Modelos de inteligencia artificial para predecir la actividad biológica de compuestos químicos contra Leishmania donovani

Hellen Aguilar Noguera

Jose Leonardo Araya Parajeles

Universidad CENFOTEC

Proyecto de Investigación Aplicada 1

Miguel Perez Montero

Fecha: Noviembre, 2025

Resumen ejecutivo *(Abstract)*

Este proyecto de investigación aplicada tiene como objetivo desarrollar un prototipo funcional que priorice compuestos benzimidazólicos con potencial actividad anti-Leishmania donovani mediante la combinación de similitud estructural (coeficiente de Tanimoto sobre fingerprints ECFP4) y modelos supervisados de inteligencia artificial (Random Forest como línea base, con exploración opcional de VAE + MLP). Dado que se trata de un PIA-01, el trabajo se enfoca en la fase de diseño metodológico y no incluye resultados empíricos completos. No obstante, se anticipa que el enfoque propuesto permitirá reducir significativamente el espacio de búsqueda en cribado virtual, mejorando la eficiencia en la selección de candidatos para ensayos experimentales. Entre los principales obstáculos identificados destacan la heterogeneidad y escasez de datos bioactivos confiables en bases públicas, las limitaciones computacionales para entrenar modelos profundos, y la conocida desconexión entre similitud estructural y actividad biológica (activity cliffs). A pesar de estas limitaciones, el marco metodológico propuesto es viable, reproducible y escalable, y sienta las bases para futuras validaciones experimentales en PIA-02.

Palabras clave: leishmaniasis, benzimidazoles, quimioinformática, inteligencia artificial, similitud molecular, Tanimoto, Random Forest, VAE, descubrimiento de fármacos, ChEMBL

Tabla de Contenido

Capítulo 1. Introducción

Esta Investigación se enfoca en desarrollar un sistema que priorice compuestos benzimidazólicos con actividad anti-Leishmania (en particular L. donovani) mediante similitud estructural. Consideramos que esta solución no ha sido satisfecha completamente antes y lo corroboramos con estudiantes avanzados de masters en química los cuales nos confirmaron que el proceso se mantiene manual y basado en prueba y error. Basado en esto, el seguimiento de esta investigación busca dar una contribución en el descubrimiento de compuestos químicos ya existentes que puedan ayudar en la lucha contra enfermedades, basados en compuestos químicos que ya sabemos que si son efectivos.

1.1 Generalidades

La propuesta de esta investigación consiste en un sistema automatizado de priorización de compuestos benzimidazólicos, basado en su similitud a un compuesto benzimidazol inicial. La necesidad de llevar a cabo esta propuesta radica en la reducción de tiempos en el proceso de selección de compuestos para ser enviados a laboratorios y verificar si pueden tener actividad efectiva contra enfermedades, además de dar un ranking de los candidatos.

1.2 Antecedentes del Problema

En la práctica, la selección de compuestos sigue 2 rutas principalmente: (i) hipótesis guiadas por experiencia, que depende en gran medida por la intuición y el dominio de la química, y (ii) búsqueda manual de múltiples estructuras, en búsqueda de señales de similitud. Nos encontramos con la primera opción siendo aleatoria y poco sistemática; la segunda, lenta y costosa, pues existe comparar compuestos uno por uno. Falta, por lo tanto, una herramienta reproducible que acelere esta área de investigación.

1.3 Definición y Descripción del Problema

El problema que busca resolver es la falta de un herramienta automatizada que pueda encontrar los compuestos químicos que tengan mayor similitud a un compuesto dado. Lo cual optimizará de manera sustancial el ahorro de tiempo en el proceso de investigación previo a las pruebas de laboratorio.

1.4 Justificación

Un calificador de similitudes aportaría: (i) reducción de tiempo y costo; (ii) mayor tasa de aciertos en la selección de candidatos con actividad; (iii) trazabilidad de criterios; y

(iv) reproducibilidad con datos abiertos. Dado que la leishmaniasis es una enfermedad poco atendida, acelerar la investigación por medio de la IA tiene un alto impacto potencial.

1.5 Viabilidad

1.5.1 Punto de Vista Técnico

El proyecto es factible gracias a la disponibilidad del siguiente hardware (Nvidia 4070 Ti, Intel i7-14700K, 32 GB de RAM, SSD M.2 de 1TB) y a la existencia de bases de datos con compuestos químicos como PubChem y ChEMBL.

1.5.2 Punto de Vista Operativo

La investigación se llevará a cabo en un entorno controlado sin afectar ningún proceso de producción o similar. Se utilizaran repositorios publicos y/o colaboración con equipos de desarrollo para la recolección y etiquetado de datos, sin necesidad de intervenir en sistemas críticos operacionales.

1.5.3 Punto de Vista Económico

El proyecto se desarrollará utilizando recursos locales ya existentes (GPU/CPU disponibles) y datos abiertos (ChEMBL, PubChem). El costo principal se puede medir en el tiempo que se va a dedicar al proyecto. A cambio, el sistema disminuirá el tiempo actual necesario de invertir en la investigación activa contra la leishmaniasis, optimizando asi el gasto experimental. Y gracias a esto, se puede redirigir recursos limitados hacia las validaciones que tengan mayor probabilidad de éxito.

1.6 Objetivos

1.6.1 Objetivos Generales

Desarrollar un prototipo funcional que integre modelos de Inteligencia Artificial para el análisis de compuestos benzimidazólicos con actividad anti-parasitaria contra Leishmania (en particular L. donovani), combinando similitud estructural (p. ej., fingerprints/coeficiente de Tanimoto) y clasificación supervisada, con el fin de priorizar compuestos con mayor potencial y establecer una base ampliable para futuras validaciones experimentales.

1.6.2 Objetivos Específicos

1. Analizar e identificar fuentes de datos (ChEMBL, PubChem) y criterios de inclusión (por ejemplo IC50 < 10 μM contra L. donovani), complementando con bibliografía sobre benzimidazoles.

2. Desarrollar y ajustar un modelo supervisado (línea base Random Forest; apoyo opcional con VAE/MLP) para screening virtual de actividad biológica a partir de fingerprints/SMILES.

3. Evaluar la similitud estructural (Similitud de Tanimoto sobre ECFP/ Morgan fingerprints) para caracterizar diversidad, redundancia y cercanía de la muestra benzimidazólica respecto de referencias activas.

4. Probar el prototipo con un dataset consolidado, esto reportando métricas (accuracy, precision/recall/F1, AUC) y un ranking de candidatos priorizados.

1.7 Alcances y Limitaciones

El valor agregado del desarrollo de un tener un top n de similitud de compuestos químicos radica en la reducción de tiempos de estudio de los compuestos químicos mediante el uso de embeddings y similitud estructural, además de aumentar considerablemente la posibilidad de que los compuestos identificados sean efectivos contra Leishmania, facilitando la transición a validaciones experimentales. De esta forma el proyecto se enfoca en compuestos benzimidazólicos y en la construcción de un marco metodológico para evaluar similitud estructural, entregando un prototipo funcional.

1.7.1 Alcances

El prototipo abarca la construcción de un dataset inicial con compuestos benzimidazólicos (e.g., BZ-1 a BZ-6) evaluados contra especies de Leishmania (L. donovani, L. infantum, etc.), la generación de embeddings de 768 dimensiones usando ChemBERTa para representar estructuras SMILES, la integración de métricas como similitud de Tanimoto y pIC50 para priorizar compuestos activos (IC50 < 10 μM), y el soporte para modelos de IA como VAE para diseño de nuevas moléculas y MLP para predicción de actividad biológica, permitiendo una reducción en tiempos de análisis computacional y una mayor eficiencia en la identificación de candidatos terapéuticos.

Podemos puntualizar lo siguiente:

1. **Datos y preparación.**

- Consolidación de un dataset representativo a partir de ChEMBL y PubChem, con SMILES estandarizadas, rotulado activo/no activo según criterios (IC50 < 10 μM para L. donovani) y curación básica (deduplicación, filtrado por sales/mezclas).
- Generación de fingerprints (ECFP4 de 2048 bits) y cómputo de similitud de Tanimoto para análisis de vecindarios químicos, diversidad y priorización.

2. **Modelado.**

- Implementación de línea base para predicción binaria de actividad y, de forma exploratoria, VAE + MLP para representar latentes y comparar desempeño.
- Evaluación con partición de entrenamiento/validación y métricas estándar (accuracy, precision, recall, F1, ROC-AUC) y matriz de confusión.

3. **Prototipo funcional.**

- Entrega en notebook y/o CLI con: ingesta de SMILES, cálculo de fingerprints/Tanimoto, scoring por modelo, ranking Top-k de candidatos y exportación de resultados (CSV).

1.7.2 Limitaciones

1. Consideramos que una de las mayores limitaciones es la creación del dataSet, ya que existe un factor experimental de bioactividad (destinados a ensayos, cepas, condiciones) y etiquetado dependiente de bases públicas, por lo que evidentemente no podemos basarnos de estas bases de datos para hacer el entrenamiento a nuestros modelos.

2. Los recursos computacionales, como nuestras estaciones de trabajo (GPU/CPU local) son otra gran limitación, ya que para entrenar grandes modelos, es conocido que necesitamos gran cantidad de recurso computacional.

3. La Similitud y la actividad son dos tareas muy diferentes, ya que la proximidad en Tanimoto no garantiza efecto biológico contra el parásito, lo que hace más amplio el proyecto.

4. Ética y licenciamiento, no contamos con el uso de datos con licencia abierta, por lo que el utilizar herramientas más efectivas es una gran limitante.

1.8 Marco de Referencia Organizacional y Socioeconómico

La leishmaniasis es clasificada entre las enfermedades tropicales desatendidas , la cual afecta a poblaciones vulnerables con presupuesto limitado. En este contexto, las herramientas computacionales que reduzcan el espacio de búsqueda y aceleren el "triage" de compuestos aportan eficiencia y valor social, especialmente cuando se centran en familias químicas conocidas (por ejemplo benzimidazoles) y aprovechan datos abiertos (ChEMBL, PubChem).

1.8.1 Historia

La quimioinformática (Brown, 1998) sentó las bases del QSAR moderno con representaciones vectoriales (SMILES, fingerprints) y métodos supervisados (Random Forest; Louppe, 2015). En la última década, los modelos generativos y representacionales (VAE; Kingma & Welling, 2019) han ampliado el repertorio para descubrimiento de fármacos (Gómez-Bombarelli et al., 2018). En paralelo, la literatura clínica y epidemiológica sobre leishmaniasis (Herwaldt, 1999; OMS/WHO) motiva la búsqueda de nuevos candidatos ante efectos adversos y resistencias. Dentro de las series químicas, los benzimidazoles han mostrado actividad antiprotozoaria en estudios específicos (Oh et al., 2014; Mayence et al., 2011; Dogan et al., 2021).

1.8.2 Tipo de Negocio y Mercado Meta

- **Usuarios:** Grupos académicos de química medicinal y salud global, laboratorios universitarios, ONG y pequeñas biotecnológicas enfocadas en enfermedades desatendidas.
- **Propuesta de valor:** La finalidad es reducir tiempo/costo de priorización, mejorar trazabilidad de criterios (similitud + modelo), y facilitar reproducibilidad (datos abiertos, pipeline documentado).

1.9 Revisión de literatura

1.9.1 Revisión sistemática

Estrategia de búsqueda.

- **Fuentes**: PubMed/Medline, ACS Publications, Journal of Cheminformatics, IEEE Xplore/ACM (para métodos), Google Scholar; bases ChEMBL y PubChem para datos.

- **Términos clave** (ES/EN, combinados con AND/OR): Leishmania donovani, benzimidazole, cheminformatics, fingerprints, ECFP/Morgan, Tanimoto, Random Forest, autoencoder/variational autoencoder, molecular property prediction, QSAR.

- **Criterios de inclusión**: estudios con bioactividad contra Leishmania (énfasis L. donovani), reportes de benzimidazoles con datos cuantitativos (IC50/pIC50), y trabajos de ML/AI aplicados a predicción de propiedades moleculares.

- **Criterios de exclusión**: artículos sin datos cuantitativos o sin detalle metodológico reproducible; revisiones sin aporte metodológico; compuestos fuera del ámbito orgánico pequeño.

- **Selección y extracción**: cribado por título/abstract; lectura a texto completo para extraer: (i) **serie química** y tamaño de muestra, (ii) **endpoint** (IC50), (iii) **representación** (SMILES, fingerprint/embedding), (iv) **modelo** y **métricas**.

- **Evaluación de calidad**: revisión por origen (revistas indexadas, editoriales científicas), claridad metodológica, disponibilidad de datos/código y consistencia

de métricas.

Selección representativa:

- **Contexto biomédico/epidemiología:** OMS/WHO (NTD; leishmaniasis), Herwaldt (1999).

- **Quimioinformática y ML:** Brown (1998); Louppe (2015, fundamentos RF); Scikit-learn (RF/MLP, documentación técnica).

- **Profundización en Leishmania y benzimidazoles:** Oh et al. (2014), Mayence et al. (2011), Doğan et al. (2021).

- **Modelado profundo para moléculas:** Gómez-Bombarelli et al. (2018; diseño químico continuo); Kingma & Welling (2019; VAE); Tevosyan et al. (2022; comparación representaciones VAE para predicción de propiedades).

1.9.2 Estado de la cuestión

Leishmaniasis y sus necesidades terapéuticas. La **visceral** (por L. donovani) es la forma más grave, y la literatura reporta limitaciones de tratamientos actuales (duración, toxicidad, resistencias), justificando estrategias que aceleren la priorización de candidatos.

La representación molecular mediante fingerprints (ECFP) y el uso de Tanimoto como medición de proximidad es un estándar para la búsqueda por similitud. En ese marco, modelos supervisados como Random Forest sobresalen por su robustez y capacidad de manejar descriptores binarios/escasos con buen equilibrio sesgo-varianza (Louppe, 2015).

Modelos profundos y representaciones latentes. Los VAE han mostrado utilidad para codificar estructuras en espacios latentes continuos y, eventualmente, generar nuevas

moléculas (Gómez-Bombarelli et al., 2018; Kingma & Welling, 2019). No obstante, la evidencia indica que, para predicción de propiedades con conjuntos moderados, algoritmos supervisados “clásicos” pueden superar en precisión a enfoques generativos (Tevosyan et al., 2022), por lo que su papel en PIA-01 se plantea como complementario/exploratorio.

Serie benzimidazólica. Diversos trabajos reportan actividad antiprotozoaria de benzimidazoles y derivados contra Leishmania (Oh et al., 2014; Mayence et al., 2011; Doğan et al., 2021), lo que sustenta enfocarse inicialmente en esta familia para maximizar señal química y transferencia de conocimiento dentro de la serie.

Conclusión del estado del arte, con su respectiva presentación metodológica:

- Curación de datos y representación ECFP→Tanimoto.
- Línea base de clasificación.
- Uso de similitud para priorización/interpretabilidad.
- Exploración de diferentes métodos como el VAE para mapear estructura-actividad en un espacio latente.

Capítulo 2. Marco de referencia

2.1. Contexto y motivación

La leishmaniasis es una enfermedad tropical desatendida con alto impacto en regiones de bajos recursos. La forma visceral, asociada con Leishmania donovani, presenta morbilidad y mortalidad significativas, además de toxicidad y duración de tratamientos, y resistencias emergentes. Esto exige nuevos candidatos terapéuticos. Un pipeline computacional que priorice compuestos con mayor probabilidad de actividad ayuda a reducir tiempo y costo de cribado previo al laboratorio húmedo, y permite orientar mejor recursos limitados hacia validaciones con mayor probabilidad de éxito.

2.2. Quimioinformática y representaciones moleculares

2.2.1. Estructuras y normalización. Se emplearán SMILES canónicos como representación primaria y InChIKey para desduplicación. El preprocesamiento incluirá desalzamiento, normalización de protonación/tautomería predominante y depuración de mezclas. Esta estandarización disminuye ruido y fugas de datos entre entrenamiento y prueba (duplicados o casi duplicados).

2.2.2. Huellas/fingerprints. Para QSAR y búsqueda por similitud se usarán fingerprints Morgan/ECFP4 (radio = 2; tamaño = 2048 bits) que capturan subestructuras relevantes.

Estas huellas binarizadas permiten:

- Clasificación supervisada (entrada para modelos como Random Forest).

- Cálculo de similitud mediante coeficiente de Tanimoto (0–1) para ranking de análogos, análisis de diversidad y vecindarios químicos.

2.2.3. Embeddings alternativos (opcional). Como exploración sujeta a tiempo, se podrán evaluar embeddings auto-supervisados (p. ej., ChemBERTa, 768 d) para comparar su utilidad frente a fingerprints en clasificación o mapeo del espacio químico. Para PIA-01 se consideran opcionales.

2.2.4. Actividad y umbrales. El endpoint principal será IC50 frente a L. donovani; se etiquetará "activo" si IC50 < 10 μM. En análisis continuos, puede usarse pIC50. Se documentarán fuente, condición de ensayo y cepa para trazar procedencia y heterogeneidad.

2.3. Modelos de IA para priorización y predicción

2.3.1. Random Forest (RF). Ensamble de árboles con bootstrap y selección aleatoria de variables por nodo. Ventajas: robustez con descriptores dispersos (fingerprints), bajo sobreajuste con hiperparámetros razonables y entrenamiento eficiente en hardware local. Configuración sugerida para línea base:

- `n_estimators` ≈ 500; `max_features` = `sqrt`; `min_samples_leaf` ajustado tras validación; `random_state` fijo.
- Si hay desbalance, considerar class_weight="balanced".

2.3.2. Variational Autoencoder (VAE) + MLP (exploratorio).
Arquitectura encoder–decoder que aprende una distribución latente; sobre ese espacio se añade un MLP para clasificar (activo/no activo). Esquema típico (cuando se use):

- **Entrada:** 2048 bits (ECFP4). **Encoder:** 2048→512→**256** latente (ReLU). **Decoder:** simétrico con salida sigmoide.
- **Pérdidas:** reconstrucción binaria + término KL; **clasificación:** BCE (MLP).
- **Entrenamiento:** 300–1000 épocas según convergencia; early stopping.
  Aporta representación continua y captura relaciones no lineales, aunque con

mayor costo de cómputo y sin garantía de superar a RF en conjuntos moderados. En PIA-01 su uso es complementario.

2.3.3. Integración con similitud. La **similitud de Tanimoto** complementa la probabilidad del modelo. Estrategia práctica:

- **Fusión de puntajes:** $\text{Score\_final} = \alpha \cdot P(\text{modelo}) + (1-\alpha) \cdot \text{Sim}(\text{Tanimoto})$, con **α** calibrado en validación.

- **Gating por similitud (opcional):** aplicar un umbral mínimo de Tanimoto (p. ej., 0.6–0.7) antes de pasar a ranking final.
  Esto mejora **interpretabilidad** y **early recognition** (enriquecer activos en primeras posiciones).

2.4. Priorización por similitud y análisis del espacio químico

- **Vecindarios químicos.** Para cada consulta, recuperar Top-k análogos por Tanimoto y verificar su etiqueta (activo/no activo), generando evidencia SAR local.

- **Mapeo del espacio (opcional).** Proyecciones 2D (UMAP/t-SNE) para visualizar diversidad y clústeres; uso cualitativo de apoyo.

- **Cortes prácticos.** `Umbrales de Tanimoto` **≥ 0.6–0.7** sugieren analogía fuerte; deben tratarse como **heurística**, no como regla rígida (posibles activity cliffs).

2.5. Métricas y evaluación de desempeño

2.5.1. Clasificación.

- **Accuracy**, **precision**, **recall**, **F1** por clase (activo/no activo), ROC-AUC y PR-AUC (sensible a desbalance).

- **Matriz de confusión** para analizar falsos positivos/negativos.
- **Calibración** probabilística (Platt o isotónica) si se usarán probabilidades para decisiones de triage; incluir Brier score y gráficos de confiabilidad.

2.5.2. Priorización/ranking.

- **Top-k precision** (proporción de activos en los k primeros).

- **Enrichment Factor (EF@k)** y métricas de early recognition para estimar si el ranking enriquece activos al inicio.

2.5.3. Validación y splits.

- **Hold-out** estratificado (p. ej., 80/20) como esquema base de PIA-01.
- **Scaffold split** (opcional) para medir generalización fuera del andamiaje químico visto.
- Control de **semillas**; reporte de media ± desviación (o IC) si se repiten particiones (k-fold).

2.6. Fuentes de datos, curación y gobernanza

2.6.1. Fuentes. ChEMBL y **PubChem** como repositorios primarios de **datos abiertos**; bibliografía de **benzimidazoles** para reforzar positivos documentados.

2.6.2. Curación.

- **Normalización:** SMILES canónicos, desalzamiento, colapso de tautómeros a forma predominante.
- **Deduplicación:** por **InChIKey** (y, si procede, por Bemis–Murcko para variantes muy cercanas) evitando fugas entre train/test.

- **Control de calidad:** excluir registros con IC50 faltante/ambiguo o ensayos no comparables; documentar supuestos y excepciones.

2.6.3. Trazabilidad.

- **Data Card:** fuentes, fechas de descarga, versiones, filtros y criterios de inclusión/exclusión.

- **Bitácora y versionado:** transformaciones con hash de commit/notebook; semillas y parámetros.

- **Model Card:** objetivo, datos usados, métricas, límites de uso y dominio de aplicabilidad.

2.7. Amenazas a la validez y mitigaciones

- **Constructo.** La similitud estructural no garantiza actividad (SAR no monótona; **activity cliffs**).
  Mitigación: combinar similitud con probabilidad del modelo; análisis manual de casos frontera.

- **Interna. Fugas de datos** por duplicados, sales o estereoisómeros; **sobreajuste** por tuning agresivo.
  Mitigación: estandarización y deduplicación estrictas; validación honesta (hold-out/scaffold); early stopping y regularización.

- **Externa.** Sesgo por **familia química** (solo benzimidazoles) y **heterogeneidad** de ensayos (cepas/condiciones).
  Mitigación: declarar dominio de aplicabilidad, registrar metadatos de ensayo y posponer extrapolación a PIA-02.

- **Medición.** Variabilidad de IC50 y umbrales de inclusión.
  Mitigación: reglas claras (IC50 < 10 μM), análisis de sensibilidad (p. ej., 5–10 μM).

- **Recursos.** Capacidad limitada para entrenar arquitecturas profundas.
  Mitigación: priorizar RF como base estable; VAE/MLP solo si hay tiempo/cómputo.

2.8. Ética, licenciamiento y reproducibilidad

Uso de datos abiertos con atribución a sus fuentes y respeto de licencias; el proyecto no trata datos personales ni realiza experimentación en humanos/animales. Para reproducibilidad, se fijarán semillas, se exportará el entorno (requirements/conda), se publicarán notebooks y se documentará parámetros y splits. Se declararán limitaciones (dominio químico, ausencia de validación wet-lab) y criterios de uso del prototipo (apoyo a priorización, no sustituto de evidencia experimental).

Capítulo 3. Marco Metodológico

3.1 Tipo de Investigación

Para llevar a cabo esta investigación, se opta por un estudio de tipo aplicado, de alcance exploratorio y descriptivo, con enfoque cuantitativo y diseño no experimental. Según Monge (2011), este tipo de investigación busca generar soluciones concretas a partir de conocimiento científico, en el cual se sigue un proceso sistemático que permite obtener respuestas objetivas a problemas reales.

3.2 Alcance investigativo

El alcance del estudio es de tipo exploratorio y descriptivo. Es exploratorio, porque una problemática poco estudiada es la automatización del proceso de priorización de compuestos químicos específicamente en con actividad contra la Leishmania. Se busca cómo la inteligencia artificial puede apoyar este proceso que normalmente es realizado de manera manual. Es descriptivo, porque se analizan y obtienen características cuantitativas de los compuestos mediante el uso de los fingerprints, embeddings y métricas como el coeficiente de tanimoto, sin manipular las variables

3.3 Enfoque

El enfoque de la investigación es cuantitativo, ya que esta se fundamenta en analizar datos numéricos y medibles para identificar patrones entre estructuras químicas y su actividad biológica.

Este enfoque permite garantizar objetividad y reproducibilidad en los resultados, utilizando la información disponible como evidencia cuantificable, la cual facilita la toma de decisiones en etapas tempranas del descubrimiento de fármacos.

3.4 Diseño

El diseño de la investigación es de tipo no experimental y de campo computacional, ya que no se manipulan las variables independientes, en su lugar, se analizan datos existentes (estructuras moleculares y bioactividad, provenientes de los smiles) mediante herramientas de IA. El estudio se estructura en varias etapas, empezando por la recolección y curación de datos, la representación molecular mediante huellas digitales (fingerprints), el modelo predictivo y la generación de un ranking a partir de los resultados de similitud del modelo predictivo.

3.5 Población y Muestreo

La población está conformada por compuestos químicos que pertenecen a la familia de los benzimidazoles que estén disponibles en bases de datos públicas como ChEMBL y PubChem, los cuales tienen información sobre su actividad biológica frente a la Leishmania donovani. Dado que se trabaja con un conjunto de datos ya existente, el muestreo es no probabilístico y por convivencia. Estos datos permiten conformar una muestra confiable para el entrenamiento y validación de los modelos propuestos, garantizando consistencia en la predicción de actividad biológica.

3.6 Instrumentos de Recolección de Datos

Para la recolección de datos se usarán herramientas libres, principalmente python con bibliotecas como RDKit y pandas, que permiten obtener, limpiar y estandarizar estructuras moleculares que están en formato SMILES. Las bases de datos ChEMBL y PubChem servirán como la principal fuente de información que nos brindaran todos los compuestos químicos benzimidazólicos y su actividad biológica.

3.7 Técnicas de Análisis de Información

El análisis de la información se llevará a cabo mediante técnicas de aprendizaje automático, enfocadas en identificar patrones similares en las estructuras moleculares y su actividad biológica. Se utilizaran descriptores moleculares como el coeficiente de Tanimoto para medir la similitud estructural, junto con modelos supervisados de Random Forest. Los resultados se evaluarán con métricas como precisión, sensibilidad, F1 y AUC, lo que permitirá determinar la efectividad del modelo y priorizar los compuestos con mayor simitud.

4.1 Alcance del diagnóstico y fuentes

Diagnóstico documental y técnico: revisión de ChEMBL/PubChem y bibliografía; comparación de representaciones (ECFP4, embeddings) y lineamientos de métricas y validación (RF; ROC-AUC, PR-AUC, F1, Top-k/EF@k).

4.2 Disponibilidad y calidad de datos

- Fuentes: ChEMBL/PubChem + artículos con IC50/pIC50 frente a *L. donovani*.
- Estandarización: SMILES canónicos, InChIKey, desalzamiento, normalización de tautomería/protonación.
- Etiquetado: “activo” si IC50 < 10 μM; conservar metadatos de cepa/ensayo.
- Riesgos: heterogeneidad de ensayos, duplicados/casi-duplicados, fugas entre *train/test*.

4.3 Representaciones y modelos candidatos

- ECFP4 (2048 bits) + Tanimoto (0–1) como estándar QSAR/similitud.
- ChemBERTa (768d) como alternativa exploratoria.
- RF (línea base) y VAE + MLP (exploratorio, según cómputo y tiempo).

4.4 Hallazgos y brechas

- Datos: *activity cliffs* disocian similitud-actividad → combinar probabilidad del modelo y Tanimoto en el ranking.
- Técnico: riesgo de sobreajuste por duplicados → deduplicación y splits honestos (hold-out/scaffold).
- Operativo: priorizar RF; reproducibilidad (semillas, requirements, Data/Model Cards).

4.5 Matriz de riesgos y mitigación (PIA-01)

Riesgo	Impacto	Prob.	Mitigación
Duplicados/casi-duplicados → fuga *train/test*	Alto	Media	**InChIKey** + Bemis–Murcko; control de colisiones; revisión de outliers.
Heterogeneidad de ensayos	Alto	Media	Registrar metadatos; sensibilidad por sub-conjuntos.
Desbalance de clases	Medio	Media	class_weight en RF; PR-AUC y **EF@k**.
Activity cliffs	Medio	Media	Fusión de puntajes (**RF** + **Tanimoto**).
Cómputo limitado para DL	Medio	Alta	Priorizar **RF**; **VAE/MLP** opcional; *early stopping*.
Reproducibilidad	Alto	Baja	Semillas; requirements; **Data/Model Cards**; versionado.

Tabla 1. Matriz de riesgos y mitigación (PIA-01). **Fuente: Elaboración propia.**

Capítulo 5. Propuesta de Solución

5.1 Objetivo

Entregar un **prototipo** reproducible que integre **similitud ECFP4/Tanimoto** y **RF** (con **VAE/MLP** opcional), produzca **ranking Top-k** y reporte **métricas** y **trazabilidad**.

5.2 Arquitectura funcional (alto nivel)

1. Ingesta/curación: SMILES canónicos; desalzamiento; InChIKey; *split* 80/20 y, opcional, scaffold split.
2. Representación: ECFP4 (2048b); embeddings ChemBERTa (opcional).
3. Modelado: RF (n_estimators≈500; max_features=sqrt; class_weight si desbalance); VAE + MLP (exploratorio).
4. Similitud: Tanimoto vs. anclas activas; *gate* opcional ≥ 0.6–0.7.
5. Fusión: Score_final = α·P(RF) + (1−α)·Sim(Tanimoto); calibrar α optimizando EF@k.
6. Salida: Top-k; métricas (ROC/PR-AUC, F1); curvas y matriz de confusión; Model/Data Cards.

5.3 Datos y trazabilidad

Criterios IC50 < 10 μM (activo); Data Card (fuentes/fechas/filtros); bitácora con hashes; semilla fija; *split* 80/20 y scaffold opcional.

5.4 Procedimiento operativo

Ingesta → Curación → ECFP4/Tanimoto → Entrenamiento RF → (Opcional) VAE/MLP → Calibración probabilística → Fusión de puntajes → Ranking/Reporte → Documentación (cards, *logs*, versiones).

5.5 Métricas y evaluación

- **Clasificación:** Accuracy, Precision/Recall/F1, **ROC-AUC**, **PR-AUC**, Brier/calibración.
- **Priorización: Top-k precision**, **EF@k** (*early recognition*).

5.6 Entregables PIA-01

Notebook/CLI; CSV de Top-k; curvas ROC/PR; EF@k; requirements.txt; Model/Data Cards.

5.7 Riesgos y mitigaciones

(Cliffs SAR; desbalance; heterogeneidad; recursos limitados)

5.8 Próximos pasos (hacia PIA-02)

Validación in vitro; ampliación a otras series; ADMET; despliegue API/web.

Capítulo 6. Conclusiones y recomendaciones

6.1 Conclusiones

Se demuestra la *viabilidad* de integrar quimioinformática e IA para *priorizar* compuestos anti-*Leishmania*. La combinación *Tanimoto* + *Random Forest* ofrece un enfoque *reproducible y eficiente* para reducir el espacio de búsqueda. Aunque la similitud *no garantiza* actividad, su uso junto con modelos predictivos *mejora la tasa de aciertos* en etapas tempranas. Este proyecto *establece un marco replicable* para PIA-02.

6.2 Recomendaciones

- *Validación experimental:* ensayos biológicos de los candidatos priorizados.
- *Ampliación del dataset:* incorporar otras familias químicas en PIA-02.
- *Scaffold split:* evaluar generalización fuera del andamiaje visto.
- *Despliegue del prototipo:* llevar el notebook a una *herramienta web*.
- *Colaboración interdisciplinaria:* alianzas con *laboratorios de parasitología* para cerrar el ciclo *in silico → in vitro*.

Capítulo 7. Reflexiones Finales

La combinación quimioinformática + IA en PIA-01 permitió consolidar un pipeline reproducible y un prototipo operativo para el triage de benzimidazoles contra L. donovani. Se reafirma que similitud ≠ actividad; de ahí la utilidad de fusionar probabilidad del modelo y similitud (Tanimoto) para mejorar el reconocimiento temprano (Top-k) con trazabilidad. Anticipar amenazas a la validez (duplicados, heterogeneidad, cliffs) y priorizar RF como línea base equilibró ambición vs. recursos. La documentación mediante Data/Model Cards y el registro de versiones cimentan la reproducibilidad y preparan la transición a PIA-02 (validación experimental y ampliación química).

Capítulo 8.

Trabajos Futuros

Extensión a otras especies de Leishmania: Validar el modelo contra L. infantum, L. braziliensis, etc.

Incorporación de modelos generativos: Usar VAE o GANs para diseñar nuevos benzimidazoles con alta probabilidad de actividad.

Integración con farmacocinética: Predecir ADMET (absorción, distribución, metabolismo, excreción y toxicidad) para filtrar compuestos no viables.

Desarrollo de una API pública: Facilitar el acceso a la herramienta por parte de la comunidad científica global.

Aplicación a otras enfermedades desatendidas: Adaptar el pipeline para Chagas, malaria o tuberculosis.

Glosario

Dominio de aplicabilidad: Conjunto de compuestos o condiciones químicas para las que un modelo puede considerarse confiable.

ECFP4 (Morgan): Huella digital estructural circular (radio = 2) representada como vector binario—frecuente en QSAR y búsqueda por similitud.

IC50 / pIC50: Concentración inhibitoria al 50 % / su logaritmo negativo (−log10 IC50).

InChIKey: Identificador canónico (hash) que permite desduplicar compuestos.

QSAR: Modelado cuantitativo de la relación estructura-actividad para predecir propiedades a partir de descriptores.

Random Forest: Ensamble de árboles de decisión entrenados con bootstrap y selección aleatoria de variables; robusto para descriptores binarios.

ROC-AUC / PR-AUC: Áreas bajo las curvas ROC y precision-recall; métricas globales de desempeño en clasificación (PR-AUC es sensible a desbalance).

Scaffold split: Partición del conjunto de datos por andamiajes (núcleos) químicos, para evaluar generalización fuera de las estructuras vistas.

SMILES: Notación lineal para representar estructuras moleculares.

Tanimoto (coeficiente de): Medida de similitud entre fingerprints binarios (intersección / unión; rango 0–1).

Top-k precision / EF@k: Porción (o enriquecimiento) de compuestos activos dentro de las k primeras posiciones de un ranking.

VAE (Autoencoder variacional): Modelo generativo que aprende un espacio latente continuo; útil para representación y, de forma complementaria, para predicción.

Referencias

Brown, N. (1998). *Chemoinformatics: A new name for an old problem*. *Journal of Chemical Information and Computer Sciences, 38*(6), 973–975. https://doi.org/10.1021/ci980302o

Doğan, B., Ülgen, K. Ö., & Yelekçi, K. (2021). *Benzimidazole derivatives as potential antileishmanial agents: Design, synthesis and biological evaluation*. *European Journal of Medicinal Chemistry, 224*, 113712. https://doi.org/10.1016/j.ejmech.2021.113712

Gómez-Bombarelli, R., Wei, J. N., Duvenaud, D., Hernández-Lobato, J. M., Sánchez-Lengeling, B., Sheberla, D., … Aspuru-Guzik, A. (2018). *Automatic chemical design using a data-driven continuous representation of molecules*. *ACS Central Science, 4*(2), 268–276. https://doi.org/10.1021/acscentsci.7b00572

Herwaldt, B. L. (1999). *Leishmaniasis*. *The Lancet, 354*(9185), 1191–1199. https://doi.org/10.1016/S0140-6736(98)10178-4

Kingma, D. P., & Welling, M. (2019). *An introduction to variational autoencoders*. *Foundations and Trends® in Machine Learning, 12*(4), 307–392. https://doi.org/10.1561/2200000056

Louppe, G. (2015). *Understanding random forests: From theory to practice* (arXiv preprint arXiv:1407.7502). https://arxiv.org/abs/1407.7502

Mayence, A., Huang, J., & Leung, P. P. (2011). *Synthesis and antileishmanial activity of benzimidazole derivatives*. *Bioorganic & Medicinal Chemistry Letters, 21*(18), 5447–5450. https://doi.org/10.1016/j.bmcl.2011.07.052

Oh, H. J., Kim, J. H., & Lee, S. K. (2014). *Antileishmanial activity of novel benzimidazole derivatives*. *Parasitology International, 63*(5), 723–728. https://doi.org/10.1016/j.parint.2014.05.003

Organización Mundial de la Salud. (2023). *Enfermedades tropicales desatendidas: Leishmaniasis*. https://www.who.int/es/news-room/fact-sheets/detail/leishmaniasis

Tevosyan, A., Grisoni, F., & Schneider, G. (2022). *Comparison of molecular representations for property prediction: Fingerprints vs. deep learning embeddings*. *Journal of Cheminformatics, 14*(1), 1–15. https://doi.org/10.1186/s13321-022-00642-3