



**TECNOLÓGICO
DE MONTERREY®**

MAESTRÍA EN INTELIGENCIA ARTIFICIAL APLICADA

Materia:

Proyecto Integrador

Avance 6

EQUIPO 29

Miguel Angel Favela Corella A00818504

Kevin Balderas Sánchez A01795149

José Manuel García Ogarrio A01795147

Sponsor del Proyecto

- **Dr. Juan Arturo Nolazco** – Director del Data Science Hub, Tecnológico de Monterrey.

Colaborador Invitado

- **Dr. Carlos Alberto Brizuela Rodríguez** – Investigador, CICESE (Centro de Investigación

Contexto del proyecto

El presente proyecto tiene como objetivo aplicar herramientas de inteligencia artificial y biología computacional para el diseño y análisis de proteínas con potencial aplicación biomédica. Nuestro foco principal se centra en el estudio del receptor GLP-1R, una proteína de membrana que desempeña un papel esencial en la regulación de la glucosa y del metabolismo energético.

El GLP-1R es activado por el péptido GLP-1, lo que estimula la secreción de insulina y reduce el apetito. Por esta razón, este receptor constituye una diana terapéutica importante en el tratamiento de la diabetes tipo 2 y la obesidad. Comprender su estructura, sus interacciones y posibles moduladores ofrece oportunidades para el diseño racional de nuevas moléculas bioactivas.

Nuestro enfoque combina distintas plataformas computacionales de vanguardia: RFDiffusion, MPNN, AlphaFold, y próximamente Rosetta y REF15, con el propósito de obtener una visión integral del diseño estructural, la viabilidad funcional y las propiedades bioquímicas de las proteínas generadas.

Herramientas

- **RFDiffusion**

RFDiffusion es un modelo generativo basado en el principio de difusión probabilística, diseñado para la creación de nuevas estructuras proteicas de forma controlada y coherente. Este sistema aprende patrones estructurales a partir de proteínas reales y posteriormente puede generar diseños que cumplan con restricciones espaciales o funcionales específicas.

En este proyecto, RFDiffusion se emplea para proponer estructuras candidatas que potencialmente interactúen con el receptor GLP-1R. La finalidad es generar proteínas que puedan actuar como ligandos o moduladores del receptor, conservando estabilidad estructural y compatibilidad con el entorno biológico.

- **ProteinMPNN**

ProteinMPNN sirve para procesar grafos moleculares, donde los nodos representan átomos y las aristas los enlaces químicos. Esta arquitectura permite modelar las interacciones locales y globales entre los componentes de una molécula o proteína.

Su función principal es predecir la secuencia de aminoácidos que mejor se ajusta a una estructura tridimensional predefinida. El modelo recibe como entrada una estructura generada por RFDiffusion y, mediante el intercambio de información entre los nodos del grafo (que representan los residuos de aminoácidos), aprende qué combinaciones son más estables y coherentes con la geometría original.

El resultado son una o varias secuencias candidatas que podrían adoptar la forma estructural deseada. Posteriormente, estas secuencias se validan mediante AlphaFold,

que predice si la secuencia realmente se pliega de manera similar a la estructura original.

- **AlphaFold**

AlphaFold, es una herramienta de predicción estructural que utiliza aprendizaje profundo para determinar la conformación tridimensional de proteínas a partir únicamente de su secuencia de aminoácidos.

Su integración en el proyecto tiene como propósito verificar la estabilidad estructural de las secuencias diseñadas con RFDiffusion y ProteinMPNN. Esta herramienta nos permite visualizar las estructuras resultantes y evaluar si adoptan configuraciones energéticamente estables y coherentes con la función prevista. De esta forma, actúa como una herramienta de validación y análisis estructural.

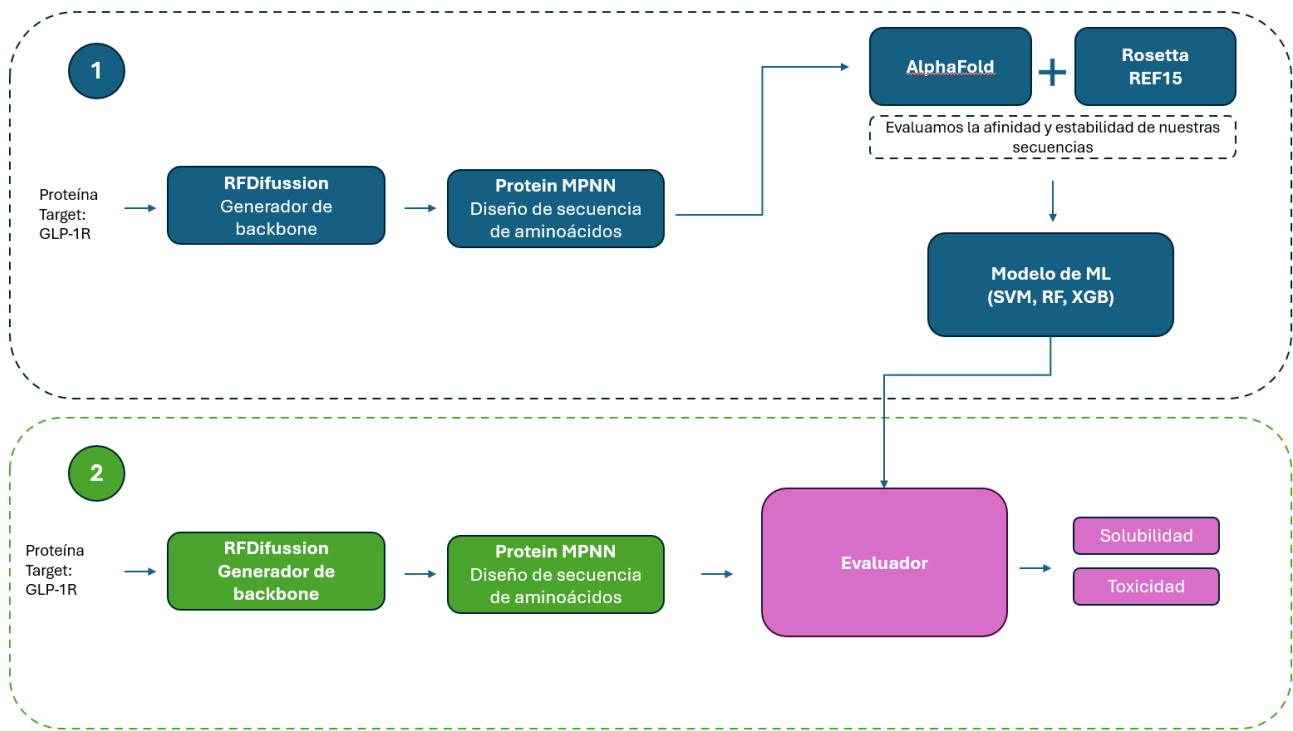
- **Rosetta y REF15**

Rosetta es una de las plataformas computacionales más avanzadas y versátiles para el modelado, diseño y simulación de proteínas. Permite realizar predicciones estructurales, análisis de estabilidad, acoplamiento molecular y optimización energética de biomoléculas. Su funcionamiento se basa en un conjunto de funciones de energía que estiman la estabilidad y la plausibilidad física de una estructura proteica.

Dentro de Rosetta, una de las funciones energéticas más utilizadas es REF15. Esta función integra parámetros físicos y estadísticos que permiten calcular con precisión las energías de interacción, enlaces de hidrógeno, fuerzas hidrofóbicas, interacciones electrostáticas, y energías de solvatación.

En el contexto de este proyecto, Rosetta y REF15 se emplearán conjuntamente para analizar y refinar las proteínas diseñadas mediante RFDiffusion y ProteinMPNN. El objetivo es evaluar no solo la estabilidad estructural y energética, sino también incorporar criterios adicionales relacionados con la toxicidad y la solubilidad de las proteínas generadas.

Etapas y planificación.



Descripción del Flujo de Trabajo

El flujo general del proyecto se divide en dos etapas principales, representadas en el diagrama anterior. Ambas tienen como punto de partida la proteína target GLP-1R, sobre la cual se generan y evalúan nuevas secuencias peptídicas con potencial afinidad estructural y funcional.

Etapla 1: Generación y evaluación inicial de secuencias

En la primera fase del proyecto (azul), se utiliza RFDiffusion como generador del backbone proteico, es decir, la estructura tridimensional base sobre la cual se diseñarán nuevas secuencias. Posteriormente, ProteinMPNN predice en este caso como ejemplo ~1,000/5,000 secuencias de aminoácidos que podrían adoptar dicha estructura.

Estas secuencias se someten a una evaluación estructural y energética mediante AlphaFold y Rosetta (REF15), con el fin de determinar propiedades como la afinidad, estabilidad y compatibilidad estructural.

El principal desafío en esta etapa es el alto costo computacional asociado al cálculo detallado de estas propiedades para miles de secuencias, lo que limita la escalabilidad del proceso.

Para superar esta limitación, se propone entrenar un modelo de ML, empleando algoritmos como SVM, Random Forest o XGBoost, usando un subconjunto representativo de las secuencias generadas.

Este modelo actuará como un predictor de afinidad y estabilidad, reduciendo significativamente el tiempo necesario para evaluar grandes volúmenes de datos sin sacrificar precisión.

Etapas 2: Evaluación masiva mediante modelo predictivo

En la segunda fase (indicada en verde y rosa), se repite el proceso de diseño estructural mediante RFDiffusion y ProteinMPNN, pero esta vez a mayor escala, generando muchas secuencias candidatas.

En lugar de evaluar individualmente cada una mediante simulaciones intensivas, las secuencias se procesan con el modelo evaluador desarrollado en la primera etapa. Este evaluador integra el modelo de ML y permite estimar de manera rápida propiedades clave como la solubilidad y toxicidad.

De esta forma, se priorizan las secuencias con mejores características, reduciendo la carga computacional y enfocando los recursos en aquellas con mayor potencial biotecnológico.

Anexo de Resultados Visuales y Evidencia Computacional

1. Propósito del anexo

Este anexo documenta de manera independiente los resultados visuales obtenidos durante la ejecución del pipeline computacional para el diseño de proteínas orientadas al tratamiento de la diabetes. Se presentan las capturas de ejecución, visualizaciones estructurales y métricas derivadas de los modelos RFDiffusion, ProteinMPNN y AlphaFold, con una descripción técnica clara y contextualizada.

2. Flujo visual de trabajo

2.1 Generación de esqueletos proteicos con RFDiffusion

Imágenes correspondientes a la ejecución local del modelo RFDiffusion, implementado en VSCode mediante notebooks de Jupyter.

Descripción técnica: El modelo generó 10 esqueletos (“backbones”) en formato .pdb a partir de parámetros relacionados con péptidos candidatos para diabetes.

Tiempo de procesamiento: ~35 minutos en GPU GeForce GTX 1650.

Visualizador utilizado: Mol* Viewer.

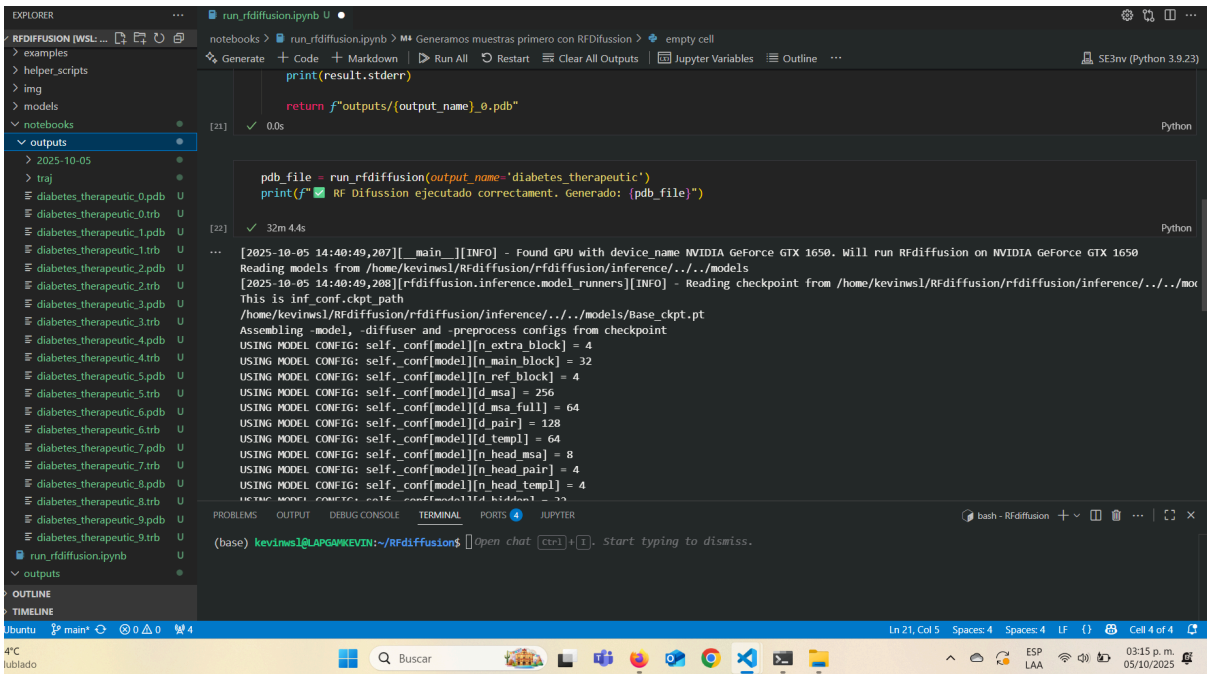


Figura 1. Ejecución del modelo RFdiffusion en entorno local.

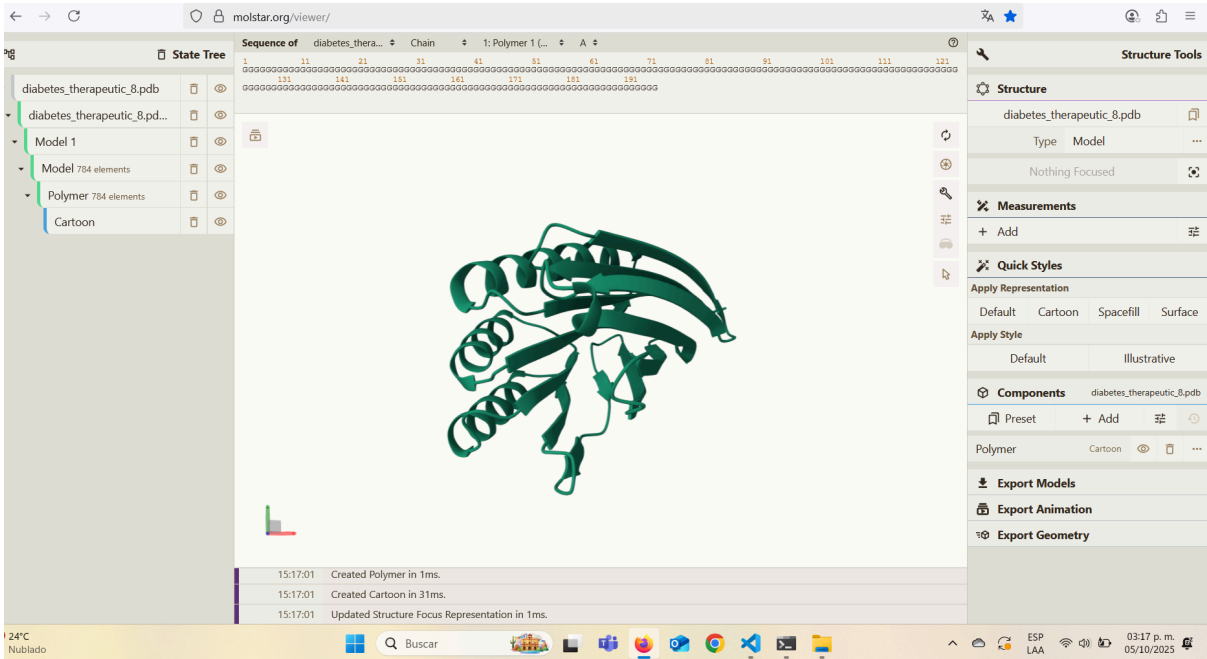


Figura 2. Visualización de estructuras generadas (.pdb) en MolViewer.

2.2 Generación de secuencias aminoacídicas con ProteinMPNN

Descripción técnica: Se utilizó el script `protein_mpnn_run.py` dentro del entorno Conda configurado. A partir de los archivos .pdb generados por RFdiffusion, se obtuvieron secuencias en formato .FASTA.

Resultados: Se generaron 10 secuencias con scores entre 1.009 y 1.2303 (valores de log-probabilidad negativa).

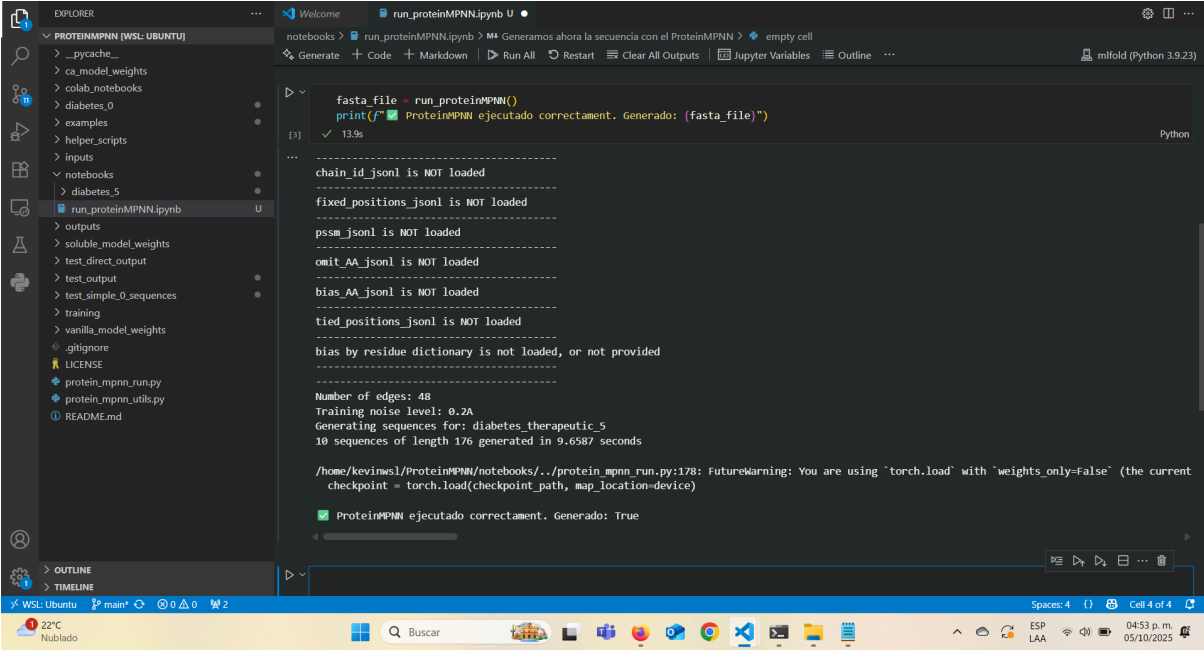


Figura 3. Ejecución de ProteinMPNN para obtención de secuencias.

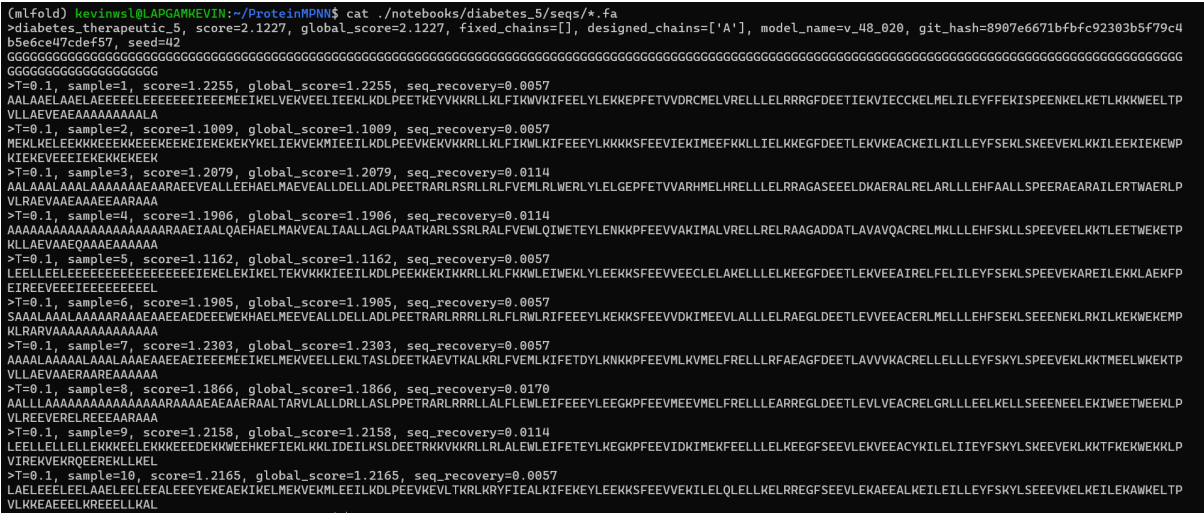


Figura 4. Visualización de resultados y scores obtenidos.

2.3 Validación estructural con AlphaFold2

Descripción técnica: Se empleó colabfold_batch para validar las secuencias generadas. Para cada muestra, se produjeron cinco modelos estructurales.

Parámetros de rendimiento:

- Total de combinaciones: 55 validaciones.
- Duración total: ~1h 14min.
- Métricas utilizadas: pLDDT, pTM y RMSD.

Resultados destacados:

- RMSD inicial: 9.16 Å (muestra compleja).
- RMSD optimizado en monómero simple: 0.31 Å.

```
kevin@LAPGAMKEVIN: ~$ cat log.txt
2025-10-06 00:37:42,335 alphafold2_ptm_model_3_seed_000 recycle=2 pLDDT=78.6 pTM=0.653 tol=1
2025-10-06 00:38:01,057 alphafold2_ptm_model_3_seed_000 recycle=3 pLDDT=76.8 pTM=0.64 tol=0.278
2025-10-06 00:38:01,057 alphafold2_ptm_model_3_seed_000 took 76.2s (3 recycles)
2025-10-06 00:38:21,357 alphafold2_ptm_model_4_seed_000 recycle=0 pLDDT=72.2 pTM=0.585
2025-10-06 00:38:40,316 alphafold2_ptm_model_4_seed_000 recycle=1 pLDDT=78.8 pTM=0.669 tol=1.07
2025-10-06 00:38:59,645 alphafold2_ptm_model_4_seed_000 recycle=2 pLDDT=76.9 pTM=0.639 tol=0.427
2025-10-06 00:39:20,699 alphafold2_ptm_model_4_seed_000 recycle=3 pLDDT=77.8 pTM=0.633 tol=0.221
2025-10-06 00:39:20,700 alphafold2_ptm_model_4_seed_000 took 79.6s (3 recycles)
2025-10-06 00:39:40,135 alphafold2_ptm_model_5_seed_000 recycle=0 pLDDT=82.1 pTM=0.72
2025-10-06 00:40:00,743 alphafold2_ptm_model_5_seed_000 recycle=1 pLDDT=80.9 pTM=0.696 tol=0.287
2025-10-06 00:40:20,067 alphafold2_ptm_model_5_seed_000 recycle=2 pLDDT=78.3 pTM=0.674 tol=0.157
2025-10-06 00:40:39,467 alphafold2_ptm_model_5_seed_000 recycle=3 pLDDT=78.7 pTM=0.674 tol=0.168
2025-10-06 00:40:39,467 alphafold2_ptm_model_5_seed_000 took 78.7s (3 recycles)
2025-10-06 00:40:39,495 reranking models by 'plddt' metric
2025-10-06 00:40:39,495 rank_001_alphafold2_ptm_model_1_seed_000 pLDDT=78.9 pTM=0.65
2025-10-06 00:40:39,495 rank_002_alphafold2_ptm_model_5_seed_000 pLDDT=78.7 pTM=0.674
2025-10-06 00:40:39,496 rank_003_alphafold2_ptm_model_2_seed_000 pLDDT=78 pTM=0.688
2025-10-06 00:40:39,496 rank_004_alphafold2_ptm_model_4_seed_000 pLDDT=77.8 pTM=0.633
2025-10-06 00:40:39,497 rank_005_alphafold2_ptm_model_3_seed_000 pLDDT=76.8 pTM=0.64
2025-10-06 00:40:40,780 Query 7/11: T_0.1__sample_6__score_1.1905__global_score_1.1905__seq_recovery_0.0057 (length 176)
2025-10-06 00:40:41,341 Sleeping for 5s. Reason: PENDING
2025-10-06 00:40:46,787 Sleeping for 7s. Reason: RUNNING
COMPLETE: 100%| 150/150 [elapsed: 00:13 remaining: 00:00]
2025-10-06 00:41:14,367 alphafold2_ptm_model_1_seed_000 recycle=0 pLDDT=77.8 pTM=0.685
2025-10-06 00:41:32,808 alphafold2_ptm_model_1_seed_000 recycle=1 pLDDT=81.8 pTM=0.743 tol=1
2025-10-06 00:41:52,854 alphafold2_ptm_model_1_seed_000 recycle=2 pLDDT=81.1 pTM=0.739 tol=0.406
2025-10-06 00:42:11,625 alphafold2_ptm_model_1_seed_000 recycle=3 pLDDT=81.5 pTM=0.738 tol=0.8987
2025-10-06 00:42:11,625 alphafold2_ptm_model_1_seed_000 took 75.9s (3 recycles)
2025-10-06 00:42:32,183 alphafold2_ptm_model_2_seed_000 recycle=0 pLDDT=78.1 pTM=0.713
2025-10-06 00:42:51,044 alphafold2_ptm_model_2_seed_000 recycle=1 pLDDT=81.8 pTM=0.76 tol=1.23
2025-10-06 00:43:10,007 alphafold2_ptm_model_2_seed_000 recycle=2 pLDDT=82.4 pTM=0.766 tol=0.456
2025-10-06 00:43:30,670 alphafold2_ptm_model_2_seed_000 recycle=3 pLDDT=82.6 pTM=0.769 tol=0.24
2025-10-06 00:43:30,671 alphafold2_ptm_model_2_seed_000 took 79.0s (3 recycles)
2025-10-06 00:43:49,771 alphafold2_ptm_model_3_seed_000 recycle=0 pLDDT=77.8 pTM=0.719
2025-10-06 00:44:10,156 alphafold2_ptm_model_3_seed_000 recycle=1 pLDDT=81.4 pTM=0.756 tol=0.785
2025-10-06 00:44:30,268 alphafold2_ptm_model_3_seed_000 recycle=2 pLDDT=82.1 pTM=0.76 tol=0.179
2025-10-06 00:44:49,046 alphafold2_ptm_model_3_seed_000 recycle=3 pLDDT=82.7 pTM=0.768 tol=0.117
2025-10-06 00:44:49,047 alphafold2_ptm_model_3_seed_000 took 78.3s (3 recycles)
2025-10-06 00:45:08,094 alphafold2_ptm_model_4_seed_000 recycle=0 pLDDT=80.5 pTM=0.74
2025-10-06 00:45:28,461 alphafold2_ptm_model_4_seed_000 recycle=1 pLDDT=82.8 pTM=0.773 tol=1.69
```

Figura 5. Ejecución de AlphaFold y generación de modelos.



Figura 6. Superposición de estructuras generadas (RFdiffusion vs AlphaFold).

2.4 Integración de pipeline

Flujo de integración logrado:

RFdiffusion → ProteinMPNN → AlphaFold2

Conclusión visual: Se demuestra la viabilidad del pipeline completo para la generación y validación de candidatos peptídicos, dejando como etapa futura la optimización de parámetros y el entrenamiento con Rosetta.

2.5 Obtención energética con Rosetta

Descripción técnica: Se empleó el módulo de PyRosetta nativo de Python para evaluar las nuevas secuencias (archivos pdb) generados previamente por AlphaFold. El script `calculate_stability.py` crea una función de energía y luego la calcula en base a nuestro .pdb, obteniendo métricas útiles. Pero el segundo script `calculate_stability_completed.py` aplicó la FastRelax para así obtener métricas importantes como el RMSD que servirán para nuestros modelos futuros evaluadores.

Antes se presenta el pdb utilizado para esta prueba:

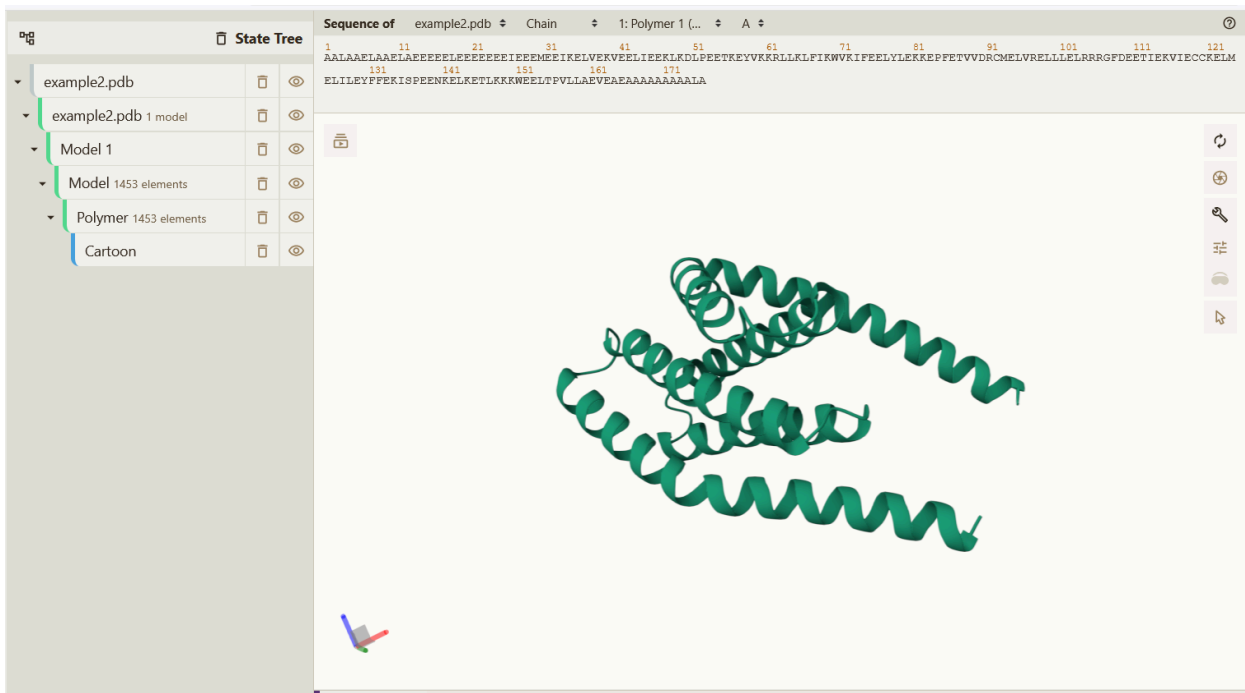


Figura 7. Archivo pdb utilizado para el proceso de Rosetta.

```
File Edit Selection View Go Run ...
roseta_scripts_estabilidad [WSL: Ubuntu]
EXPLORER
  ROSETTA_SCRIPTS_ESTABILIDAD [WSL: UB...
    .vscode
    scripts
      calculate_stability.py
      calculate_stability_completed.py
      example.pdb
      example2_relaxed.pdb
      example2.pdb
scripts > calculate_stability.py > ...
5
6 # Inicializar PyRosetta
7 pyrosetta.init()
8 init_time = datetime.now()
9
10 # Cargar tu estructura de AlphaFold2
11 pose = pose_from_pdb("example2.pdb")
12
13 # Crear función de scoring (energía)
14 scorefxn = get_fa_scorefxn() # Función de energía full-atom
15
16 # Calcular energía total
17 total_score = scorefxn(pose)
18
19 print(f"50")
20 print("ANÁLISIS DE ESTABILIDAD")
21 print(f"50")
22 print(f"Energía Total (REU): {total_score:.2f}")
23 print()
PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS
(rosetta_env) kevin@LAPGAMKEVIN:~/roseta_scripts_estabilidad/scripts$ python calculate_stability.py
Energía Total (REU): 951.89

Desglose de términos energéticos:
-----
fa_atr      : -1092.81
fa_rep      : 2596.34
fa_sol      : 757.54
fa_elec     : -290.08
hbond_sc    : -0.73
hbond_bb_sc : -1.48
omega       : 20.21
rama_prepro : -57.43
p_aa_pp     : -44.49
-----

Energía por residuo (top 10 más inestables):
```

Figura 8. Ejecución del script que calcula función de energía full-atom

```
=====
Energía por residuo (top 10 más inestables):
-----
Residuo#  Tipo  Energía (REU)
      65  LYS      174.40
      22  GLU      173.90
      29  GLU       98.55
      72  LYS       96.83
     108  PHE       73.63
      55  THR       70.00
     106  ARG       61.14
      58  TYR       58.10
     132  PHE       40.43
     137  PRO       28.23
=====
Tiempo de ejecución: 0:00:00.697646
Tiempo en segundos: 0.697646
(rosetta env) kevinwsl@LAPGAMKEVIN:~/roseta_scripts_estabilidad/scripts$
```

Figura 9. Detalles técnicos de la función de energía full-atom

Parámetros de rendimiento:

- PDB utilizado: Archivo de 171 cadenas
- Duración total: 0.69 segundos
- Métricas utilizadas: Atracción de van der Waals , Solvatación, Electrostatica

Resultados destacados:

- Resumen de métricas a nivel global o general del pdb utilizado
- Análisis de tiempo utilizado durante su ejecución

Ahora se presenta el segundo script creado donde ya usamos Fast-Relax y obtenemos más métricas importantes.

```
(rosetta_env) kevinws1@LAPGAMKEVIN:~/roseta_scripts_estabilidad/scripts$ python calculate_stability_completed.py
(C) Copyright Rosetta Commons Member Institutions

NOTE: USE OF PyRosetta FOR COMMERCIAL PURPOSES REQUIRES PURCHASE OF A LICENSE
See LICENSE.PyRosetta.md or email license@uw.edu for details

PyRosetta-4 2025 [Rosetta PyRosetta4.Release.python39.ubuntu.cxx11thread.serialization 2025.39+release.c389ea27189ed431c5f30718cf7d86843f
eeab8e 2025-09-24T10:26:10] retrieved from: http://www.pyrosetta.org
Ejecutando FastRelax para optimizar estructura...

=====
RESUMEN DE ESTABILIDAD
=====
Archivo: example2.pdb
Número de residuos: 176

Energía inicial:      951.89 REU
Energía post-relax:   -556.77 REU
Cambio energético:    -1508.66 REU
Energía/residuo:      5.41 REU/res
RMSD (Cα):           2.22 Å

INTERPRETACIÓN:
-----
⚠ Estructura MUY INESTABLE (cambio > -5 REU)
⚠ RMSD alto (2.22 Å) - estructura se deformó mucho
=====

Estructura relajada guardada en: example2.pdb
Tiempo de ejecución: 0:01:11.347443
Tiempo en segundos: 71.347443
(rosetta_env) kevinws1@LAPGAMKEVIN:~/roseta_scripts_estabilidad/scripts$
```

Figura 10. Resultados luego de aplicar Fast-Relax.

```
EXPLORER
ROSETA_SCRIPTS_EST...
> .vscode
  scripts
    calculate_stability_completed.py
    calculate_stability.py 9
    example.pdb
    example2.pdb
    example2_relaxed.pdb
    example2.pdb

scripts > calculate_stability_completed.py > analizar_estructura
def analizar_estructura(pdb_file):
10
61     elif energy_change < 0:
62         print("✓ Estructura aceptable (cambio < -2 REU)")
63     else:
64         print("✓ Estructura MUY ESTABLE (energía aumentó o igual)")
65
66     if rmsd > 2.0:
67         print(f"⚠ RMSD alto ({rmsd:.2f} Å) - estructura se deformó mucho")
68     elif rmsd > 1.0:
69         print(f"⚠ RMSD moderado ({rmsd:.2f} Å)")
70     else:
71         print(f"✓ RMSD bajo ({rmsd:.2f} Å) - estructura estable")
72
73     print("=*60)
74
75     # Guardar estructura relajada
76     output_file = pdb_file.replace('.pdb', 'relaxed.pdb')
77     relaxed_pose.dump_pdb(output_file)
78     print(f"\nEstructura relajada guardada en: {output_file}")
79
80     return {
81         'total_energy': total_energy,
```

Figura 11. Estructura de la etapa de Rosetta.

Parámetros de rendimiento:

- PDB utilizado: Archivo de 171 cadenas.
- Duración total: 1.11 minutos.
- Método utilizado: Fast-Relax

Resultados destacados:

- Energía inicial, Energía Post-Relax, Cambio Energético, RMSD
- Análisis de tiempo utilizado durante su ejecución

Por último , en esta etapa se puede concluir que PyRosseta es un módulo muy completo para calcular métricas relacionadas con la energía, así como el método Fast-Relax, el cual es un protocolo que optimiza la geometría molecular. Es decir, luego de los resultados obtenidos mediante AlphaFold, puede existir detalles en su geometría de enlace (no óptima) o energías locales subóptimas. Así, podemos decir que tenemos ya el siguiente flujo logrado:

Flujo de integración logrado:
RFdiffusion → ProteinMPNN → AlphaFold2 → Rosseta

3. Desarrollo de Predictores (Alphafold y Rosetta)

3.1 Desarrollo de predictor de Alphafold

Para el desarrollo de este modelo predictor primero es necesario identificar entradas y salidas de modelo. Una vez que ya pudimos hacer funcionar los ambientes y el ciclo normal de generación de prospectos para el target que nos interesa, la entrada de alphafold son las cadenas de secuencias (que serían las salidas luego de correr ProteinMPNN) y de salida tenemos que la variable target es el ipTM (Interface predicted template modelling), la cual es una métrica que nos da alphafold luego de utilizarlo y que aproximadamente para uno de los primero ciclos de 100 muestras que generamos, duró cerca de 6.30 horas.

Una vez identificado eso y que justamente queremos hacer un sustituto para encontrar esa métrica, se proponen los siguientes diagramas, que son para la parte de entrenamiento, inferencia y de deployment del predictor.

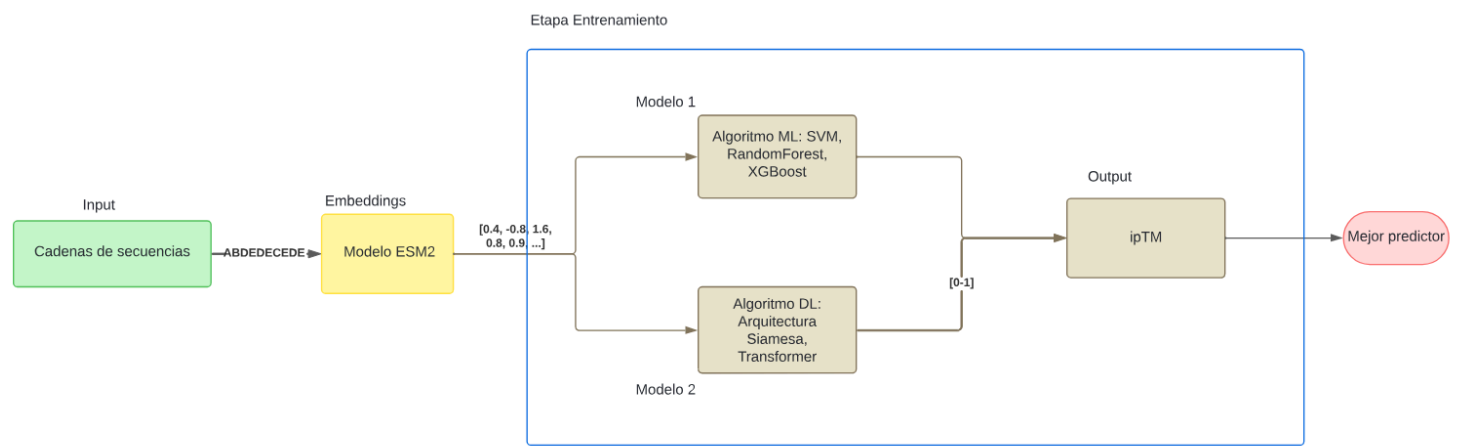


Figura 12.Diagrama de bloques para entrenamiento del predictor de AlphaFold

Es importante subrayar que la conversión de secuencias proteicas a representaciones numéricas no puede realizarse mediante técnicas simples como *one-hot encoding*, dado que éstas no capturan las relaciones biológicas ni las dependencias contextuales entre residuos. En su lugar, se propone el uso de modelos de embeddings especializados, capaces de preservar las interacciones latentes y las propiedades bioquímicas implícitas de las cadenas.

Posteriormente, se plantean dos estrategias complementarias de entrenamiento:

- Enfoque clásico de aprendizaje automático, utilizando algoritmos como *Support Vector Machines (SVM)*, *Random Forest* o *XGBoost*, para obtener modelos interpretables y con bajo costo computacional.
- Enfoque de aprendizaje profundo, explorando arquitecturas como *Transformers* o redes *Siamesas*, con el fin de capturar dependencias de largo alcance y relaciones no lineales entre secuencias.

Los resultados de ambos enfoques se evaluarán comparando las predicciones del ipTM frente a los valores obtenidos experimentalmente, utilizando métricas de error como el Root Mean Squared Error (RMSE) para cuantificar el desempeño predictivo.

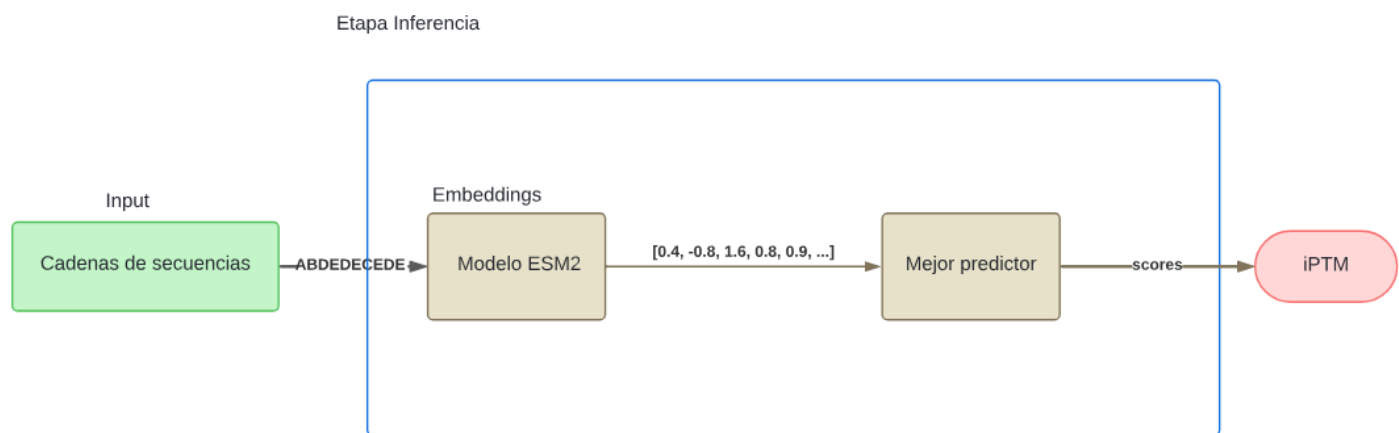


Figura 13. Diagrama de bloques para inferencia del predictor de Alphafold

Finalmente, en la fase de implementación (deployment), se contempla la creación de una API o contenedor modular, capaz de recibir una secuencia de aminoácidos como entrada, ejecutar el proceso de embedding y retornar de forma automatizada la predicción del ipTM. Esta arquitectura permitirá la integración del modelo en flujos de trabajo bioinformáticos más amplios.

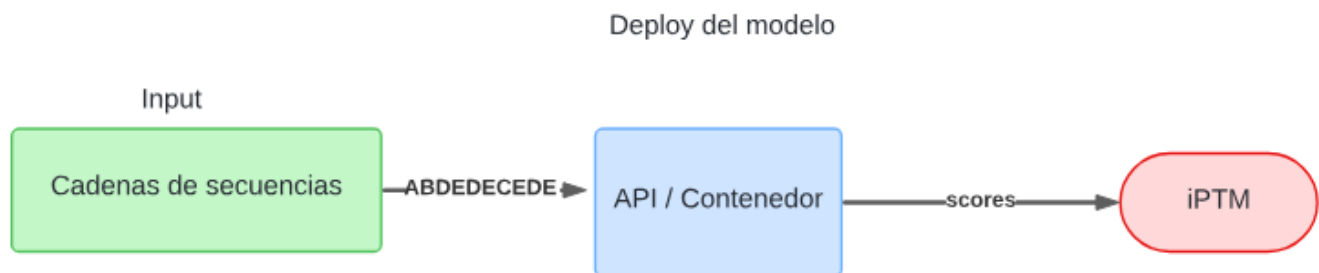


Figura 14. Diagrama de bloques del deploy del predictor de Alphafold

3.2 Desarrollo de organización del deploy del predictor de Alphafold

Ahora, una vez que ya tenemos los prospectos del modelo que sería el predictor de Alphafold, es necesario hacer un ajuste en el repositorio de código que ya tenemos. Esto para facilitar el entendimiento y mejorar la organización de archivos. Como vamos a tener un modelo ganador final, se tiene pensado hacer el deploy mediante un contenedor de Docker. También haciendo notar que uno de los entregables finales será el contenedor que hace todas las etapas para replicación de backbones, es lógico proponer un contenedor para consumir el modelo ganador para esta etapa del predictor de Alphafold. Por último, también una vez con el conocimiento de que el modelo tomará como entrada la cadena y como salida las dos métricas pLDDT e ipTM, es fácil intuir que se proponen dos modos de operación: Uno de tipo cadena y otro de tipo batch. El tipo cadena únicamente tiene como objetivos dar las métricas para esa única cadena que se la da como entrada mediante la consola de comandos, y el tipo batch como su nombre lo indica, cuando tienes un archivo CSV con el formato adecuado, y a la salida tendrás también un CSV pero con las métricas predichas.

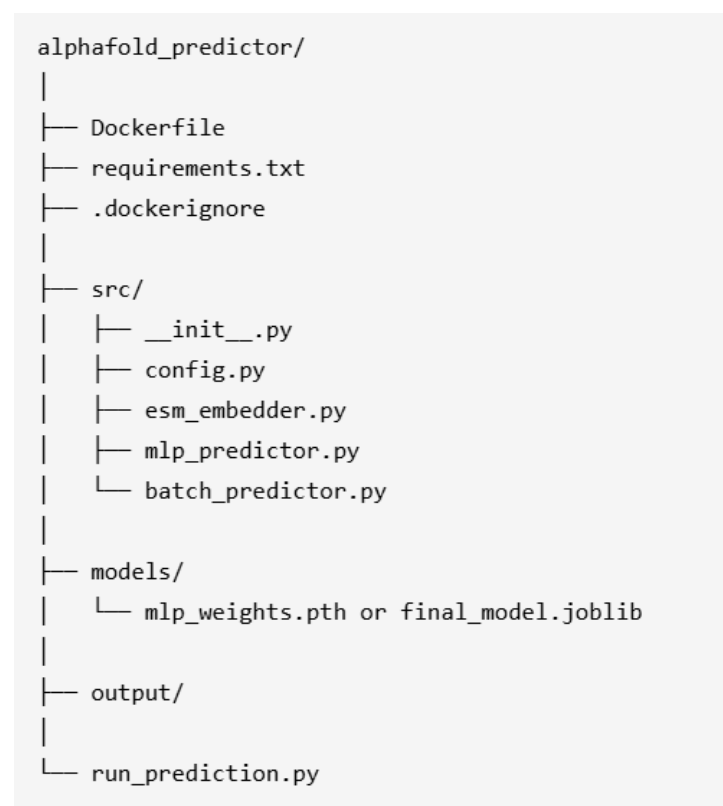


Figura 15. Estructura de archivos para tener el contenedor predictor de Alphafold

Explicando ahora cada archivo y su finalidad tenemos:

- config.py: Define rutas y parámetros centralizados, también variables de entorno como BATCH_SIZE, MAX_SEQ_LENGTH
- requirements.txt: Paquetes o dependencias de python.
- Dockerfile: Es de donde se tomará el build de docker y cuida que para el momento del modelo de embeddings ESM, sólo se descarga en build time y así no cuando se haga run o se hagan duplicidades innecesarias.
- run_prediction.py: De acuerdo al modo de ejecución sea single o batch, hace el proceso correspondiente
- outputs: Guarda los resultados en csv
- models: Tiene el valor de los pesos o parámetros de acuerdo al modelo ganador final.

- mlp_predictor.py: Clase de tipo Arquitectura del Multi Layer Perceptron
- esm_embedder.py: Clase que toma como entrada las cadenas y las pasa por este modelo de embeddings y así obtener los formatos numéricos que necesita el predictor
- batch_predictor.py: Es el archivo que hace la inferencia y únicamente pasa el resultado a run_prediction, para seguir manteniendo modularidad.

Y al final el comando para ejecutar tanto el build como el run en docker serían:

```
docker build -t batch-predictor:latest .
```

Para ver ayuda sobre los comandos y flags disponibles:

```
docker run batch-predictor:latest --help
```

Luego para el modo secuencia para una predicción única:

```
docker run batch-predictor:latest --mode single --sequence "MKTLILAFLFASA"
```

Y para el modo batch:

```
docker run batch-predictor:latest --mode batch --input /app/input.csv
```

Para validar el resultado antes de ejecutar el contenedor, vemos como se usaría una vez esté corriendo de local el predictor de pLDDT e ipTM de AlphaFold. Primero vemos que pasa si no pasamos ninguna cadena o ningún modo de operación:

```
(mi_env_gpu) C:\Users\balde\Desktop\MAESTRIA MNA\PI\drug-design-diabetes-team29\alphafold_predictor>python run_prediction.py
usage: run_prediction.py [-h] --mode {single} --sequence SEQUENCE
run_prediction.py: error: the following arguments are required: --mode, --sequence
```

Y cuando damos la flag -h o --help de ayuda:

```
(mi_env_gpu) C:\Users\balde\Desktop\MAESTRIA MNA\PI\drug-design-diabetes-team29\alphafold_predictor>python run_prediction.py -h
usage: run_prediction.py [-h] --mode {single} --sequence SEQUENCE

Predictor de proteínas usando ESM + MLP

options:
  -h, --help            show this help message and exit
  --mode {single}       Modo de predicción: single (una secuencia)
  --sequence SEQUENCE   Secuencia de aminoácidos para predecir

Ejemplos de uso:

Predicción individual:
  python run_prediction.py --mode single --sequence "MKTLILAFLFASA"

Desde Docker:
  docker run batch-predictor:latest --mode single --sequence "MKTLILAFLFASA"
```

Y empezamos dando una cadena, viendo los logs de todo el proceso:

```
(mi_env_gpu) C:\Users\balde\Desktop\MAESTRIA MNA\PI\drug-design-diabetes-team29\alphafold_predictor>python run_prediction.py --mode single --sequence "ADEDE"
2025-11-03 22:37:57 - __main__ - INFO - =====
2025-11-03 22:37:57 - __main__ - INFO - Iniciando Batch Predictor
2025-11-03 22:37:57 - __main__ - INFO - =====
2025-11-03 22:37:57 - __main__ - INFO - Cargando modelos...
2025-11-03 22:37:57 - src.batch_predictor - INFO - Inicializando BatchPredictor...
2025-11-03 22:37:57 - src.esm_embedding - INFO - Cargando modelo ESM: facebook/esm2_t33_650M_UR50D
Some weights of EsmModel were not initialized from the model checkpoint at facebook/esm2_t33_650M_UR50D and are newly initialized: ['pooler.dense.bias', 'pooler.dense.weight']
You should probably TRAIN this model on a down-stream task to be able to use it for predictions and inference.
2025-11-03 22:38:02 - src.esm_embedding - INFO - Modelo ESM cargado exitosamente
2025-11-03 22:38:02 - src.batch_predictor - INFO - Cargando modelo MLP desde C:\Users\balde\Desktop\MAESTRIA MNA\PI\drug-design-diabetes-team29\alphafold_predictor\models\mlp_weights.pth
2025-11-03 22:38:02 - src.mlp_predictor - INFO - Cargando modelo MLP completo desde: C:\Users\balde\Desktop\MAESTRIA MNA\PI\drug-design-diabetes-team29\alphafold_predictor\models\mlp_weights.pth
C:\Users\balde\Desktop\MAESTRIA MNA\PI\drug-design-diabetes-team29\alphafold_predictor\src\mlp_predictor.py:33: FutureWarning: You are using `torch.load` with `weights_only=False` (the current default value), which uses the default pickle module implicitly. It is possible to construct malicious pickle data which will execute arbitrary code during unpickling (See https://github.com/pytorch/pytorch/blob/main/SECURITY.md#untrusted-models)
torch.load(MLP_WEIGHTS_PATH, map_location=DEVICE)
2025-11-03 22:38:02 - src.mlp_predictor - INFO - Modelo MLP cargado exitosamente en cpu
2025-11-03 22:38:02 - src.mlp_predictor - INFO - Arquitectura del modelo: MLP_pLDDT(
  (network): Sequential(
    (0): Linear(in_features=1280, out_features=1280, bias=True)
    (1): ReLU()
    (2): Dropout(p=0.2, inplace=False)
    (3): Linear(in_features=1280, out_features=512, bias=True)
    (4): ReLU()
    (5): Dropout(p=0.2, inplace=False)
    (6): Linear(in_features=512, out_features=254, bias=True)
    (7): ReLU()
    (8): Dropout(p=0.2, inplace=False)
    (9): Linear(in_features=254, out_features=1, bias=True)
  )
)
```

Se continúa con la arquitectura ganadora que toma el embedding:

```
C:\Users\balde\Desktop\MAESTRIA MNA\PI\drug-design-diabetes-team29\alphafold_predictor\src\mlp_predictor.py:33: FutureWarning: You are using `torch.load` with `weights_only=False` (the current default value), which uses the default pickle module implicitly. It is possible to construct malicious pickle data which will execute arbitrary code during unpickling (See https://github.com/pytorch/pytorch/blob/main/SECURITY.md#untrusted-models)
torch.load(MLP_WEIGHTS_PATH, map_location=DEVICE)
2025-11-03 22:38:02 - src.mlp_predictor - INFO - Modelo MLP cargado exitosamente en cpu
2025-11-03 22:38:02 - src.mlp_predictor - INFO - Arquitectura del modelo: MLP_pLDDT(
  (network): Sequential(
    (0): Linear(in_features=1280, out_features=1280, bias=True)
    (1): ReLU()
    (2): Dropout(p=0.2, inplace=False)
    (3): Linear(in_features=1280, out_features=512, bias=True)
    (4): ReLU()
    (5): Dropout(p=0.2, inplace=False)
    (6): Linear(in_features=512, out_features=254, bias=True)
    (7): ReLU()
    (8): Dropout(p=0.2, inplace=False)
    (9): Linear(in_features=254, out_features=1, bias=True)
  )
)
```

Siguiendo el proceso para predicción de cadena única:

```
(7): ReLU()
(8): Dropout(p=0.2, inplace=False)
(9): Linear(in_features=254, out_features=1, bias=True)
)
)
2025-11-03 22:38:02 - src.batch_predictor - INFO - BatchPredictor listo para predicciones
2025-11-03 22:38:02 - __main__ - INFO - Modo: Predicción individual
2025-11-03 22:38:02 - __main__ - INFO - Secuencia: ADEDE
2025-11-03 22:38:02 - __main__ - INFO - =====
2025-11-03 22:38:02 - src.batch_predictor - INFO - Procesando predicción individual para secuencia: ADEDE...
2025-11-03 22:38:02 - src.esm_embedding - INFO - Generando embedding para secuencia de longitud 5
2025-11-03 22:38:07 - src.esm_embedding - INFO - Embedding generado con dimensión: torch.Size([1, 1280])
C:\Users\balde\Desktop\MAESTRIA MNA\PI\drug-design-diabetes-team29\alphafold_predictor\src\batch_predictor.py:24: UserWarning: To copy construct from a tensor, it is recommended to use sourceTensor.clone().detach() or sourceTensor.clone().detach().requires_grad_(True), rather than torch.tensor(sourceTensor).
embedding_tensor = torch.tensor(embedding, dtype=torch.float32)
2025-11-03 22:38:07 - src.batch_predictor - INFO - Predicción exitosa: 44.6744
2025-11-03 22:38:07 - __main__ - INFO - =====
2025-11-03 22:38:07 - __main__ - INFO - PREDICCIÓN EXITOSA
2025-11-03 22:38:07 - __main__ - INFO - =====

2025-11-03 22:38:07 - src.esm_embedding - INFO - Embedding generado con dimensión: torch.Size([1, 1280])
C:\Users\balde\Desktop\MAESTRIA MNA\PI\drug-design-diabetes-team29\alphafold_predictor\src\batch_predictor.py:24: UserWarning: To copy construct from a tensor, it is recommended to use sourceTensor.clone().detach() or sourceTensor.clone().detach().requires_grad_(True), rather than torch.tensor(sourceTensor).
```

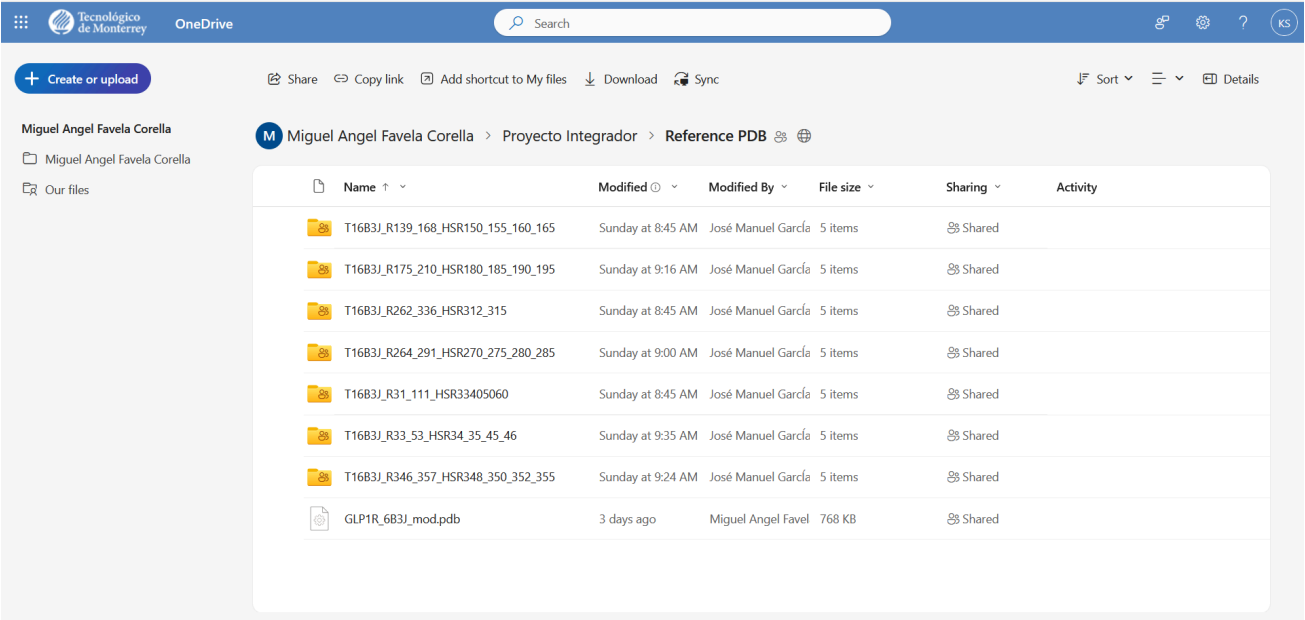
Y el resultado final es la predicción:


```
(9): Linear(in_features=254, out_features=1, bias=True)
)
)
2025-11-03 22:51:16 - src.batch_predictor - INFO - BatchPredictor listo para predicciones
2025-11-03 22:51:16 - __main__ - INFO - Modo: Predicción individual
2025-11-03 22:51:16 - __main__ - INFO - Secuencia: AD
2025-11-03 22:51:16 - __main__ - INFO - -----
2025-11-03 22:51:16 - src.batch_predictor - INFO - Procesando predicción individual para secuencia: AD...
2025-11-03 22:51:16 - src.esm_embedding - INFO - Generando embedding para secuencia de longitud 2
2025-11-03 22:51:19 - src.esm_embedding - INFO - Embedding generado con dimensión: torch.Size([1, 1280])
C:\Users\balde\Desktop\MAESTRIA MNA\PI\drug-design-diabetes-team29\alphafold_predictor\src\batch_predictor.py:24: UserWarning: To copy construct from a tensor, it is recommended to use sourceTensor.clone().detach() or sourceTensor.clone().detach().requires_grad_(True), rather than torch.tensor(sourceTensor).
  embedding_tensor = torch.tensor(embedding, dtype=torch.float32)
2025-11-03 22:51:19 - src.batch_predictor - INFO - Predicción exitosa: 50.8442
2025-11-03 22:51:19 - __main__ - INFO - =====
2025-11-03 22:51:19 - __main__ - INFO - PREDICCIÓN EXITOSA
2025-11-03 22:51:19 - __main__ - INFO - -----

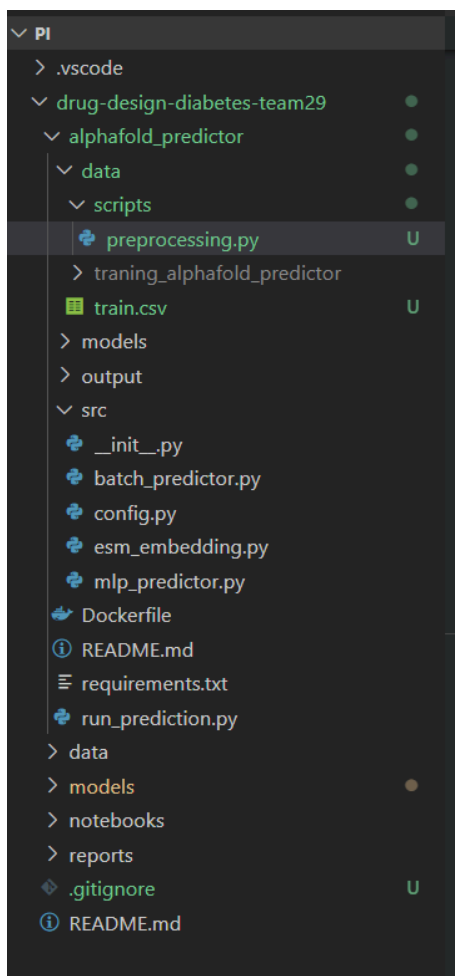
Secuencia: AD
Predicción: 50.844189
Timestamp: 2025-11-03T22:51:19.890144
```

3.3 Ciclo de entrenamiento completo. Desde captura de datos hasta modelo predictor de Alphafold

Primero, recordando que se tuvo que correr el contenedor que es el encargado de generar las muestras, en este caso, todo el primer ciclo de RF-Difussion -> ProteinMPNN -> Alphafold. Se almacenaron cerca de 840 registros primero:



Es importante mencionar que los experimentos y el formato de salida aquí es íntegro, tal cual se obtienen se suben al workspace de almacenamiento. Luego de eso los descargamos y lo ponemos al en la ruta del repositorio del proyecto: /alphafold_predictor/data/traning_alphafold_predictor. Sólo se ponen tal cual las carpetas descargadas:



Luego de eso , se abre el archivo /alphafold_predictor/data/scripts/[processing.py](#) que es el encargado de buscar todos los csv generados, obtener las métricas ipTM y pLDDT y de ahí hacer un df final que servirá para entrenamiento. Resaltando los registros por carpeta encontrados y una sumatoria final para ver los 840 registros totales:

```

Dockerfile preprocessing.py X esm_embedding.py mlp_predictor.py batch_predictor.py run_prediction.py
drug-design-diabetes-team29 > alphafold_predictor > data > scripts > preprocessing.py > ...
27 except KeyError:
28     # Error común: si una de las 'columnas deseadas' no existe
29     print(f" [AVISO] Saltando archivo: {archivo.name}. No tiene todas las columnas: {columns}")
30 except Exception as e:
31     # Captura cualquier otro error de lectura
32     print(f" [ERROR] No se pudo leer el archivo: {archivo.name}. Detalle: {e}")
33
34 #3
35 if list_data:
36     df_final = pd.concat(list_data, ignore_index=True)
37     print(f"\nTotal de registros: {df_final.shape[0]}")
38     final_path = os.path.join(main_path_raw, output_file)
39     df_final.to_csv(final_path, index=False, encoding='utf-8-sig')
40     print(f"\n¡ÉXITO! Se ha creado el archivo consolidado en:")
41     print(final_path)
42     print(f"Total de filas consolidadas: {len(df_final)}")
43 else:
44     print("\nNo se procesó ningún archivo con éxito. No se generó el archivo de salida.")
45
PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS POSTMAN CONSOLE
Se encontraron 7 archivos. Procesando...
[OK] Procesado: mpnn_results.csv con 120 registros
[OK] Procesado: mpnn_results.csv con 120 registros
[OK] Procesado: mpnn_results.csv con 120 registros
[OK] Procesado: mpnn_results.csv con 120 registros
[OK] Procesado: mpnn_results.csv con 120 registros
[OK] Procesado: mpnn_results.csv con 120 registros
[OK] Procesado: mpnn_results.csv con 120 registros
Total de registros: 840
¡ÉXITO! Se ha creado el archivo consolidado en:
C:\Users\balde\Desktop\MAESTRIA MNA\PI\drug-design-diabetes-team29\alphafold_predictor\data\train.csv
Total de filas consolidadas: 840
PS C:\Users\balde\Desktop\MAESTRIA MNA\PI>
```

Y el csv generado va de acuerdo al DataLoader presentado la semana pasada:

```
drug-design-diabetes-team29 > alphafold_predictor > data > train.csv
1  plddt,i_ptm,seq
2  0.5481196641921997,0.0626161694526672,EQLFLFYIITYTVGYALSFSALVIASAILLG/EQLLALLGALLALLA
3  0.5196810364723206,0.0496557764708995,EQLFLFYIITYTVGYALSFSALVIASAILLG/LQLLILLLLILLALLS
4  0.5258078575134277,0.0561077147722244,EQLFLFYIITYTVGYALSFSALVIASAILLG/EQLLLLLLILLLLLS
5  0.5323011875152588,0.0528381243348121,EQLFLFYIITYTVGYALSFSALVIASAILLG/EQLLALLLILLALLA
6  0.5184487700462341,0.0691659227013588,EQLFLFYIITYTVGYALSFSALVIASAILLG/EQLLILGLIVLALLA
7  0.4950191974639892,0.0486248545348644,EQLFLFYIITYTVGYALSFSALVIASAILLG/EQLLILLLIVLALLG
8  0.5387964248657227,0.0614663623273372,EQLFLFYIITYTVGYALSFSALVIASAILLG/EQLLILGLILLALLA
9  0.5226511359214783,0.0764601454138755,EQLFLFYIITYTVGYALSFSALVIASAILLG/EQLLILGLVLVLLLG
10 0.5199462175369263,0.061865746974945,EQLFLFYIITYTVGYALSFSALVIASAILLG/SQLLLLLLIVLALLA
11 0.4993141293525696,0.0545775257050991,EQLFLFYIITYTVGYALSFSALVIASAILLG/EQLLILLLQILLSLLG
12 0.5510651469230652,0.0805610045790672,EQLFLFYIITYTVGYALSFSALVIASAILLG/LGLLILLLLILLLLLS
13 0.4978887438774109,0.0535683296620845,EQLFLFYIITYTVGYALSFSALVIASAILLG/EQLLILLLLVLVLLLL
14 0.5101308226585388,0.0570815689861774,EQLFLFYIITYTVGYALSFSALVIASAILLG/EQLLILLLGIVLALLA
15 0.5365804433822632,0.0639857277274131,EQLFLFYIITYTVGYALSFSALVIASAILLG/EQLLILLLQLLSLLS
16 0.517942488193512,0.0539916343986988,EQLFLFYIITYTVGYALSFSALVIASAILLG/EQLLILLLGILLALLG
17 0.5101308226585388,0.0570815689861774,EQLFLFYIITYTVGYALSFSALVIASAILLG/EQLLILLLGIVLALLA
18 0.5101308226585388,0.0570815689861774,EQLFLFYIITYTVGYALSFSALVIASAILLG/EQLLILLLGIVLALLA
19 0.5169110298156738,0.0566046014428138,EQLFLFYIITYTVGYALSFSALVIASAILLG/EQLLILLLLILGLLLG
20 0.5392673015594482,0.068948820233345,EQLFLFYIITYTVGYALSFSALVIASAILLG/EQLLALLGLVLALLA
```

Luego de eso, abrimos el notebook generate_embeddings.ipynb para ver cómo se comportan comporta ahora el DataLoader con datos de entrenamiento reales, resaltando que si bien tenemos dos cadenas, para el pLDDT formamos una sola cadena:

2. Datos Entrada

```
from custom_dataloader import CustomProteinDataset
from torch.utils.data import DataLoader

CSV_PATH = '../alphafold_predictor/data/train.csv'
BATCH_SIZE = 8
dataset = CustomProteinDataset(CSV_PATH)
data_loader = DataLoader(dataset, batch_size=BATCH_SIZE, shuffle=True)

[5] ✓ 0.5s
... LFLFYIITYTVGYALSFSALVIASAILLG EQLLALLGALLALLA', 'EQLFLFYIITYTVGYALSFSALVIASAILLGLQLLILLLLILLALLS', 'EQLFLFYIITYTVGYALSFSALVIASAILLG EQLLLLLLILLLLLS',
```

Y el batch que servirá de entrada para los embeddings queda:

```
next(iter(data_loader))

[6] ✓ 0.0s
... [('TRNYIHLNLFASFILRALSVFIKDAALKWMYSTAAQFDEALAAILEALARA',
      'SLWETVQKWREYRRQCQRSLTEDPPPATDLFCNRTFDEYACWPDGEPGSFVNVSCPWYLPWASSVPQGHVYRFCTAEGLWLAAEVAALAAAAAAAAA',
      'WETVQKWREYRRQCQRSLTEDWETVRYRRRQREEE',
      'WETVQKWREYRRQCQRSLTEDWEETVRYLRRQREEE',
      'WETVQKWREYRRQCQRSLTEDWETVRYLRRQREEE',
      'KCRLAKSTLTLISKEKMLAKSTLTLI',
      'SLWETVQKWREYRRQCQRSLTEDPPPATDLFCNRTFDEYACWPDGEPGSFVNVSCPWYLPWASSVPQGHVYRFCTAEGLWLSAEVAEELRQLLEELRLAE',
      'KCRLAKSTLTLISKEKERLAKSTLTLI'),
      tensor([[0.5243],
              [0.3976],
              [0.5350],
              [0.6752],
              [0.6201],
              [0.5191],
              [0.5228],
```

Confirmamos dimensiones luego de haber pasado por el embedding y obtenemos los 1280 valores por secuencia

```
#Luego de pasar por la última capa tenemos [num_secuencias, longitud_max_secuencia, dimension_emb]
outputs.last_hidden_state.shape

[10] ✓ 0.0s

... torch.Size([8, 94, 1280])

embedding_por_residuo = outputs.last_hidden_state
embedding_por_secuencia = embedding_por_residuo.mean(dim=1) #Promedio por cada secuencia

[11] ✓ 0.0s

embedding_por_secuencia.shape

[12] ✓ 0.0s

... torch.Size([8, 1280])
```

Y se ejecuta el entrenamiento y la validación para así detectar overfitting:

```
drug-design-diabetes-team29 > models > generate_embeddings.ipynb > Predictor para pLDDT de Alphafold > #MODELO DE MULTILAYER F

Generate + Code + Markdown | Interrupt Restart Clear All Outputs Go To Jupyter Variables Outline ...

total_val_loss += loss.item()

avg_epoch_loss_val = total_val_loss / len(data_loader_val)
print(f"ÉPOCA {epoch + 1} / {EPOCHS} FINALIZADA VAL")
print(f"Pérdida Promedio de la Época VAL: {avg_epoch_loss_val:.4f}")

3m 38.5s

... ÉPOCA 1 / 10 FINALIZADA
Pérdida Promedio de la Época: 0.0137
ÉPOCA 1 / 10 FINALIZADA VAL
Pérdida Promedio de la Época VAL: 0.0023
ÉPOCA 2 / 10 FINALIZADA
Pérdida Promedio de la época: 0.0046
ÉPOCA 2 / 10 FINALIZADA VAL
Pérdida Promedio de la Época VAL: 0.0010
ÉPOCA 3 / 10 FINALIZADA
Pérdida Promedio de la Época: 0.0035
ÉPOCA 3 / 10 FINALIZADA VAL
Pérdida Promedio de la Época VAL: 0.0029
```

Y luego de tener el modelo ganador se pasa a la sección **3.2 Desarrollo de organización del deploy del predictor de Alphafold** para actualizar el contenedor y estar listo para inferencias con docker. Recordando que en un futuro, los especialistas querrán realizar más experimentos para tener diversas escalas de las métricas, por lo que el entrenamiento será necesario para afinar y así tener inferencias más precisas.

3.4 Desarrollo de predictor de Rosetta (Siguiendo Semana)

Durante la siguiente etapa, se desarrollará el modelo predictor asociado a Rosetta, el cual tendrá como objetivo estimar métricas energéticas derivadas de las simulaciones *FastRelax*, tales como energía total, RMSD y estabilidad estructural. Este modelo se entrenará utilizando los valores obtenidos experimentalmente por *PyRosetta*, con el propósito de anticipar el comportamiento energético de nuevas secuencias sin necesidad de ejecutar el pipeline completo.

REPORTE DE DISEÑO COMPUTACIONAL DE PROTEÍNAS CON DOCKER (RFdiffusion, ProteinMPNN y Rosetta)

1.- Estructura de carpetas

Crea una carpeta de trabajo principal, por ejemplo:

```
RFDiffusion_Project/
|
├── Dockerfile
├── PyRosetta4.Release.python310.linux.release-387.tar.bz2
├── requirements.txt
|
├── outputs/                # Carpeta donde se guardarán los resultados
├── 1_run_rfdiffusion.py
├── 2_run_mpnn_af.py
└── 3_run_rosetta.py
```

Ingresa al siguiente enlace:

<https://graylab.jhu.edu/download/PyRosetta4/archive/release/PyRosetta4.Release.python310.linux/>

y descargue el archivo: ***PyRosetta4.Release.python310.linux.release-387***

2. Creación del contenedor Docker

Construir la imagen

Abre CMD o PowerShell en la carpeta del proyecto y ejecuta:

```
C:\Users\Jose Manuel Garcia\Documents\RFDifussion>docker build -t rfdiffusion .
```

Esto creará una imagen llamada ***rfdiffusion*** basada en tu Dockerfile. Incluye dependencias de RFdiffusion, PyRosetta y las librerías necesarias.

```
[*] Building 3.6s (28/28) FINISHED
=> [internal] load build definition from Dockerfile
=> => transferring dockerfile: 3.17kB
=> [internal] load metadata for docker.io/nvidia/cuda:12.4.0-devel-ubuntu22.04
=> [internal] load .dockerignore
=> => transferring context: 2B
=> [ 1/23] FROM docker.io/nvidia/cuda:12.4.0-devel-ubuntu22.04@sha256:28326fed36d43ecf824d7e9call175db639b931e234019b2807f35a8d1551556
=> [internal] load build context
=> => transferring context: 79B
=> CACHED [ 2/23] RUN apt-get update && apt-get install -y python3 python3-pip git wget aria2 unzip && rm -rf /var/lib/apt/lists/*
=> CACHED [ 3/23] RUN pip3 install jedi omegaconf hydra-core icecream pyrsistent pyyaml decorator psutil tqdm numpy scipy networkx requests pillow
=> CACHED [ 4/23] WORKDIR /workspace
=> CACHED [ 5/23] RUN pip3 install torch torchvision torchaudio --index-url https://download.pytorch.org/whl/cu124
=> CACHED [ 6/23] RUN pip3 install jedi omegaconf hydra-core icecream pyrsistent pyyaml decorator psutil
=> CACHED [ 7/23] RUN git clone https://github.com/sokrypton/RFdiffusion.git
=> CACHED [ 8/23] RUN pip3 install --no-dependencies dgl -f https://data.dgl.ai/wheels/torch-2.4/cu124/rep0.html
=> CACHED [ 9/23] RUN pip3 install --no-dependencies einops==0.5.0 opt_einsum==4
=> CACHED [10/23] RUN pip3 install git+https://github.com/NVIDIA/dllogger#egg=dllogger
=> CACHED [11/23] RUN cd RFdiffusion/env/SE3Transformer && pip3 install
=> CACHED [12/23] RUN wget -q https://files.ipd.uw.edu/krypton/anasas && chmod +x anasas
=> CACHED [13/23] RUN pip3 install git+https://github.com/sokrypton/ColabDesign.git@v1.1.1
=> CACHED [14/23] RUN ln -s /usr/local/lib/python3.*/dist-packages/colabdesign colabdesign
=> CACHED [15/23] RUN mkdir -p params RFdiffusion/models
=> CACHED [16/23] RUN aria2c -q -x 16 https://files.ipd.uw.edu/krypton/schedules.zip && unzip -q schedules.zip && rm schedules.zip
=> CACHED [17/23] RUN aria2c -q -x 16 -d RFdiffusion/models http://files.ipd.uw.edu/pub/RFdiffusion/6f5902ac237024bdc176cb93063dch/baseckpt.pt
=> CACHED [18/23] RUN aria2c -q -x 16 -d RFdiffusion/models http://files.ipd.uw.edu/pub/RFdiffusion/e293113f6f1bf1af98f7f9ef944b832bb/complex_baseckpt.pt
=> CACHED [19/23] RUN mkdir -p params && aria2c -q -x 16 https://storage.googleapis.com/alphafold/alphafold_params_2022-12-06.tar && tar -xvf alphafold_params_2022-12-06.tar -C params && rm
=> CACHED [20/23] COPY PyRosetta.Release.python310.Linux.release-387.tar.bz2 /workspace/
=> CACHED [21/23] RUN cd /workspace && tar -xvf PyRosetta.Release.python310.Linux.release-387.tar.bz2
=> CACHED [22/23] RUN cd /workspace/PyRosetta4.Release.python310.Linux.release-387/setup && python3 setup.py install
=> CACHED [23/23] RUN mkdir -p outputs
=> => exporting to image
=> => exporting layers
=> => writing image sha256:bfaa7ced3d19de016d93d3237af8e467b86c647fcec1354ed99c3d97928c1be5
=> => naming to docker.io/library/rfdiffusion
=> =>
```

3. Configuración del entorno virtual

Aunque Docker ya aísla el entorno, se crea un entorno virtual para posteriormente correr nuestros modelos de machine learning en el proceso predictor.

```
python3 -m venv protein_env
```

```
protein_env\Scripts\activate
```

Instalar dependencias:

```
pip install -r requirements.txt
```

4. Ejecución de los scripts

Cada script de Python se ejecutará dentro del contenedor. La idea es montar las carpetas locales de tu PC en /workspace dentro de Docker

4.1 RFdiffusion — Generación de estructuras

```
docker run -it --gpus all -v "C:\Users\Jose Manuel Garcia\Documents\RFDiffussion\outputs":/workspace/outputs -v "C:\Users\Jose Manuel Garcia\Documents\RFDiffussion\1_run_rfdiffusion.py":/workspace/1_run_rfdiffusion.py rfdiffusion python3 1_run_rfdiffusion.py
```

Esto:

- Usa tu GPU (--gpus all)
- Monta tu carpeta local de resultados (outputs)
- Ejecuta el script 1_run_rfdiffusion.py dentro del contenedor.

```
(protein_env) C:\Users\Jose Manuel Garcia\Documents\RFDiffusion>docker run -it --gpus all -v "C:\Users\Jose Manuel Garcia\Documents\RFDiffusion\outputs":/workspace/outputs -v "C:\Users\Jose Manuel Garcia\Documents\RFDiffusion\1_run_rfdiffusion.py":/workspace/1_run_rfdiffusion.py rfdiffusion python3 1_run_rfdiffusion.py

=====
== CUDA ==
=====





CUDA Version 12.4.0

Container image Copyright (c) 2016-2023, NVIDIA CORPORATION & AFFILIATES. All rights reserved.

This container image and its contents are governed by the NVIDIA Deep Learning Container License.
By pulling and using the container, you accept the terms and conditions of this license:
https://developer.nvidia.com/ngc/nvidia-deep-learning-container-license

A copy of this license is made available in this container at /NGC-DL-CONTAINER-LICENSE for your convenience.

/usr/local/lib/python3.10/dist-packages/torch/cuda/__init__.py:61: FutureWarning: The pynvml package is deprecated. Please install nvidia-ml-py instead. If you did not install pynvml directly, please report th
is to the maintainers of the package that installed pynvml for you.
  import pynvml # type: ignore[import]
=====
VERIFICACIÓN GPU:
CUDA disponible: True
GPU: NVIDIA GeForce RTX 3050 Laptop GPU
Memoria: 4.3 GB
=====
📦 Descargando 6B3J desde RCSB...
✅ PDB descargado: /workspace/RFdiffusion/6B3J.pdb
🔗 Usando PDB: /workspace/RFdiffusion/6B3J.pdb
Parameters RFDIFFUSION:
Nombre: test_6B3J
Contigs: 12-15/0 R311-337
PDB: /workspace/RFdiffusion/6B3J.pdb
Iteraciones: 200
Hotspot: R312,R313,R314,R315
Diseños: 1
Simetría: None
Orden simetría: None
Cadenas: None
Visual: image
Comando: python3 RFdiffusion/run_inference.py inference.output_prefix=outputs/test_6B3J contigmap.contigs=[12-15/0 R311-337] inference.num_designs=1 diffuser.T=200 inference.dump_pdb=True inference.dump_pdb_pa
th=/dev/shm inference.input_pdb=/workspace/RFdiffusion/6B3J.pdb ppi.hotspot_res=[R312,R313,R314,R315]
=====
Iniciando diseño de proteína...
```

Name	Date modified	Type	Size
 test_6B3J_0.trb	10/26/2025 7:17 PM	TRB File	38 KB
 test_6B3J_0.pdb	10/26/2025 7:17 PM	Program Debug D...	11 KB
 2025-10-27	10/26/2025 9:13 PM	File folder	
 traj	10/26/2025 7:17 PM	File folder	

4.2 ProteinMPNN — Generación de secuencias

Luego, usa el modelo 3D generado por RFdiffusion:

```
docker run -it --gpus all -v "C:\Users\Jose Manuel Garcia\Documents\RFDiffusion\outputs":/workspace/outputs -v "C:\Users\Jose Manuel Garcia\Documents\RFDiffusion\2_run_mpnn_af.py":/workspace/2_run_mpnn_af.py rfdiffusion python3 2_run_mpnn_af.py
```

```
(protein_env) C:\Users\Jose Manuel Garcia\Documents\RFDiffusion>docker run -it --gpus all -v "C:\Users\Jose Manuel Garcia\Documents\RFDiffusion\outputs":/workspace/outputs -v "C:\Users\Jose Manuel Garcia\Documents\RFDiffusion\2_run_mpnn_af.py":/workspace/2_run_mpnn_af.py rfdiffusion python3 2_run_mpnn_af.py

=====
== CUDA ==
=====

CUDA Version 12.4.0

Container image Copyright (c) 2016-2023, NVIDIA CORPORATION & AFFILIATES. All rights reserved.

This container image and its contents are governed by the NVIDIA Deep Learning Container License.
By pulling and using the container, you accept the terms and conditions of this license:
https://developer.nvidia.com/ngc/nvidia-deep-learning-container-license

A copy of this license is made available in this container at /NGC-DL-CONTAINER-LICENSE for your convenience.

=====
📁 INFORMACIÓN DEL ENTORNO Y ARCHIVOS
=====
📁 Directorio actual: /workspace
=====
📁 ARCHIVOS EN /workspace/outputs/ (VOLUMEN DOCKER):
- 2025-10-27/ [carpeta]
- test_6B3J_0.pdb (10.7 KB)
- test_6B3J_0.trb (37.7 KB)
- traj/ [carpeta]
=====
🔧 PROTEINMPNN + ALPHAFOLD VALIDATION
=====
📁 PARÁMETROS:
PDB: /workspace/outputs/test_6B3J_0.pdb
Output: /workspace/outputs/test_6B3J
Contigs: 12-15 R311-337
Secuencias: 32
Recycles: 1
```



```
Ejecutando ProteinMPNN + AlphaFold...
{'pdb': './workspace/outputs/test_6833_0.pdb', 'loc': './workspace/outputs/test_6833', 'contigs': '12-15R311-337', 'copies': 1, 'num_seqs': 32, 'initial_guess': False, 'use_multimer': False, 'num_recycles': 1, 'rm_aa': 'C', 'num_designs': 1, 'mpnn_sampling_temp': 0.1}
protocol=blinder
WARNING:2825-10-27 03:19:28,623:jax._src.xla_bridge:794: An NVIDIA GPU may be present on this machine, but a CUDA-enabled jaxlib is not installed. Falling back to cpu.
running proteinMPNN...
running AlphaFold...
design:0 n:0 mpnn:1.581 plddt:0.666 i_ptm:0.133 i_pae:13.229 rmsd:31.484 LPILFAIGVNFIFVRVICIVSKLKA/LLLLLLLLLLLLLA
design:0 n:1 mpnn:1.684 plddt:0.659 i_ptm:0.043 i_pae:20.914 rmsd:36.832 LPILFAIGVNFIFVRVICIVSKLKA/SELEELLEELKKA
design:0 n:2 mpnn:1.580 plddt:0.725 i_ptm:0.092 i_pae:16.583 rmsd:35.828 LPILFAIGVNFIFVRVICIVSKLKA/LLLLLLLLLLLLLA
design:0 n:3 mpnn:1.699 plddt:0.666 i_ptm:0.133 i_pae:13.229 rmsd:31.484 LPILFAIGVNFIFVRVICIVSKLKA/LLLLLLLLLLLLLA
design:0 n:4 mpnn:1.678 plddt:0.750 i_ptm:0.097 i_pae:16.324 rmsd:33.675 LPILFAIGVNFIFVRVICIVSKLKA/EELLRLLEELKKA
design:0 n:5 mpnn:1.594 plddt:0.793 i_ptm:0.093 i_pae:15.834 rmsd:35.198 LPILFAIGVNFIFVRVICIVSKLKA/EELLKLEELKKA
design:0 n:6 mpnn:1.587 plddt:0.651 i_ptm:0.102 i_pae:14.706 rmsd:33.287 LPILFAIGVNFIFVRVICIVSKLKA/LLLLITLLLLLAA
design:0 n:7 mpnn:1.537 plddt:0.611 i_ptm:0.121 i_pae:13.943 rmsd:33.899 LPILFAIGVNFIFVRVICIVSKLKA/LLLLITLLLLLAA
design:0 n:8 mpnn:1.594 plddt:0.610 i_ptm:0.096 i_pae:15.893 rmsd:31.681 LPILFAIGVNFIFVRVICIVSKLKA/SLVVLLLLLLLLLA
design:0 n:9 mpnn:1.536 plddt:0.666 i_ptm:0.133 i_pae:13.229 rmsd:31.484 LPILFAIGVNFIFVRVICIVSKLKA/LLLLLLLLLLLLLA
design:0 n:10 mpnn:1.710 plddt:0.809 i_ptm:0.071 i_pae:17.854 rmsd:36.345 LPILFAIGVNFIFVRVICIVSKLKA/SSLEKLEELKKA
design:0 n:11 mpnn:1.688 plddt:0.789 i_ptm:0.160 i_pae:12.854 rmsd:29.377 LPILFAIGVNFIFVRVICIVSKLKA/LELLLLLLLLLLLA
design:0 n:12 mpnn:1.611 plddt:0.674 i_ptm:0.101 i_pae:15.551 rmsd:33.559 LPILFAIGVNFIFVRVICIVSKLKA/ELLITLLLLLLLA
design:0 n:13 mpnn:1.684 plddt:0.702 i_ptm:0.126 i_pae:14.403 rmsd:35.977 LPILFAIGVNFIFVRVICIVSKLKA/ELLEKLEELKKA
design:0 n:14 mpnn:1.556 plddt:0.707 i_ptm:0.110 i_pae:14.700 rmsd:28.871 LPILFAIGVNFIFVRVICIVSKLKA/GLLLLLLLLLLLAA
design:0 n:15 mpnn:1.522 plddt:0.666 i_ptm:0.133 i_pae:13.229 rmsd:31.484 LPILFAIGVNFIFVRVICIVSKLKA/LLLLLLLLLLLLLA
design:0 n:16 mpnn:1.581 plddt:0.672 i_ptm:0.140 i_pae:13.353 rmsd:26.474 LPILFAIGVNFIFVRVICIVSKLKA/SLLLLLLTLLLLLA
design:0 n:17 mpnn:1.656 plddt:0.599 i_ptm:0.111 i_pae:15.418 rmsd:31.339 LPILFAIGVNFIFVRVICIVSKLKA/SLLLLLLLLLLLLA
design:0 n:18 mpnn:1.628 plddt:0.697 i_ptm:0.102 i_pae:15.168 rmsd:31.262 LPILFAIGVNFIFVRVICIVSKLKA/GLLLVLLLLLLLA
design:0 n:19 mpnn:1.599 plddt:0.659 i_ptm:0.138 i_pae:13.049 rmsd:33.258 LPILFAIGVNFIFVRVICIVSKLKA/LLLLTLLLLLLLLA
design:0 n:20 mpnn:1.632 plddt:0.773 i_ptm:0.088 i_pae:16.886 rmsd:39.189 LPILFAIGVNFIFVRVICIVSKLKA/SLLELELELLKKA
design:0 n:21 mpnn:1.597 plddt:0.666 i_ptm:0.133 i_pae:13.229 rmsd:31.484 LPILFAIGVNFIFVRVICIVSKLKA/LLLLLLLLLLLLLA
design:0 n:22 mpnn:1.558 plddt:0.697 i_ptm:0.106 i_pae:15.040 rmsd:30.557 LPILFAIGVNFIFVRVICIVSKLKA/GLLLLLLLLLLLAA
design:0 n:23 mpnn:1.588 plddt:0.688 i_ptm:0.124 i_pae:13.714 rmsd:31.567 LPILFAIGVNFIFVRVICIVSKLKA/LLLLLLLLLLLLLA
design:0 n:24 mpnn:1.525 plddt:0.659 i_ptm:0.138 i_pae:13.049 rmsd:33.258 LPILFAIGVNFIFVRVICIVSKLKA/LLLLLLLLLLLLLA
design:0 n:25 mpnn:1.670 plddt:0.672 i_ptm:0.121 i_pae:14.521 rmsd:30.486 LPILFAIGVNFIFVRVICIVSKLKA/SLLLITLLLLLLLA
design:0 n:26 mpnn:1.704 plddt:0.788 i_ptm:0.071 i_pae:17.779 rmsd:41.076 LPILFAIGVNFIFVRVICIVSKLKA/SELLKKLEELKKA
design:0 n:27 mpnn:1.600 plddt:0.713 i_ptm:0.108 i_pae:14.694 rmsd:34.528 LPILFAIGVNFIFVRVICIVSKLKA/GLLLLLLLLLLLLA
design:0 n:28 mpnn:1.516 plddt:0.666 i_ptm:0.133 i_pae:13.229 rmsd:31.484 LPILFAIGVNFIFVRVICIVSKLKA/LLLLLLLLLLLLLA
design:0 n:29 mpnn:1.595 plddt:0.685 i_ptm:0.122 i_pae:13.922 rmsd:33.482 LPILFAIGVNFIFVRVICIVSKLKA/LLVLLLLLLLLLA
design:0 n:30 mpnn:1.665 plddt:0.828 i_ptm:0.064 i_pae:19.175 rmsd:33.972 LPILFAIGVNFIFVRVICIVSKLKA/EELLKLEELKKA
design:0 n:31 mpnn:1.604 plddt:0.627 i_ptm:0.112 i_pae:15.112 rmsd:34.332 LPILFAIGVNFIFVRVICIVSKLKA/GLLLLLLLLLLLLA
✓ Validación completada en 190.5s
ARCHIVOS GENERADOS:
✓ mpnn_results.csv (4.6 KB)
✓ best.pdb (25.1 KB)
✓ best_design.pdb (25.0 KB)
✓ design.fasta (3.7 KB)
=====
TIEMPO TOTAL DE EJECUCIÓN: 190.56 segundos
```

ProteinMPNN: genera secuencias de aminoácidos que se ajustan a la estructura 3D dada.

AlphaFold: predice cómo se doblaría esa secuencia, y sirve para verificar si la secuencia propuesta realmente adopta la forma deseada.

4.3 Rosetta — Refinamiento estructural

Finalmente, ejecuta el tercer script dentro del mismo contenedor:

```
docker run -it --gpus all -v "C:\Users\Jose Manuel Garcia\Documents\RFDifussion\outputs":/workspace/outputs -v "C:\Users\Jose Manuel Garcia\Documents\RFDifussion\3_run_rosetta.py":/workspace/3_run_rosetta.py rfdiffusion python3 3_run_rosetta.py
```

```
(protein_env) C:\Users\Jose Manuel Garcia\Documents\RFDifussion>docker run -it --gpus all -v "C:\Users\Jose Manuel Garcia\Documents\RFDifussion\outputs":/workspace/outputs -v "C:\Users\Jose Manuel Garcia\Documents\RFDifussion\3_run_rosetta.py":/workspace/3_run_rosetta.py rfdiffusion python3 3_run_rosetta.py

=====
== CUDA ==
=====

CUDA Version 12.4.0

Container image Copyright (c) 2016-2023, NVIDIA CORPORATION & AFFILIATES. All rights reserved.

This container image and its contents are governed by the NVIDIA Deep Learning Container License.
By pulling and using the container, you accept the terms and conditions of this license:
https://developer.nvidia.com/ngc/nvidia-deep-learning-container-license

A copy of this license is made available in this container at /NGC-DL-CONTAINER-LICENSE for your convenience.

PyRosetta-4
Created in JHU by Sergey Lyckov and PyRosetta Team
(C) Copyright Rosetta Commons Member Institutions

NOTE: USE OF PyRosetta FOR COMMERCIAL PURPOSES REQUIRE PURCHASE OF A LICENSE
See LICENSE.PyRosetta.md or email license@um.edu for details

PyRosetta-4 2024 [Rosetta PyRosetta4.Release.python310.linux.2024.39+release.59628fbc5bc09f1221e1642f1f8d157ce49b1410.2024-09-23T07:49:48] retrieved from: http://www.pyrosetta.org
● INICIANDO ANÁLISIS ROSETTA COMPLETO

● Encontrados 32 archivos PDB
=====
➤ ANÁLISIS ROSETTA COMPLETO: design0_n0.pdb
=====

■ INFORMACIÓN BÁSICA:
Residuos: 41
Energía REFI: 190.65 REU
Energía/residuo: 4.65 REU/res

● DESGLOSE ENERGÉTICO:
fa_atr      : -177.98 REU
fa_rep      : 159.15 REU
fa_sol      : 131.09 REU
fa_elec     : -40.28 REU
hbond_sc    : 0.00 REU
hbond_bb_sc : 0.00 REU
omega       : 9.46 REU
rama_prepro : -4.99 REU
p_aa_pp     : -6.12 REU
```


🏆 REPORTE COMPARATIVO – MEJORES DISEÑOS				
Archivo	Residues	Energy/Res	RMSD	Estado
1. design0_n10.pdb	41	3.04	8.74	PROBLEMA
2. design0_n2.pdb	41	3.42	3.14	PROBLEMA
3. design0_n1.pdb	41	3.76	4.50	PROBLEMA
4. design0_n29.pdb	41	3.82	2.41	PROBLEMA
5. design0_n5.pdb	41	4.38	2.58	PROBLEMA
6. design0_n0.pdb	41	4.65	1.86	PROBLEMA
7. design0_n15.pdb	41	4.65	1.95	PROBLEMA
8. design0_n21.pdb	41	4.65	1.96	PROBLEMA
9. design0_n28.pdb	41	4.65	2.03	PROBLEMA
10. design0_n3.pdb	41	4.65	1.96	PROBLEMA
11. design0_n9.pdb	41	4.65	2.03	PROBLEMA
12. design0_n23.pdb	41	4.69	2.01	PROBLEMA
13. design0_n18.pdb	41	5.15	2.85	PROBLEMA
14. design0_n27.pdb	41	5.20	2.04	PROBLEMA
15. design0_n14.pdb	41	5.38	3.67	PROBLEMA
16. design0_n22.pdb	41	5.51	3.67	PROBLEMA
17. design0_n4.pdb	41	5.63	2.58	PROBLEMA
18. design0_n20.pdb	41	5.81	3.73	PROBLEMA
19. design0_n30.pdb	41	5.98	6.42	PROBLEMA
20. design0_n11.pdb	41	6.04	2.56	PROBLEMA
21. design0_n13.pdb	41	6.17	2.48	PROBLEMA
22. design0_n26.pdb	41	6.45	3.84	PROBLEMA
23. design0_n25.pdb	41	6.53	1.85	PROBLEMA
24. design0_n12.pdb	41	7.48	2.52	PROBLEMA
25. design0_n31.pdb	41	8.69	4.30	PROBLEMA
26. design0_n8.pdb	41	11.24	1.41	PROBLEMA
27. design0_n6.pdb	41	14.78	1.77	PROBLEMA
28. design0_n17.pdb	41	21.29	2.16	PROBLEMA
29. design0_n19.pdb	41	23.44	1.43	PROBLEMA
30. design0_n24.pdb	41	23.44	1.43	PROBLEMA
31. design0_n7.pdb	41	38.33	1.72	PROBLEMA
32. design0_n16.pdb	41	48.11	2.24	PROBLEMA

3. Conclusiones

- Se logró implementar un flujo de trabajo reproducible y aislado mediante Docker.
- RFdiffusion permitió generar geometrías 3D plausibles de proteínas según parámetros definidos (contigs, hotspots).
- ProteinMPNN generó secuencias compatibles con las estructuras diseñadas.
- Rosetta se empleó para evaluar y optimizar la estabilidad estructural.
- PyMOL permitió visualizar y verificar las partes clave del diseño.

Bibliografía

- Jumper, J., Evans, R., Pritzel, A., et al. (2021). *Highly accurate protein structure prediction with AlphaFold*. Nature, 596(7873), 583–589.
- Watson, J. L., Juergens, D., Bennett, N. R., et al. (2023). *De novo design of protein structure and function with RFdiffusion*. bioRxiv. <https://doi.org/10.1101/2023.03.20.533421>
- Anand, N., & Achim, T. (2022). *ProteinMPNN: A deep learning model for protein sequence design*. bioRxiv. <https://doi.org/10.1101/2022.07.20.500902>

