2024-06-17

citas TAPE y PEER CASP

Unified Benchmarks and Evaluation Protocols

Initiatives like TAPE [145], ProteinGym [303], ProteinShake [315], PEER [316], ProteinInvBench [317], ProteinWorkshop [318], exemplify the critical role of comprehensive benchmarking in furthering innovation. Moreover, the Critical Assessment of Protein Structure Prediction (CASP) experiments act as a crucial platform for evaluating the most recent advancements in PSP [Survey 2024]

Revisar TAPE y PEER CASP

TAPE

Task Assesing Preotein embbedings is a set of 5 biologillicaly relevant Semisupervised learning Tasks spread across different domains of protein biology:

First attempt at systematically evaluating semisupervised learning on protein sequences.

Los modelos se preentrenaron utilizando un corpus no etiquetado de secuencias de proteínas del conjunto de datos Pfam.

Para cada tarea se utilizó una arquitectura (drawing from state-of-the-art where available) y el fine-tunning en todos los modelos fue el mismo.

1. Protein Structure Prediction
2. Evolutionary Understanding
3. Protein Engineering

Table 2: Results on downstream supervised tasks

Method		Structure		Evolutionary	Engineering	
		SS	Contact	Homology	Fluorescence	Stability
No Pretrain	Transformer	0.70	0.32	0.09	0.22	-0.06
	LSTM	0.71	0.19	0.12	0.21	0.28
	ResNet	0.70	0.20	0.10	-0.28	0.61
Pretrain	Transformer	0.73	0.36	0.21	**0.68**	**0.73**
	LSTM	0.75	0.39	**0.26**	0.67	0.69
	ResNet	0.75	0.29	0.17	0.21	**0.73**
Supervised [11]	LSTM	0.73	0.40	0.17	0.33	0.64
UniRep [12]	mLSTM	0.73	0.34	0.23	0.67	**0.73**
Baseline	One-hot	0.69	0.29	0.09	0.14	0.19
	Alignment	**0.80**	**0.64**	0.09	N/A	N/A

PEER

Peer benchmark includes 14 biologically relevant tasks (Now 17, https://torchprotein.ai/benchmark) that covers diverse aspects of protein understanding, including:

1. Protein Function Prediction
2. Protein Localization Prediction
3. Protrein Structure Prediction
4. Protein-Protein interaction Prediction
5. Protein-Ligand Prediction

Para cada tarea se evaluó el performance de diferentes tipos de:

sequence-based approaches

sequence encoders: CNNs LSTMs Transformers pLLM

Se evaluaron diferentes approaches under the multitask learning setting

En cada benchmark task cada data set se diseño adecuadamente para la tarea en cuestión

Also evaluate different approaches under multi-task learning setting

Task (Acronym)	Task Category	Data Source	#Protein	Seq. len.	#Train/Validation/Test	Metric
Function Prediction						
GB1 fitness prediction (GB1)	Protein-wise Reg.	FLIP [16]	8,733	$378.6_{(0.9)}$	381/43/8,309	Spearman's ρ
AAV fitness prediction (AAV)	Protein-wise Reg.	FLIP [16]	82,583	$1033.0_{(3.4)}$	28,626/3,181/50,776	Spearman's ρ
Thermostability prediction (Thermo)	Protein-wise Reg.	FLIP [16]	7,158	$880.6_{(974.2)}$	5,149/643/1,366	Spearman's ρ
Fluorescence prediction (Flu)	Protein-wise Reg.	Sarkisyan's dataset [71]	54,025	$343.3_{(1.3)}$	21,446/5,362/27,217	Spearman's ρ
Stability prediction (Sta)	Protein-wise Reg.	Rocklin's dataset [66]	68,934	$66.6_{(5.2)}$	53,571/2,512/12,851	Spearman's ρ
β-lactamase activity prediction (β-lac)	Protein-wise Reg.	Envision [25]	5,198	$396.1_{(0.7)}$	4,158/520/520	Spearman's ρ
Solubility prediction (Sol)	Protein-wise Cls.	DeepSol [39]	71,419	$424.1_{(225.9)}$	62,478/6,942/1,999	Acc
Localization Prediction						
Subcellular localization prediction (Sub)	Protein-wise Cls.	DeepLoc [2]	13,961	$665.3_{(395.3)}$	8,945/2,248/2,768	Acc
Binary localization prediction (Bin)	Protein-wise Cls.	DeepLoc [2]	8,634	$636.5_{(396.5)}$	5,161/1,727/1,746	Acc
Structure Prediction						
Contact prediction (Cont)	Residue-pair Cls.	ProteinNet [3]	25,563	$320.0_{(275.2)}$	25,299/224/40	L/5 precision
Fold classification (Fold)	Protein-wise Cls.	DeepSF [31]	13,766	$235.4_{(155.1)}$	12,312/736/718	Acc
Secondary structure prediction (SSP)	Residue-wise Cls.	NetSurfP-2.0 [41]	11,361	$360.5_{(229.3)}$	8,678/2,170/513	Acc
Protein-Protein Interaction Prediction						
Yeast PPI prediction (Yst)	Protein-pair Cls.	Guo's dataset [26]	1,707	$726.3_{(432.0)}$	1,668/131/373	Acc
Human PPI prediction (Hum)	Protein-pair Cls.	Pan's dataset [59]	5,553	$727.7_{(438.2)}$	6,844/277/227	Acc
PPI affinity prediction (Aff)	Protein-pair Reg.	SKEMPI [56]	627	$304.9_{(193.8)}$	2,127/212/343	RMSE
Protein-Ligand Interaction Prediction						
Affinity prediction on PDBbind (PDB)	Protein-ligand Reg.	PDBbind [49]	10,607	$414.9_{(234.3)}$	16,436/937/285	RMSE
Affinity prediction on BindingDB (BDB)	Protein-ligand Reg.	BindingDB [47]	1,006	$799.8_{(417.0)}$	7,900/878/5,230	RMSE

Resultados importantes: Los modelos preentrenados como ProtBert y ESM-1b tuvieron el mejor rendimiento en la mayoría de las tareas individuales, y Entrenar múltiples tareas conjuntamente mejoró el rendimiento de los modelos

CASP

The most well-known protein Benchmark. Focuses in structure modeling

Se centra en Protein Structure Prediction y nada más

Se realiza cada dos años, el más actual es CASP XV y Comenzó en 1994

El interés central de CASP XV fue qué grupos además de DeepMind podían aproxiae el accuracy experimenta y si los limitaciones de shallow sequence alignment pudieron superarse.

Se mencionan CAMEO y CAPRI

Alpha Fold 2: supera a todos pero se combina con otros métodos para alcanzar buenos resultados. Usando los parámetros estándar del modelo, sólo produce los mejores resultados para 2/3 de los targets

El segundo mejor performance es rosettafold

Three dimensional protein structure

- domain domain interaction

En general, los modelos mostraron un rendimiento inferior en varios casos específicos, incluyendo **proteínas con baja homología**, **complejos de proteínas**, **proteínas con alta movilidad estructural y proteínas de membrana.** →Hay que mejorar los métodos de predicción para abordar la complejidad y variabilidad de las estructuras proteicas en diferentes contextos.

- **Free Modeling**

Predicción de estructuras de proteínas sin usar estructuras homólogas como plantillas, dependiendo completamente de la secuencia de aminoácidos.

- **Template-Based Modeling**

utilizando plantillas homólogas conocidas. Se basa en la similitud de secuencias con proteínas de estructura conocida para modelar la nueva proteína.

- **Refinement**

Mejora de la precisión de modelos iniciales de proteínas, ajustando las conformaciones de las proteínas para acercarlas más a su estructura nativa.

- **Contact Prediction**

Predicción de qué pares de aminoácidos en una proteína están en contacto cercano (normalmente definidos como a menos de 8 Å de distancia en el espacio tridimensional).

- **Membrane Proteins**

Predicción de la estructura de proteínas de membrana, que se encuentran incrustadas en las membranas celulares y son difíciles de estudiar experimentalmente.

- **Single-Sequence Structure Prediction**

Predicción de estructuras de proteínas a partir de una única secuencia de aminoácidos, sin utilizar alineamientos múltiples de secuencias.

Domain-Domain interactions

- **Protein-Ligand Interaction**
- **Protein-Protein Interaction**

Structure protein of assambles

- **Assembly**

Predicción de la estructura cuaternaria de complejos proteicos, es decir, cómo múltiples cadenas polipeptídicas se ensamblan para formar una estructura funcional.

- **Disorder Prediction**

Identificación de regiones en proteínas que no adoptan estructuras tridimensionales fijas y que permanecen flexibles o desordenadas.

RNA STRUCRTURE

- **RNA Structure Prediction**

Predicción de la estructura tridimensional de moléculas de ARN a partir de sus secuencias nucleotídicas.

Macromoleculas assambles

- **Complexes with Nucleic Acids**

Predicción de la estructura de complejos formados entre proteínas y ácidos nucleicos (ARN o ADN).

Resumen

Tasks

Nombre	Task group	Descripción	Paper	Métrica	State-of-art	Tipo	tarea
Secondary Structure	Structure Prediction	Predecir la estructura secundaria (hélice, hoja o otra) de cada aminoácido en una secuencia de proteína.	PEER PGlue TAPE	Accuracy		sequence-to-sequence	Prediction
Contact	Structure Prediction	Predecir si pares de aminoácidos en una secuencia de proteína están en contacto (a menos un δ de distancia)	CASPXV PEER TAPE	L/5		binary clasification $(x_i,x_j) \rightarrow y_{i,j} \in \{0,1\}$	Prediction
Remote Homology	evolutionary understanding	Clasificar secuencias de proteínas en una de 1195 posibles estructuras de pliegues	TAPE	Accuracy		multi classification $x \rightarrow y$	Detection

Nombre	Task group	Descripción	Paper	Métrica	State-of-art	Tipo	tarea
Fluorescence (Landscape)	Function Prediction engineering	(TAPE)Determina la intensidad de la log-fluoresence (PEER)Predice la intensidad de fluorescencia de mutantes de la proteína verde fluorescente	PEER TAPE	spearman		regression x→ y	Prediction
Stability (Landscape)	Function Prediction engineering	(TAPE)cada proteína de entrada x se mapea a una etiqueta y que mide las circunstancias más extremas en las que la prot mantiene su estructura por encima de un umbral de concentración (PEER) Evalúa la estabilidad de las proteínas bajo condiciones naturales. \y indica la medida experimental de estabilidad	PEER TAPE	spearman		regression x→ y	Prediction
β-Lactamase Activity	Function Prediction	Studies the activity among the first-orders mutants of the TEM-1 beta-lactamse protein	PEER	spearman		regression x→ y	Prediction
Solubility	Function Prediction	predict wheter the protein is soluble or not	PEER	Accuracy		classification x→y \in {0,1}	Prediction
Subcellular Localization	Localization Prediction	predicts where a natural protein locates in the cell. Ten possible localizations	PEER	Accuracy		multi classification x →y	Prediction
Binary Localization	Localization Prediction	simpler version of subcellular, the model predicts classify each protein to be either “membran-bound” or “soluble”	PEER	Accuracy		classification x→y \in {0,1}	Prediction
Fold Clasification	Structure Prediction	Classifies the global structural topology of a protein on the fold level, 1995 options.	PEER	Accuracy		multi classification x →y	Prediction
PPI en Levaduras	Prot-Prot interaction	predice si preteinas de levadura interactuan	PEER	Accuracy		classification x1,x2→ y \in {0,1}	Prediction

Nombre	Task group	Descripción	Paper	Métrica	State-of-art	Tipo	tarea
PPI en Humanos	Prot-Prot interaction	predice si preteinas de levadura interactuan	PEER PGlue	Accuracy		classification x1,x2→ y \in {0,1}	Prediction
Afinidad en PPI	Prot-Prot interaction	Estima la afinidad de unión y medida por pKd entre dos proteínas. pKd es una medida específica. La afinidad de unión se refiere a la fuerza con la que dos proteínas se unen entre sí. Una afinidad de unión alta indica que las proteínas se unen fuertemente, mientras que una afinidad baja indica una unión débil.	PEER	RMS		regression x1,x2→ y	Prediction
Afinidad en PDBbind	Prot-ligand interaction	Predicts the binding affinity between proteins and ligands on PDBbind-2019 dataset.	PEER	RMS		regression x1,ligand→ y	Prediction
Afinidad en BindingDB	Prot-ligand interaction	Predicts the binding affinity between proteins and ligands on BindingDB dataset.	PEER ProteinGym	RMS		regression x1,ligand→ y	Prediction
Solvent Accessibility (ASA)	engineering	measures the amount of surface area of an amino acid that is exposed to the solvent. Suponiendo que estan en un solvente, los aminoacidos externos están expuestos a iteractuar con el solvente	PGlue	MAE		Regresión x1→y \in R	
Solvent Accessibility (BUR)	engineering	measures the amount of surface area of an amino acid that is exposed to the solvent. Suponiendo que estan en un solvente, los aminoacidos externos están expuestos a iteractuar con el solvente	PGlue	MAE		clasificación x1→y \in {0,1}	
Hydrophobic Patch	engineering	dentificación de parches hidrofóbicos en	PGlue	pearson			

Nombre	Task group	Descripción	Paper	Métrica	State-of-art	Tipo	tarea
		la superficie de la proteína, importantes para interacciones con otras proteínas o membranas					
thermostability	engineering		ProteinGym				
mutations	Function Prediction engineering	anotaciones de cada mutación:maligna o benigna	ProteinGym	Accuracy			

Metrics

Métricas	Descripción	Paper
GDT_TS	evalúa la fracción de residuos en la estructura predicha que se encuentran dentro de ciertos umbrales de distancia (1, 2, 4 y 8 Å) de la estructura experimental.	CASP
Accuracy		
Spearman		
L/5		
RMSE	root-mean-square error	
ICS	Interface Contact Score. Específicamente, evalúa la precisión con la que se predicen los contactos entre los residuos de diferentes subunidades en un complejo proteico.	CASP

Otros Benchmarks

Papers	Link	Fecha	descripcion
PEER			
ProteinWorkshop			
ProteinInvBench			
ProteinGym			
Proteinglue			builds a benchmark containing 7 downstream tasks for evaluating self-supervised protein representation learning.
ATOM3D			provides benchmark datasets for 3D structure based biomolecule understanding. Muy específico
TDA			contains protein-related datasets and tasks for drug discovery. Muy específico
functional properties (nature)			focuses on the evaluation of unsupervised protein representations and evaluates 23 typical methods. Pendiente
FLIP	https://benchmark.protein.properties/home		proposes three protein landscape benchmarks for fitness prediction evaluation. Es muy específico. No muy citado
TAPE			comprehensively compare different machine learning methods, is built on five tasks spread across different domains of protein biology and evaluate the performance of protein sequence encoders (First attempt to benchmarking protein methods)
CAFA			Is held for the evaluation of PFP
CASP			Focuses on PSP (golden standard)

Prinicipales problemas en cada área:

1. **Predicción de la Estructura de Proteínas**:

* predecir con precisión las estructuras de proteínas multi-dominio y complejas.

*integrar mejor la información evolutiva y estructural.

2. **Anotación Funcional**:

*La predicción de funciones específicas de proteínas basadas en datos de secuencia y estructura

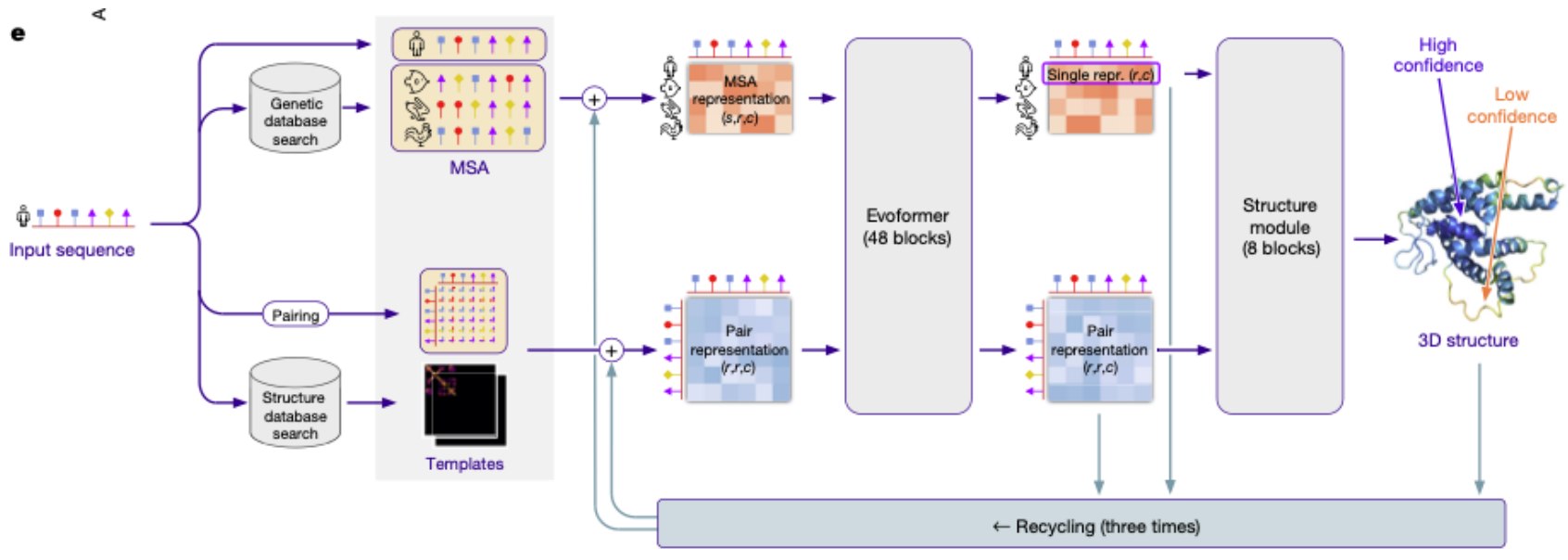
*Los modelos de lenguaje pueden ayudar a identificar sitios funcionales y prever interacciones de manera más precisa.

3. **Interpretabilidad de Modelos y Aprendizaje Multimodal**

*Mejorar la transparencia y la interpretabilidad de estos modelos puede aumentar su utilidad y confiabilidad en aplicaciones prácticas.

AF2

Utiliza características evolutivas, físicas y geométricas de las proteínas para realizar predicciones precisas de extremo a extremo.



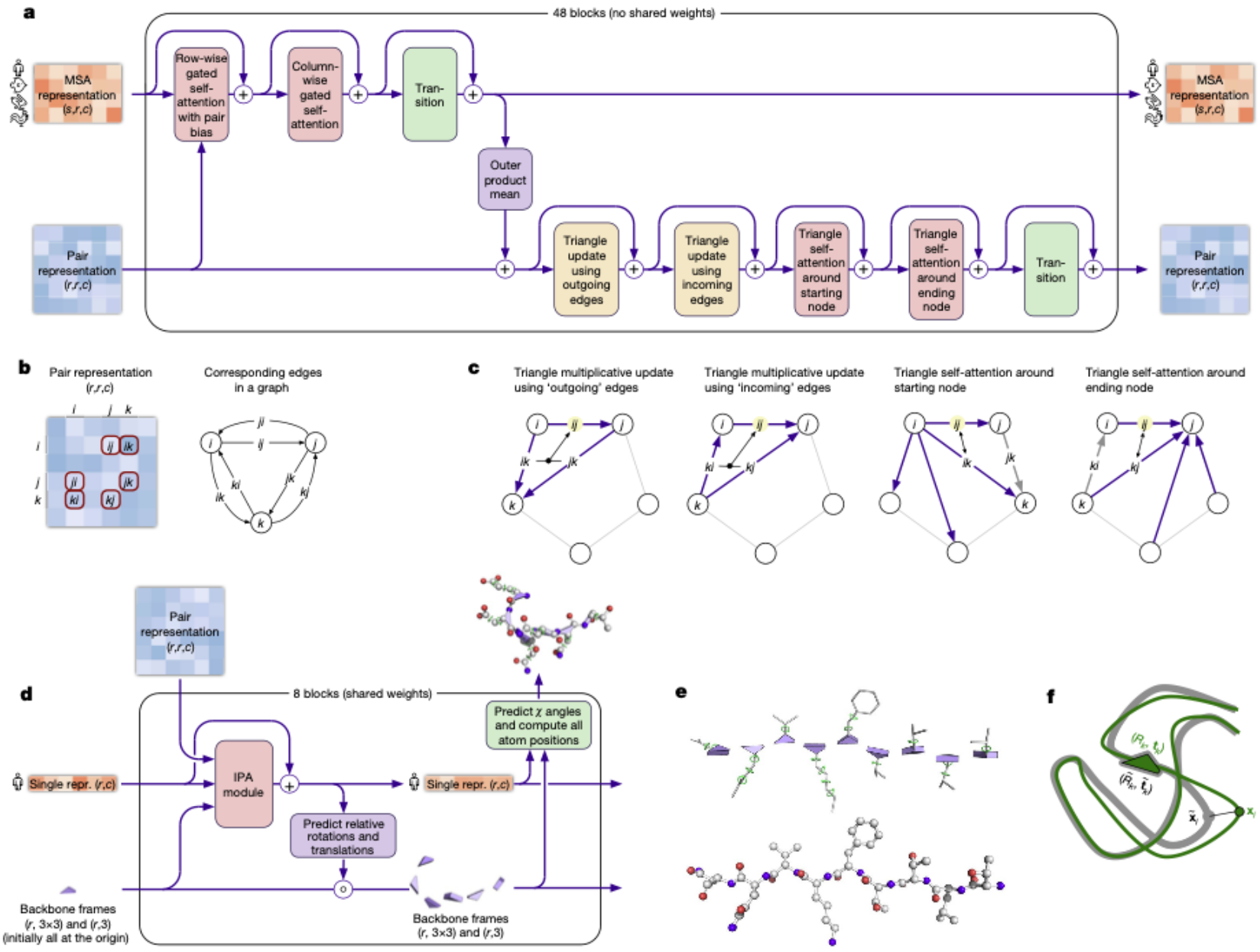
*alineación de secuencias múltiples, las filas representan diferentes secuencias y las columnas representan posiciones en la secuencia de aminoácidos.

*Las representaciones de pares de residuos se actualizan utilizando operaciones de atención.

analiza cómo interactúan diferentes pares de aminoácidos en la secuencia. Las operaciones de atención permiten a la red “enfocar” partes específicas de la secuencia para entender las relaciones entre los residuos (los bloques de construcción de las proteínas).

*Se utiliza una operación de atención geométrica que actualiza las activaciones neuronales de cada residuo en el marco local sin cambiar sus posiciones 3D. Esto se realiza proyectando puntos de consulta, claves y valores en el marco global y luego de vuelta al marco local.

Refinamiento Iterativo: La red refina continuamente esta estructura inicial en pasos sucesivos. Cada refinamiento mejora la precisión de la estructura.



TODO:

- Solubilidad
- Fluoresencia
- Bert / NLP Bert
- con CNN