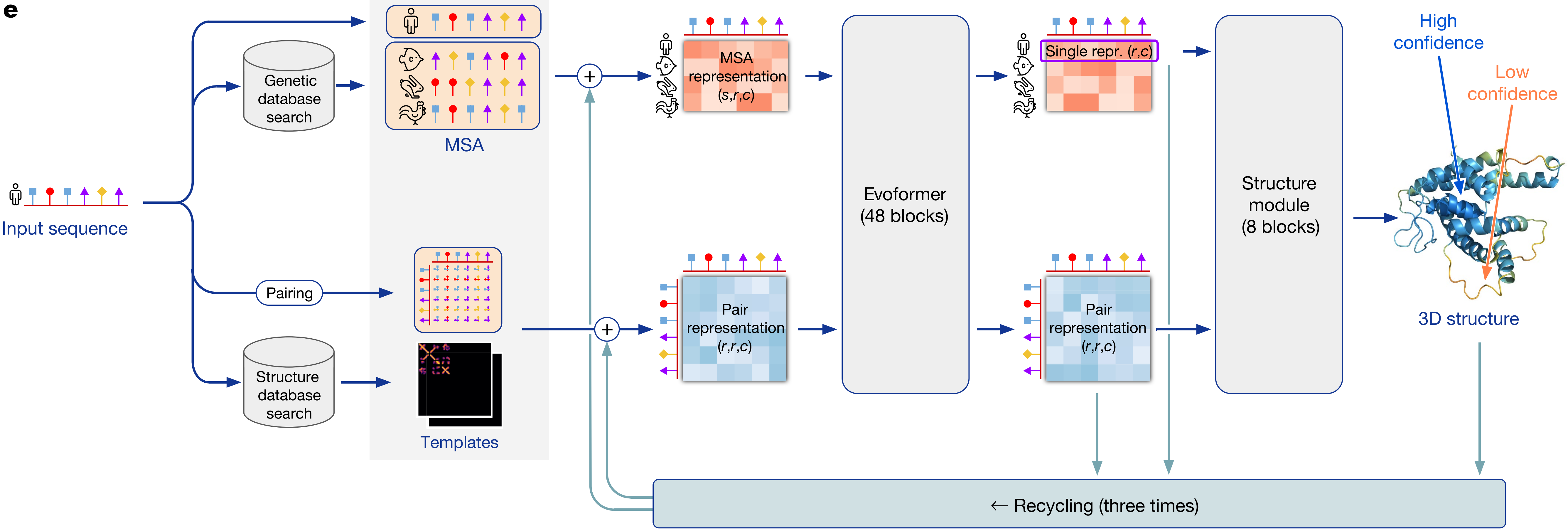


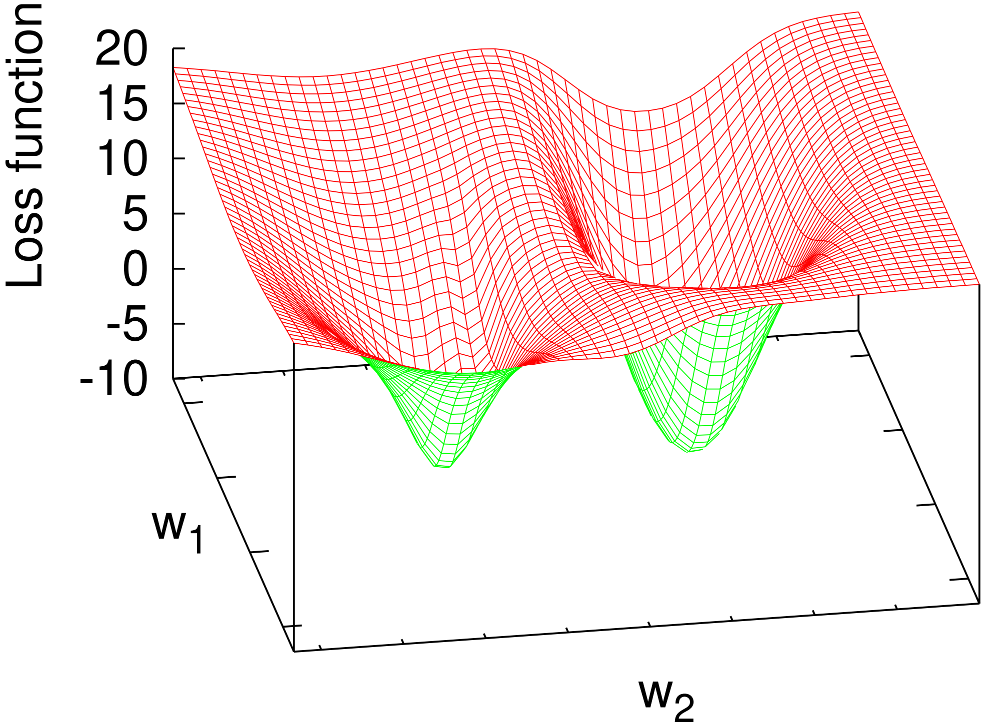
		26	116
Physeter macrocephalus	sperm whale	QDH	HSRH
Balaena mysticetus	bowhead	QDH	HSRH
Sus scrofa	pig	QEH	QSKH
Orycteropus afer afer	aardvark	QEH	QSKH
Equus caballus	horse	QEH	HSKH
Homo sapiens	man	QEH	QSKH

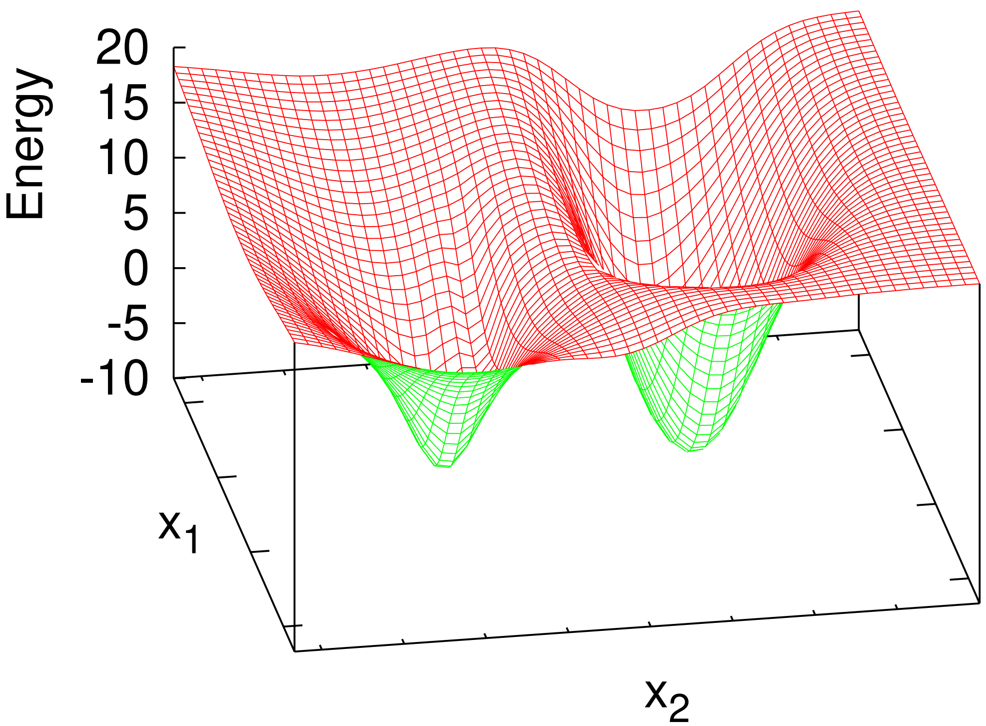
short-D $\ominus \cdots \oplus$ R $-$ long

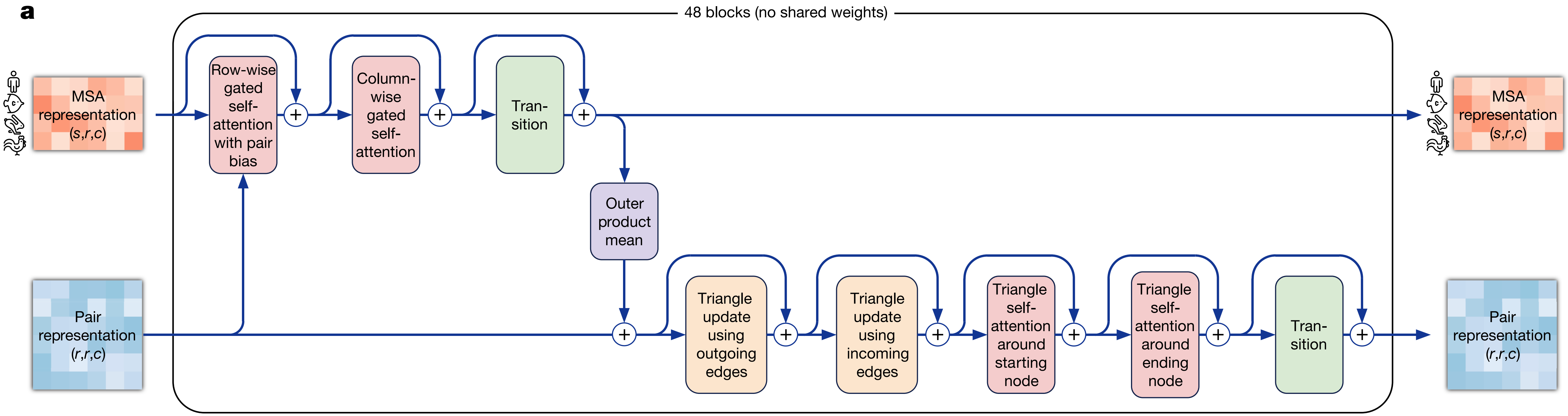
long $-$ E $\ominus \cdots \oplus$ K-short

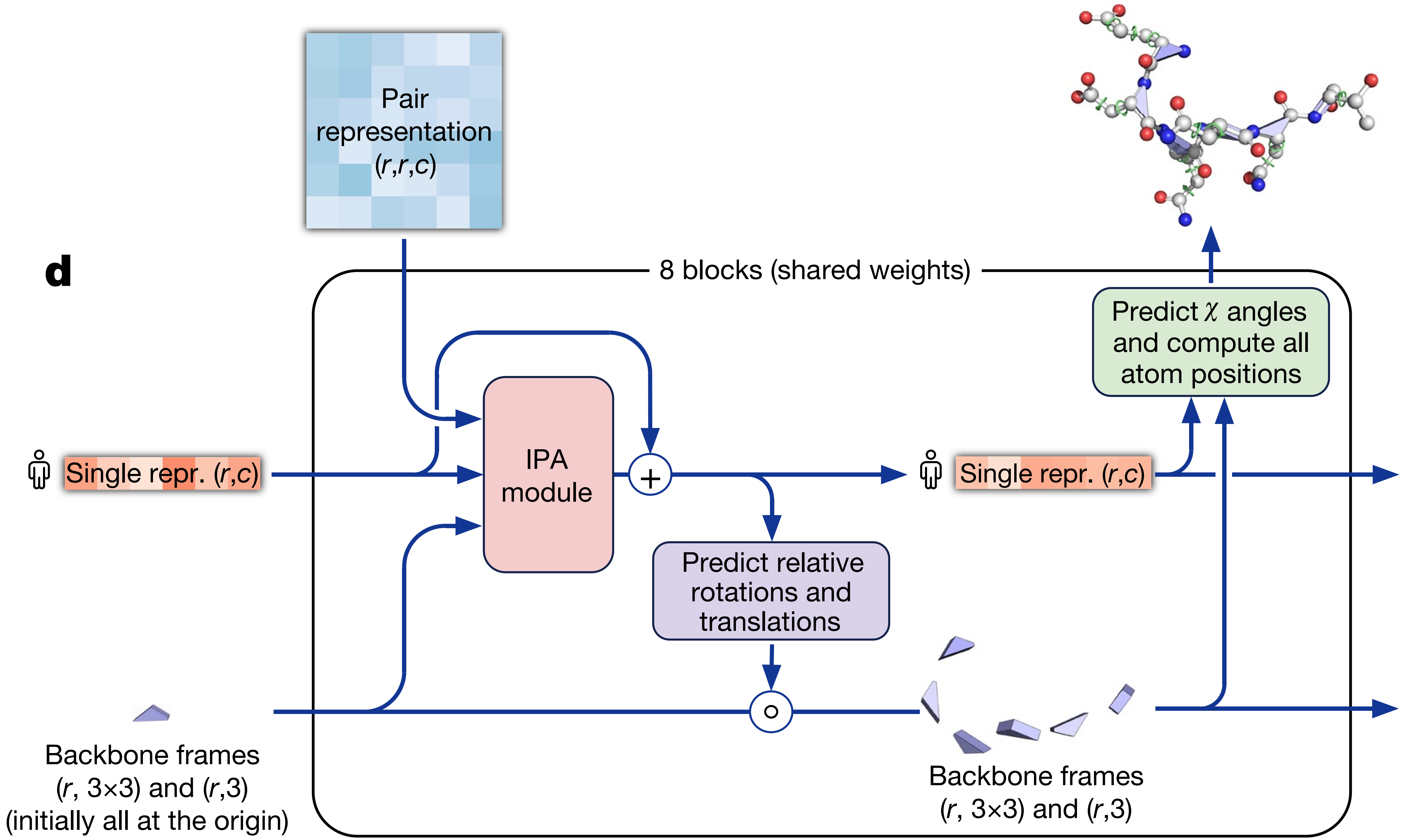

short-D ⊖ ··· ⊕ R – long
long – E ⊖ ··· ⊕ K-short


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)


Loss function
20
15
10
5
0
-5
-10
w1
w2


Energy
20
15
10
5
0
-5
-10
x1
x2


a
48 blocks (no shared weights)
MSA representation (s,r,c)
Pair representation (r,r,c)
Row-wise gated self-attention with pair bias
Column-wise gated self-attention
Tran-sition
Outer product mean
Triangle update using outgoing edges
Triangle update using incoming edges
Triangle self-attention around starting node
Triangle self-attention around ending node
Tran-sition
MSA representation (s,r,c)
Pair representation (r,r,c)


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)