Integrantes que aportaron al entregable 1:

- Chalco Adrián
- Hurtado Benjamin

Estado del arte

La predicción cuantitativa de la afinidad antígeno-anticuerpo a partir de secuencias proteicas ha experimentado avances significativos gracias a modelos de inteligencia artificial que prescinden de información estructural tridimensional. Entre los enfoques puramente basados en secuencia, DG-Affinity (Yuan et al., 2023) emplea embeddings de lenguaje entrenados sobre corpus proteicos generales para convertir antígenos y anticuerpos en vectores de alta dimensión que luego alimentan un regresor ConvNeXt, alcanzando una correlación de Pearson superior a 0.65 en conjuntos de prueba independientes . De manera similar, AttABseq (Jin et al., 2024) implementa una arquitectura híbrida de tres CNN seguidas de bloques de atención multi-cabeza que predicen cambios de afinidad (ΔΔG) en mutantes de anticuerpos, logrando coeficientes de correlación de aproximadamente 0.5 en múltiples datasets de mutaciones puntuales y múltiples, lo que supera a métodos clásicos de física computacional como FoldX o BeAtMuSiC . Por su parte, AbAgIntPre (Huan et al., 2022) se basa en una red siamés de tipo CNN para clasificar interacciones anticuerpo-antígeno, alcanzando un AUC de 0.82 en la identificación binaria de interacciones, lo que demuestra el potencial de los modelos sólo secuencia para tareas de detección .

En el campo de los modelos generativos, diversos estudios han explorado arquitecturas difusivas, de flujo y de transformers multimodales. Uçar et al. (2024) establecieron que la log-verosimilitud de modelos generativos correlaciona de forma significativa con la afinidad experimental (Kd) y que un modelo difusivo entrenado en un gran dataset sintético puede mejorar sustancialmente la predicción de Kd . En paralelo, Luo et al. (2022) presentaron un modelo difusivo co-diseñador de secuencia y estructura de Regiones Determinantes de Complementariedad (CDR), obteniendo resultados competitivos en afinidad medida por funciones energéticas de Rosetta sin depender de estructuras cristalográficas . Asimismo, AntiFold (Gao et al., 2024), una variante de inverse folding basada en ESM-IF1, demostró correlaciones “zero-shot” de ~0.42 entre log-verosimilitud de secuencia y afinidad medida en escaneos masivos mutacionales, superando a ProtMPNN y AbMPNN en recuperación de variantes de alta afinidad . Más recientemente, AffinityFlow (Chen et al., 2025) combinó un difusor estructural estilo AlphaFold con un predictor energético entrenado sobre datos de enlaces, mostrando mejoras de afinidad en anticuerpos contra SARS-CoV-2 según estimaciones de Rosetta ΔG . Finalmente, el estudio de diseño de novo de Shanehsazzadeh et al. (2023) validó experimentalmente que un millón de variantes de CDR generadas por un modelo generativo a gran escala pudieron reducir Kd por debajo de 10 nM en 71 candidatos, rivalizando con terapias establecidas como trastuzumab .

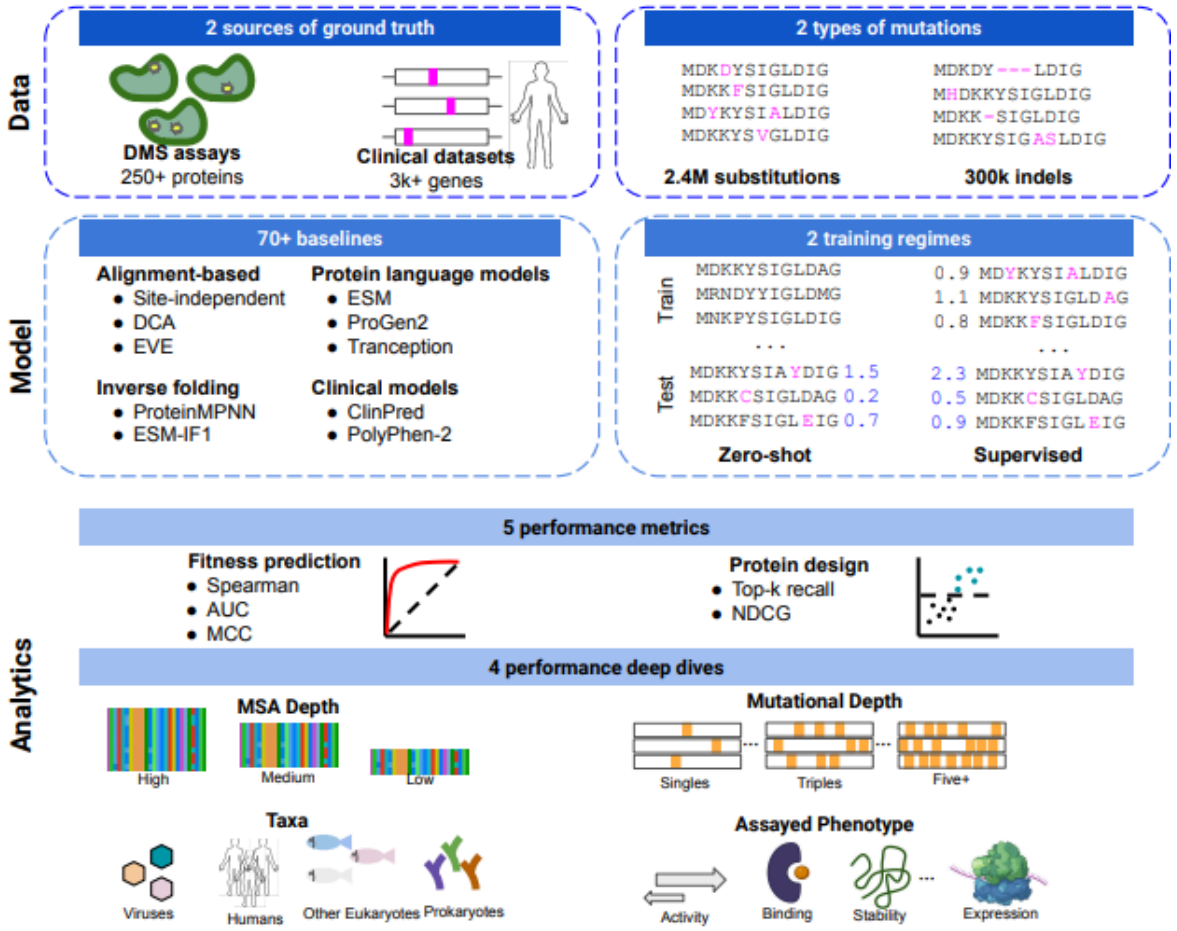
A pesar de estos logros, persiste la necesidad de un método que prediga directamente valores continuos de Kd o ΔG desde secuencias sin recurrir a datos estructurales o a benchmarks limitados, lo que motiva el desarrollo de una solución basada en transferencia de aprendizaje y modelos generativos finamente ajustados.

ESM3 (Hayes et al., 2025) es un modelo de lenguaje generativo y multimodal diseñado para razonar sobre la secuencia, estructura y función de las proteínas, aprendiendo de tokens discretos que representan cada modalidad. Se distingue de otros modelos por su gran escala, alcanzando los 98 mil millones de parámetros, lo que resulta en mejoras en la representación de estas propiedades y en sus capacidades generativas. ESM3 ha sido aplicado en la generación controlada de proteínas siguiendo prompts complejos que

combinan sus modalidades, la simulación de la evolución, el diseño racional de proteínas, la predicción de estructura, e incluso la generación de una proteína funcional novedosa como una proteína verde fluorescente (GFP). Esta herramienta puede ser prometedora para predecir la afinidad de secuencias proteicas de anticuerpos a sus antígenos porque su capacidad para razonar sobre la estructura tridimensional y las características funcionales, incluyendo sitios de unión (ligand binding sites), le permite ir más allá de la simple secuencia y capturar las complejas interacciones moleculares relevantes para el binding

En el contexto de la comparación de diferentes herramientas de predicción, ProteinGym (Notin et al., 2023) es un amplio conjunto de benchmarks diseñado específicamente para la predicción de la aptitud (fitness) y el diseño de proteínas. Fue creado para abordar las dificultades de evaluar modelos de aprendizaje automático para estas tareas debido a la diversidad de conjuntos de datos y la variabilidad en el rendimiento de los modelos. ProteinGym contiene varios tipos de bases de datos; en particular, incluye una vasta colección de más de 250 ensayos estandarizados de escaneo mutacional profundo (Deep Mutational Scanning - DMS), que abarcan millones de secuencias mutadas en más de 200 familias de proteínas con diversas funciones y taxones. También incorpora datasets clínicos curados que ofrecen anotaciones de alta calidad de expertos sobre los efectos de mutaciones en genes humanos, incluyendo tanto sustituciones como inserciones o deleciones (indels). La utilidad de ProteinGym para evaluar el desempeño del fine-tuning de un modelo de lenguaje, como podría ser ESM3, según otras fuentes, reside en su marco de evaluación estructurado que incluye métricas apropiadas para la predicción de fitness y el diseño;el cual, precisamente, soporta evaluaciones en configuraciones supervisadas. El fine-tuning es ampliamente usado para entrenamiento supervisado, por lo que los benchmarks supervisados de ProteinGym son directamente relevantes para evaluar cómo un modelo, una vez ajustado con datos específicos de fitness, es capaz de predecir los efectos de mutaciones, incluso extrapolando a posiciones no vistas durante el entrenamiento o a nuevas familias de proteínas, dependiendo de cómo se dividan los datos para la validación cruzada. Los datasets y el código base están disponibles públicamente y ProteinGym permite comparar una variedad de modelos, incluidos los modelos de lenguaje de proteínas, en estas tareas supervisadas utilizando métricas como la correlación de Spearman.

Cabe resaltar que la existencia de benchmarks como ProteinGym que evalúan el desempeño en tareas de predicción de fitness, incluyendo ensayos de binding, sugiere que ESM3 puede ser evaluado en este tipo de tareas supervisadas mediante transfer learning, lo cual es relevante para predecir la afinidad anticuerpo-antígeno.



Vacío de Conocimiento

Aún no se ha evaluado específicamente el uso de los embeddings de ESM3 para predecir los efectos mutacionales en la afinidad de anticuerpos

Objetivo General

Desarrollar un modelo de transfer learning basado en ESM3 para predecir cuantitativamente la afinidad (Kd/ΔG) de interacciones antígeno-anticuerpo directamente a partir de secuencias proteicas del anticuerpo CR9114 a unión del anígeno H1 (hemaglutinina).

Hipótesis

Los embeddings generados por ESM3 capturan patrones de unión relevantes y permiten, tras un ajuste fino supervisado, predecir valores de afinidad para CR9114 con correlación significativa frente a datos experimentales.

Bases de datos

Para el preentrenamiento multimodal se recurrirá a los 2.78 mil millones de secuencias naturales y sus 236 millones de estructuras anotadas que componen el corpus de ESM3, incluyendo fuentes como UniRef, MGnify, JGI, OAS, PDB, AlphaFoldDB y ESMAtlas . Además, para la fase de ajuste fino se empleará un conjunto de datos específico de mutagénesis de anticuerpos: más de **65 000 variantes** del anticuerpo CR9114 dirigidas contra hemaglutininas virales (subtipos H1, H3 y H9), cada una con su Kd experimental . Como fuentes complementarias se considerarán **SAbDab** y **PDBbind** para ampliar pares anticuerpo-antígeno con mediciones de Kd o ΔG . ProteinGym y UniProt seguirán siendo esenciales para diversificar el entrenamiento y validar la generalización . La base de datos de ProteinGym servirá como base de trabajo por su estructura de análisis de benchmarks. Por su parte, UniProt, con más de 227 millones de secuencias anotadas de proteínas, incluye reconstrucciones detalladas de inmunoglobulinas y antígenos clínicamente relevantes (The UniProt Consortium, 2023) . La riqueza y diversidad de estas bases de datos permiten tanto el pre-entrenamiento de modelos generales de proteínas como la adaptación específica a la predicción de afinidad antígeno-anticuerpo.

Fundamentos teóricos

El proyecto se sustenta en dos pilares conceptuales. Primero, la transferencia de aprendizaje (Transfer Learning) sobre modelos de lenguaje de proteínas—como ESM-3 o ProGen—permite explotar patrones evolutivos aprendidos en corpus de secuencias generales y luego afinar estos modelos para tareas específicas de afinidad antígeno-anticuerpo, tal como ha demostrado ThermoMPNN para estabilidad proteica . Segundo, los modelos generativos basados en arquitecturas de transformers, flujos y difusores capturan la distribución subyacente de secuencias funcionales, posibilitando la generación de variantes plausibles y la estimación indirecta de afinidad mediante log-verosimilitud o métricas energéticas. La representación de proteínas como secuencias de tokens de aminoácidos facilita la aplicación de técnicas de NLP, donde los embeddings resultantes reflejan contexto biológico y estructural sin necesidad de resolver la estructura tridimensional. Estos fundamentos abordan la limitación de los métodos tradicionales de docking y química computacional, que dependen de estructuras precisas y no escalan bien a grandes bibliotecas de mutantes.

Metodología

El flujo de trabajo propuesto integra secuencia y predicción IA en un esquema de cuatro etapas: en primer lugar, se recopilarán secuencias de anticuerpos con afinidades proxy extraídas de ProteinGym junto con secuencias de antígenos representativas de UniProt. A continuación, se empleará un modelo de lenguaje de proteínas pre-entrenado (p. ej., ESM-3) para extraer embeddings de cada secuencia, los cuales se concatenarán a nivel de anticuerpo y antígeno. En la tercera fase, estos vectores alimentarán un regresor—por ejemplo, una red neuronal CNN/MLP—entrenado para minimizar la diferencia entre la afinidad predicha (Kd o ΔG) y los valores experimentales proxy. Finalmente, se evaluará el desempeño mediante validez cruzada, cálculo de correlación de Pearson y error absoluto medio (MAE), complementado con validación externa usando datasets no vistos. Para profundizar en la interpretabilidad, se analizarán los pesos de atención y la importancia de cada posición de CDR, lo que permitirá visualizar mapas de atención y logos de secuencia que resalten residuos críticos.

El flujo de trabajo se articula en cuatro etapas integradas:

1. **Recopilación y curación** de datos:
    - Extracción de las 65 000 secuencias de CR9114 con Kd medida y de datasets de SAbDab/PDBbind.
    - Incorporación de embeddings de ProteinGym/UniProt para enriquecer la representación.
2. **Extracción de embeddings** con ESM3:
    - Preentrenamiento multimodal y generación de vectores continuos que capturan contexto secuencial, estructural y funcional.
3. **Zero-shot:**
    - Uso directo de los embeddings intrínsecos de ESM3 para predecir valores de Kd de las secuencias mutantes de CR9114
    - Evaluación de la capacidad intrínseca del modelo preentrenado
4. **Ajuste fino (Transfer Learning)**:
    - Entrenamiento de una capa predictiva (CNN/MLP) que reciba como entrada la concatenación de embeddings de anticuerpo y antígeno.
    - Búsqueda de hiperparámetros optimizando la correlación Pearson y minimizando MAE.
5. **Validación y evaluación**:
    - Splits de validación basados en protocolos de ProteinGym :
        - *Random-wise :* Cada mutación en el conjunto de datos se asigna aleatoriamente a uno de los cinco folds (divisiones) diferentes utilizados para la validación cruzada
        - *Contiguous-wise*: Implica dividir la secuencia de la proteína en segmentos contiguos a lo largo de su longitud, las mutaciones se asignan a cada fold dependiendo de la posición en la que ocurren. Se usa para casos de mutaciones simples de un nucleótido.
        - *Modulo-wise* : Las posiciones se asignan a cada fold utilizando el operador módulo del número de posición por el número total de folds. Se usa para casos de mutaciones simples de un nucleótido.
    - Validación cruzada y test independiente para medir rendimiento realista

Resultados y métricas

Para cuantificar la eficacia del modelo se utilizarán:

- **Correlación de Pearson (r):** se espera $r > 0.6$ en test independiente.
- **Error medio absoluto** (MAE) o **RMSE normalizado** (NDCG RMSE): objetivo < 1 kcal/mol en ΔG.
- **Top-K candidates:** evaluados según su NDCG y r, para priorizar anticuerpos con

mayor afinidad .

Estas métricas asegurarán que el prototipo no solo aprende patrones generales, sino que identifica variantes con afinidad real y relevante para aplicaciones biotecnológicas.

Innovación tecnológica

El desarrollo de un prototipo bioinformático que infiera afinidad antígeno-anticuerpo directamente desde la secuencia representa una innovación disruptiva al eliminar la dependencia de métodos de cristalografía y ensayos in vitro iniciales. Esta herramienta permitirá un screening automatizado y escalable de bibliotecas de anticuerpos, acelerando el diseño de candidatos terapéuticos con afinidad optimizada y la generación de kits diagnósticos rápidos para patógenos emergentes. La integración de modelos generativos y de transferencia de aprendizaje ofrece una ventaja competitiva al combinar la automatización de generación de variantes con la precisión de predicción cuantitativa, posibilitando actualizaciones continuas del prototipo con nuevos datos experimentales.

Reconocimiento de patrones

Los modelos de lenguaje y de atención descubren automáticamente motivos funcionales en las regiones CDRs de anticuerpos que correlacionan con alta afinidad. Por ejemplo, AttABseq (Jin et al., 2024) permite identificar a nivel residuo las mutaciones que más influyen en la constante de disociación, generando mapas de atención que destacan posiciones fundamentales para la unión . De forma análoga, los difusores secuencia-estructura y los modelos de flujo capturan patrones locales y globales que pueden representarse mediante logos de secuencia o mapas de calor, proporcionando insight sobre la combinación de residuos que optimizan ΔG. Este reconocimiento de patrones no solo guía la interpretación de resultados, sino que también alimenta iteraciones de diseño generativo para explorar regiones inexploradas del espacio de secuencias.

Conclusiones

La combinación de enfoques basados en secuencia y modelos generativos finamente ajustados llena un vacío crítico en la predicción de afinidad antígeno-anticuerpo al prescindir de datos estructurales y ofrecer estimaciones cuantitativas directas de Kd o ΔG. El uso de bases de datos como ProteinGym y UniProt garantiza un entrenamiento sólido y representativo, mientras que los fundamentos de transferencia de aprendizaje y generación de secuencias respaldan la viabilidad técnica. El prototipo bioinformático propuesto promete acelerar el desarrollo de anticuerpos terapéuticos y diagnósticos, al tiempo que aporta interpretabilidad a través de mapas de atención y logos de secuencia. Los próximos pasos incluirán la incorporación de nuevos ensayos de DMS específicos de anticuerpos y la validación experimental in vitro de las variantes de mayor afinidad generadas.

Referencias

- Yuan, Y., Chen, Q., Mao, J., Li, G., & Pan, X. (2023). DG-Affinity: Predicting antigen–antibody affinity with language models from sequences. *BMC Bioinformatics, 24*, Article 430. https://doi.org/10.1186/s12859-023-05562-z
- Jin, R., Ye, Q., Wang, J., Cao, Z., Jiang, D., Wang, T., Kang, Y., Xu, W., Hsieh, C.-Y., & Hou, T. (2024). AttABseq: An attention-based deep learning prediction method for antigen–antibody binding affinity changes based on protein sequences. *Briefings in*

Bioinformatics, 25(4), bbae304. https://doi.org/10.1093/bib/bbae304

- Huang, Y., Zhang, Z., & Zhou, Y. (2022). AbAgIntPre: A deep learning method for predicting antibody–antigen interactions based on sequence information. *Frontiers in Immunology, 13*, 1053617. https://doi.org/10.3389/fimmu.2022.1053617
- Notin, P., Kollasch, A. W., Ritter, D., van Niekerk, L., Paul, S., Spinner, H., Rollins, N., Shaw, A., Orenbuch, R., Weitzman, R., Frazer, J., Dias, M., Franceschi, D., Gal, Y., & Marks, D. S. (2023). ProteinGym: Large-scale benchmarks for protein design and fitness prediction [Preprint]. *bioRxiv*. https://doi.org/10.1101/2023.12.07.570727
- The UniProt Consortium. (2023). UniProt: The universal protein knowledgebase in 2023. *Nucleic Acids Research, 51*(D1), D523–D531. https://doi.org/10.1093/nar/gkac1052
- Uçar, T., Malherbe, C., & Gonzalez, F. (2024). Benchmarking generative models for antibody design & exploring log-likelihood for sequence ranking [Preprint]. *bioRxiv*. https://doi.org/10.1101/2024.10.07.617023
- Luo, S., Su, Y., Peng, X., Wang, S., Peng, J., & Ma, J. (2022). Antigen-specific antibody design and optimization with diffusion-based generative models for protein structures. In A. H. Oh, A. Agarwal, D. Belgrave, & K. Cho (Eds.), *Advances in Neural Information Processing Systems 35* (NeurIPS 2022). Curran Associates. https://openreview.net/forum?id=jSorGn2Tjg
- Høie, M. H., Hummer, A. M., Olsen, T. H., Aguilar-Sanjuan, B., Nielsen, M., & Deane, C. M. (2025). AntiFold: Improved antibody structure design using inverse folding. *Bioinformatics Advances, 5*(1), vbae202. https://doi.org/10.1093/bioadv/vbae202
- Hayes, T., Rao, R., Akin, H., Sofroniew, N. J., Oktay, D., Lin, Z., Verkuil, R., Tran, V. Q., Deaton, J., Wiggert, M., Badkundri, R., Shafkat, I., Gong, J., Derry, A., Molina, R. S., Thomas, N., Khan, Y., Mishra, C., Kim, C., … Rives, A. (2025). *Simulating 500 million years of evolution with a language model*. *Science, 387*(6736), 850–858. https://doi.org/10.1126/science.ads0018
- Chen, C., Herpoldt, K.-L., Zhao, C., Wang, Z., Collins, M., Shang, S., & Benson, R. (2025). AffinityFlow: Guided flows for antibody affinity maturation [Preprint]. *arXiv*. https://doi.org/10.48550/arXiv.2502.10365
- Shanehsazzadeh, A., Bachas, S., McPartlon, M., et al. (2023). Unlocking de novo antibody design with generative artificial intelligence [Preprint]. *bioRxiv*. https://doi.org/10.1101/2023.01.08.523187
- Dunbar, J., Krawczyk, K., Leem, J., Baker, T., Fuchs, A., Georges, G., Shi, J., & Deane, C. M. (2014). SAbDab: The structural antibody database. *Nucleic Acids Research, 42*(D1), D1140–D1146. https://doi.org/10.1093/nar/gkt1043
- Liu, Z., Li, Y., Han, L., Li, J., Liu, J., Zhao, Z., & Wang, R. (2015). PDB-wide collection of binding data: Current status of the PDBbind database. *Bioinformatics, 31*(3), 405–412. https://doi.org/10.1093/bioinformatics/btu626
- Phillips, A. M., Lawrence, K. R., Moulana, A., Dupic, T., Chang, J., Johnson, M. S., Cvijovic, I., Mora, T., Walczak, A. M., & Desai, M. M. (2021). Binding affinity landscapes constrain the evolution of broadly neutralizing anti-influenza antibodies. *eLife, 10*, e71393. https://doi.org/10.7554/eLife.71393