Escuela de Bio y Nanotecnologías (EByN)
Universidad Nacional de de San Martín

Doctorado en Biología Molecular y Biotecnología

Tesis de Doctorado

Quimiogenómica aplicada a la identificación y reposicionamiento de compuestos bioactivos para la Enfermedad de Chagas

Abril, 2024

Autor:
Lionel Urán Landaburu

Director:
Fernán Agüero

Índice general

Nota: todos los links, en azul en la versión electrónica, son activos y permiten navegar hacia las respectivas secciones, referencias, notas al pie y sitios en internet.

Índice de figuras

Nota: todos los links, en azul en la versión electrónica, son activos y permiten navegar hacia las respectivas secciones, referencias, notas al pie y sitios en internet.

Índice de tablas

Nota: todos los links, en azul en la versión electrónica, son activos y permiten navegar hacia las respectivas secciones, referencias, notas al pie y sitios en internet.

1. Resumen

Actualmente existe una necesidad urgente de desarrollo de nuevas drogas para combatir enfermedades infecciosas tropicales asociadas a la pobreza, tales como la Malaria, la Enfermedad de Chagas y la Tripanosomiasis Africana, entre otras. Incluso en los casos en los que se cuenta con drogas para estas enfermedades, su uso se ve limitado por su alto costo, baja eficacia, toxicidad y aparición de resistencias. A partir del conocimiento de la secuencia genómica completa de varios patógenos, se iniciaron acciones tendientes a aprovechar estos datos para la identificación de nuevos blancos terapéuticos.

En el laboratorio de Genómica y Bioinformática de la UNSAM se desarrolló una base de datos, (TDR Targets) que contiene información genómica de patógenos prioritarios para el Special Programme for Research and Traning in Tropical Diseases (TDR) de la *World Health Organization* (WHO). Esta base de datos puede ser utilizada como una herramienta computacional para priorizar potenciales blancos de drogas, siguiendo distintas estrategias, o como herramienta de consulta. Recientemente se ha incorporado a esta base de datos información relacionada a >1,5 millones de compuestos bioactivos con potencial de uso como drogas en el tratamiento de estas enfermedades; conjuntamente con una serie de herramientas quimioinformáticas que permiten explorar esta información. La mayor parte de estos compuestos han sido ensayados para otras enfermedades o indicaciones, de manera que hay gran potencial para realizar estrategias de reposicionamiento.

El presente trabajo propone distintas estrategias para la integración de datos bioinformáticos, quimioinformáticos y quimiogenómicos. En el capítulo 3 se exponen los desafíos y oportunidades que presenta dicha integración, y se presentan las distintas actualizaciones al repositorio quimiogenómico (TDR Targets) que permiten la exploración y explotación de los mismos para asistir al proceso de reposicionamiento de moléculas bioactivas hacia distintas enfermedades desatendidas. En el capítulo 4 se describe un flujo de priorización de blancos y moléculas bioactivas que culmina en la comprobación experimental de las inferencias obtenidas, probando compuestos identificados en el análisis computacional en ensayos *in vitro* en tripanosomátidos para evaluar su capacidad tripanocida. Surgen de este capítulo múltiples hipótesis de trabajo, entre las que destaca la validación experimental de la esencialidad de la monoacilglicérido lipasa (MAGL) en *Trypanosoma cruzi* (un potencial blanco terapéutico completamente nuevo para este patógeno), y la determinación del mecanismo de acción los 5 *hits* hallados durante este trabajo.

En su conjunto, esta tesis demuestra el gran potencial que albergan los datos quimiogenómicos generados hasta el momento para brindar apoyo al desarrollo de nuevas drogas para enfermedades desatendidas, en general, y para la Enfermedad de Chagas en particular.

Palabras clave: *Desarrollo de drogas*, *Enfermedades desatendidas*, *Enfermedad de Chagas*, *Tripanosomiasis americana*, *Big Data*, *Base de datos*, *Compuestos bioactivos*, *Reposicionamiento de drogas*, *Quimiogenómica*, *Quimioinformática*, *Bioinformática*

Keywords: *Drug development*, *Neglected diseases*, *Chagas disease*, *American trypanosomiasis*, *Big Data*, *Databases*, *Bioactive Compounds*, *Drug repurposing*, *Chemogenomics*, *Chemoinformatics*, *Bioinformatics*

1.1. Publicaciones

Las siguientes publicaciones contienen algunos de los resultados y revisiones presentados en esta tesis:

- Urán Landaburu, L., Didier-Garnham, M. & Aguero, F. Targeting trypanosomes: how chemogenomics and artificial intelligence can guide drug discovery. *Biochem Soc Trans* **51**, 195–206 (2022). URL https://portlandpress.com/biochemsoctrans/article-abstract/doi/10.1042/BST20220618/232416/Targeting-trypanosomes-how-chemogenomics-and

- Urán Landaburu, L. *et al.* TDR Targets 6: driving drug discovery for human pathogens through intensive chemogenomic data integration. *Nucleic Acids Research* gkz999 (2019). URL https://academic.oup.com/nar/advance-article/doi/10.1093/nar/gkz999/5611677

- Salas-Sarduy, E. *et al.* Novel scaffolds for inhibition of cruzipain identified from high-throughput screening of anti-kinetoplastid chemical boxes. *Sci. Rep.* **7**, 12073 (2017). URL http://www.nature.com/articles/s41598-017-12170-4

- Salas-Sarduy, E. *et al.* Potent and selective inhibitors for M32 metallocarboxypeptidases identified from high-throughput screening of anti-kinetoplastid chemical boxes. *PLOS Neglected Tropical Diseases* **13**, e0007560 (2019). URL https://dx.plos.org/10.1371/journal.pntd.0007560

1.2. Objetivos

Objectivos Generales

1. Crear soluciones quimiogenómicas guiadas por datos (*data-driven*) para asistir en el reposicionamiento de drogas para enfermedades desatendidas

Objetivos Específicos

1. **Armonización de los datos**: Definir lineamientos estandarizados para la integración de datos quimiogenómicos de distintas fuentes y en distintos formatos.

2. **Priorización de blancos**: Establecer una propiedad cuantitativa que dé cuenta de la drogabilidad potencial de los blancos proteicos.

3. **Priorización de drogas**: Desarrollar algoritmos que permitan pasar de la colección de datos integrados a un conjunto de blancos o drogas de interés, con potencial terapéutico.

4. **Democratización de los datos**: Poner a disposición los datos integrados y todos los cálculos derivados de éstos a través de un software consultable (base de datos) con interfaz web.

5. **Reposicionamiento**: Utilizar los datos de priorización para obtener una lista reducida de especies químicas y probar su actividad tripanocida *in vitro*

2. Introducción

2.1. Justificación

La Organización Mundial de la Salud ha identificado 20 enfermedades infecciosas tropicales desatendidas [5]. Estas enfermedades, entre las cuales se encuentra la leishmaniasis, filariasis, esquistosomiasis, la enfermedad del sueño y la enfermedad de Chagas, afectan a 1,000 millones de personas, y hay una cantidad similar en riesgo de contraerlas [6]. En lineas generales, estas dolencias han afectado históricamente a personas que viven en condiciones de pobreza en África, Asia y América Latina [7]. En la última década, no obstante, algunas de estas — como la enfermedad de Chagas — han visto un incremento en el número de casos reportados en países Canada, Australia, Japón, EE.UU. y algunos países de Europa [8].

Los tratamientos actuales para estas enfermedades presentan limitaciones debido a su costo, dificultades en la administración, alta toxicidad, y aparición de resistencia, entre otros [7]. Sin embargo históricamente no ha habido demasiado interés comercial en el desarrollo de nuevas formas de tratamiento, principalmente debido a la baja expectativa de ganancia, ya que se trata de enfermedades que afectan mayormente a poblaciones de bajos recursos [9]. Como consecuencia, solamente el 1 % del total de nuevas drogas que llegaron al mercado en los últimos 25 años fueron para el tratamiento de estas enfermedades [7, 10]. En los últimos años surgieron asociaciones entre organizaciones públicas y privadas, tales como *Drugs for Neglected Diseases initiative* y *Medicines for Malaria Venture* con el objetivo de llevar a la clínica moléculas prometedoras que surjan de centros industriales o académicos de descubrimiento de drogas [9].

En cuanto a la selección de blancos para el descubrimiento de drogas, tradicionalmente se ha enfocado en estudiar si la alteración de la actividad normal de un blanco potencial puede tener algún efecto terapéutico, sin tener en cuenta la probabilidad de descubrir nuevos ligandos para ese blanco en particular [7, 11]. El concepto de "drogabilidad" (*druggability*) evalúa la probabilidad de un blanco de unirse y ser modulado por un ligando. Las predicciones de drogabilidad se basan en datos empíricos de relaciones estructura-actividad, los cuales requieren una gran cantidad de datos para entrenar y validar[11]. Además, esta estrategia solo permite explorar blancos terapéuticos arduamente estudiados (o blancos relacionados a éstos) y, consecuentemente, carece de la capacidad de descubrir nuevos blancos. El uso de plataformas quimioproteómicas puede ayudar a salvar este punto ciego, mediante *screenings* de alto rendimiento que permiten ensayar múltiples ligandos contra un centenar de proteínas recombinantes [12]. No obstante, el costo de este tipo de estrategias resulta prohibitivo para buena parte de los esfuerzos de *drug discovery* en enfermedades desatendidas.

El proceso de descubrimiento de drogas ha ido cambiando en el tiempo. Hace unos 50 años, el potencial de un compuesto como posible droga estaba principalmente determinado por el resultado de ensayos en modelos animales, lo cual hoy en día se considera una prueba pre-clínica avanzada. El foco buscaba determinar si la droga causaba el efecto deseado, prestando poca importancia a otros aspectos como la afinidad de la droga por su blanco o su especificidad [13, 14].

Sin embargo, en los años 80 se produjo un cambio de estrategia en el proceso de descubrimiento de nuevas drogas. Esto comenzó con el desarrollo de la biología molecular, el conocimiento de los mecanismos de acción de muchas drogas, y más recientemente con la capacidad de secuenciar genomas completos. El conocimiento en mayor profundidad de diversos procesos biológicos y

de los agentes moleculares involucrados produjo este cambio de estrategia, ya que ahora resulta posible y lógico elegir proteínas con alto potencial de convertirse en buenos blancos de drogas – en el sentido de potencial de modulación química – y luego racionalmente diseñar drogas que pudieran interferir con su actividad. Es decir, que el criterio para evaluar si un compuesto podía ser utilizado como droga dejó de ser estrictamente fisiológico para incorporar también un criterio molecular, donde aquellos compuestos con mayor potencial eran los que mostraban una buena afinidad y especificidad para un determinado blanco [13].

Más recientemente se han incorporado métodos más globales y comparativos que integran información de varios blancos, permitiendo relacionar globalmente el espacio químico del conjunto de blancos. En este contexto han surgiendo métodos *in silico* que hacen posible estudiar a gran escala tanto la probabilidad de distintos compuestos de modular blancos como encontrar nuevas indicaciones terapéuticas para viejas drogas (reposicionamiento). Por otro lado, los enfoques quimiogenómicos están surgiendo como una nueva disciplina en cuanto a predicción de blancos mediante la exploración de bases de datos [15].

El paradigma moderno de descubrimiento y desarrollo de drogas se divide en entonces en 4 etapas bien diferenciadas: identificación (de blancos terapéuticos), ensayos pre-clínicos, ensayos clínicos y registro/aprobación. Todo el proceso puede durar entre 15 y 18 años, y costar hasta $ 12.000 M de dólares [16] (Figura 2.1).

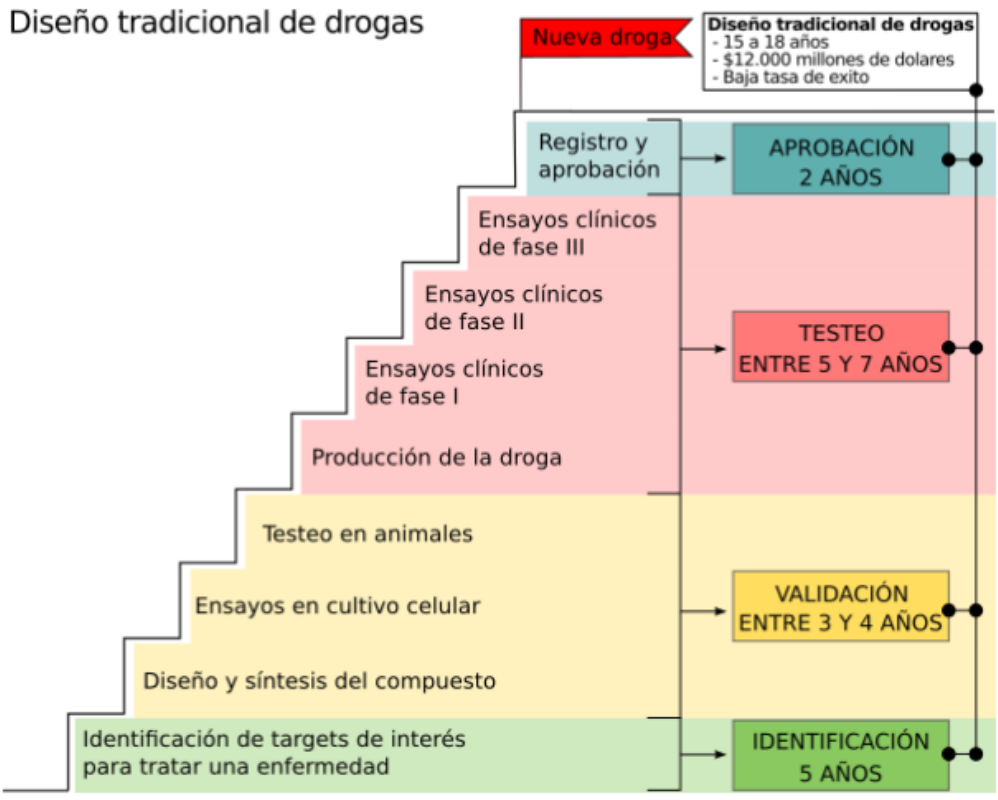


Figura 2.1 – Etapas comunmente halladas en procesos de *drug discovery* convencionales.

2.2. Desarrollo o descubrimiento de drogas *de novo*

La estrategia de descubrimiento *de novo* se focaliza en la identificación de nuevas moléculas, tanto compuestos sintéticos como productos naturales [17]. Se trata de un enfoque a largo plazo e integra la investigación basada en ensayos a gran escala centrados en la modulación de blancos, y ensayos a mediana escala a realizados sobre parásitos enteros o proteínas específicas. Cada una de estas alternativas constituyen la primera etapa (identificación) de un proceso de *drug discovery de novo* y, de resultar exitosas, son sucedidas por los pasos de validación, testeo y aprobación (como puede apreciarse en la figura 2.1).

En particular, para algunas enfermedades desatendidas (y muy frecuentemente para aquellas exploradas en este trabajo), se ha visto poco éxito con estrategias basadas en blancos: muchos compuestos activos en ensayos sobre proteínas son inactivos en ensayos de células enteras. Esto puede deberse a diversos problemas: baja permeabilidad de las drogas, que los blancos no sean esenciales, o no sean expresados por el estadío de vida de interés del parásito, o sí sean expresados pero en un exceso tal que sea imposible lograr una inhibición que afecte considerablemente al patógeno (un ejemplo es la pantotenato quinasa de *M. tuberculosis* [18]).

Por lo tanto, la estrategia de ensayos a gran escala basada en blancos es válida, pero necesita ser mejorada en cuanto a la validación de los mismos y la calidad de las bibliotecas de compuestos elegidas para las primeras etapas de la investigación. Se trata de una estrategia complementaria a los ensayos de células enteras, y no un reemplazo de las mismas [17].

Como contraparte, la estrategia de *screening* sobre parásitos enteros ha dado a luz a buena parte de los tratamientos (no reposicionados) que usamos en la actualidad. Como el benznidazol y el nifurtimox (enfermedad de Chagas) [19], la artemisinina [20] (Malaria), y la suramina o el melarsoprol (tripanosomiasis africana) [21]. Sin embargo, estos éxitos pueden ser discutibles, dado que han sido logros de hace más de 50 años e incluso estando actualmente en el repertorio de herramientas para tratar estas enfermedades, algunas de estas moléculas difícilmente podrían ser aprobadas bajo los estándares modernos de desarrollo de drogas por su perfil de toxicidad [5].

2.3. Estrategias de reposicionamiento

Tanto el factor temporal como el factor económico resultan prohibitivos para la mayoría de las enfermedades desatendidas, por lo que históricamente han sido pocas las iniciativas de descubrimiento *de novo* [17], y han primado estrategias que aprovechan esfuerzos conjuntos (o directamente ajenos) a la enfermedad que intenta tratarse, tales como la extensión de la indicación original o el *piggy-back drug discovery*.

La principal estrategia de "descubrimiento" de drogas para enfermedades tropicales se basó en extender la indicación de tratamientos existentes para otras enfermedades en humanos o animales a enfermedades tropicales [22]. Este enfoque ha sido exitoso, y tuvo como resultado algunas de las drogas antiparasitarias más importantes de uso en la actualidad, como ivermectina o albendazol para la filariasis/oncocercosis [23, 24]; el praziquantel para la esquistosomiasis [25]; y la pentamidina, originalmente prescripta para el tratamiento de tripanosomiasis equina y actualmente usada para el tratamiento de tripanosomiasis africana en su etapa aguda [21].

En cuanto al descubrimiento *piggy-back*, este es el caso de compuestos que estén siendo investigados como potenciales drogas para una enfermedad de interés comercial, cuyo blanco es una proteína de otro organismo que también se encuentra presente en el parásito de interés.

Un ejemplo de esta estrategia son las drogas inhibidoras de desacetilasas de histonas, que fueron desarrolladas originalmente para el tratamiento de cáncer,o los inhibidores de cistein proteasas que están siendo desarrollados para la osteoporosis; ambos fueron probados como drogas antimalaria [17].

Las estrategias discutidas arriba constituyen ejemplos de reposicionamiento, y pueden, a su vez, enmarcarse dentro de distintas formas, categorizándolas según la entidad reposicionada (Figura 2.2), sea ésta un agente químico o un blanco terapéutico; y según el conocimiento acumulado para dicha entidad. Esta clasificación es especialmente útil porque permite identificar rápidamente la instancia o el paso del *pipeline* de *drug discovery* en el que se insertan estas entidades, lo que se traduce en una estiamación casi intuitiva de la reducción de los costos que propone una estrategia u otra.

Así, reposicionar una familia o clase de blancos tendrá un menor impacto que reposicionar un blanco (o *target*) cuyo ortólogo ya ha sido identificado y caracterizado en el parásito de interés. Reposicionar un *lead* impone la necesidad de un proceso de optimización (ya sea para mejorar su actividad, su biodisponibilidad, o su perfil de toxicidad) que no es necesario cuando lo que se reposiciona es una droga comercial, siendo este último el caso el ideal.

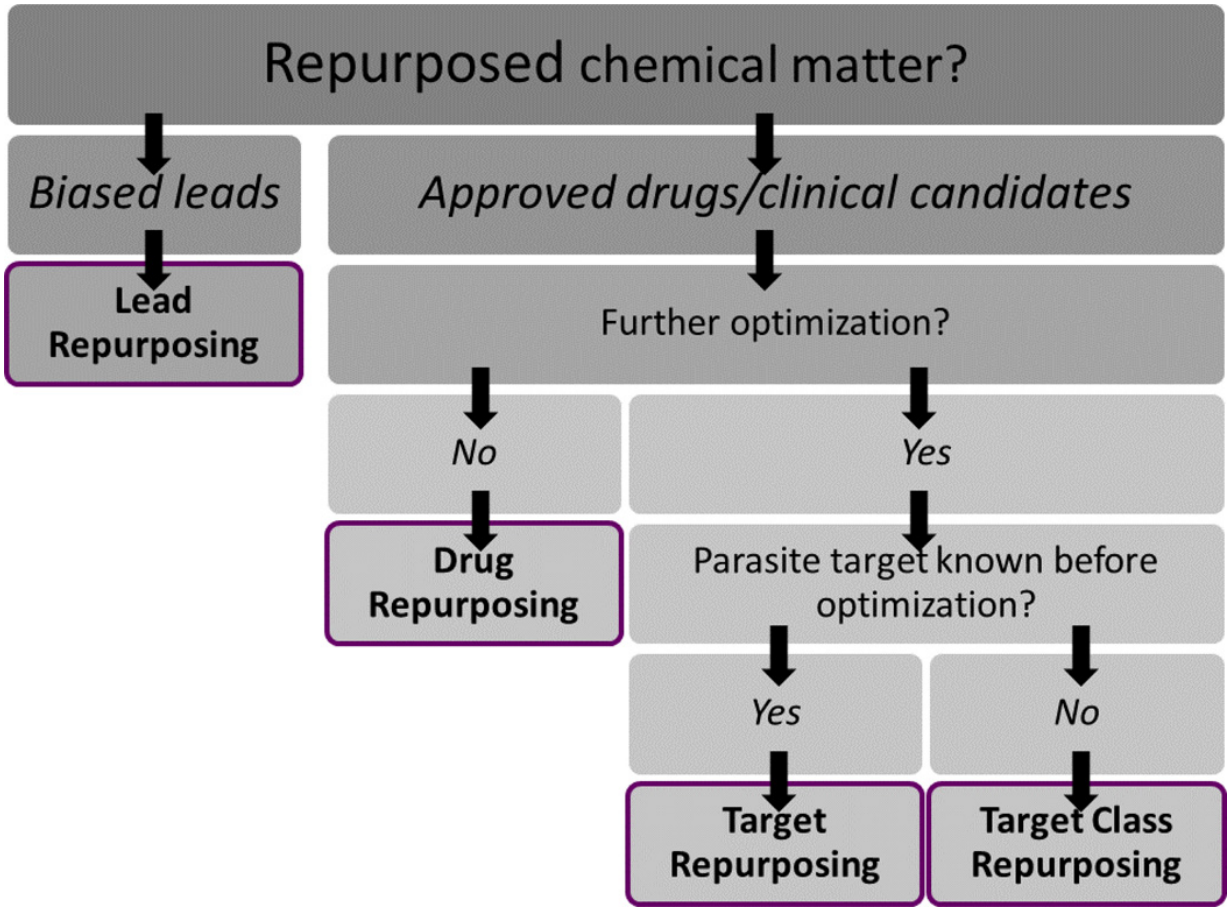


Figura 2.2 – Casos de reposicionamiento según la entidad reposicionada y el conocimiento disponible. Adaptado de Klug et al, 2016.

2.3.1. Reposicionamiento de drogas

El reposicionamiento de drogas se caracteriza por requerir mínimas adecuaciones para que el agente químico pueda ser utilizado en su nueva indicación. Es el caso de drogas que ya han pasado por todas las pruebas de validación y ensayos clínicos subsiguientes, por lo que se ha comprobado que son seguras y que poseen un perfil farmacocinético deseable.

La eflornitina, un inhibidor de la síntesis de poliaminas que inactiva irreversiblemente de la ornitina decarboxilasa [26], fue originalmente desarrollado por Merrell Dow Research Institute hacia final de 70s con fines oncológicos. A pesar de ser poco efectivo a este fin, fue reposicionado para el tratamiento de tripanosomiasis africana causada *Tryipanosoma brucei gambiense*, luego de mostrar capacidad curativa tanto *in vitro* como *in vivo*, en modelo de ratones infectados con dicho parásito [27]. En la actualidad, la indicación de eflornitina+nifurtimox es el tratamiento estándar para el tratamiento de esta enfermedad, siendo efectiva incluso en el estadío que afecta el sistema nervioso central.

Otro ejemplo de reposicionamiento de drogas, con un camino recorrido similar al de la eflornitina, es el del tamoxifeno. Originalmente aprobado para el tratamiento de cáncer de mama con marcador de receptor de estrógeno positivo (ER) en 1977 [28], el tamoxifeno mostró una potente actividad contra amastigotes intracelulares de *Leishmania amazonensis*. El mecanismo por el cual la droga mata a los amastigotes no es directo, sino que lo logra por alcalinización de la vacuolas parasitóforas (PVs) de los macrófagos en la éstos residen.

Ambos ejemplos comparten el haber sido reposicionados sin necesidad alguna de optimización (aunque las dosis y por lo tanto los esquemas de tratamiento sí se fueron ajustando). El camino regulatorio ya recorrido por una droga permite aventurar el uso de la misma para otras indicaciones con muy bajo riesgo de que estas resulten tóxicas para el paciente. Es importante notar que el reposicionamiento de drogas *per se* ignora si el mecanismo de acción es el mismo (o no) que el buscado en la indicación original.

2.3.2. Reposicionamiento de *targets*

El reposicionamiento de *targets* o blancos terapéuticos es quizás el caso de más abundante entre los casos de reposicionamiento. La capacidad de secuenciar genomas completos que trajo consigo la era post-genómica de la biología abrió el juego para estrategias de búsqueda de drogas basadas en mecanismos de acción definidos y la conservación de genes entre dos o más especies. Esto resultó especialmente útil para enfermedades desatendidas [29].

En esta estrategia, se parte de un mecanismo de acción conocido y un blanco terapéutico bien caracterizado, que a su vez se encuentra presente en otras especies y que haya ha sido modulado químicamente con éxito en el pasado. Lo que se busca en estos casos no es un reposicionamiento directo: se espera una etapa de optimización para mejorar la actividad o especificidad. Típicamente, como la modulación química está demostrada y ha sido exitosa, los *targets* se reposicionan en conjunto con sus respectivos moduladores químicos, cuya toxicidad, perfil Absorción, Distribución, Metabolsimo y Excreción (ADME) y farmacocinética ya han sido aprobados (o están en vías de serlo) por entidades regulatorias. Esto, como hemos discutido anteriormente, es especialmente conveniente en enfermedaes desatendidas [22]

En adición a las obvias ventajas que ofrece esta estrategia en relación al marco regulatorio, el hecho de que el blanco terapéutcio ya esté caracterizado supone que es posible pensar en ensayos de co-cristalización o, mínimamente, de modelado por homología y docking entre el blanco de

interés y el conjunto de entidades químicas reposicionadas con éste. También, el conocimiento acumulado sobre le blanco en el organismo de interés significa que existen una o más formas de estudiarlo, por lo que montar ensayos bioquímicos es intrínsecamente posible.

2.3.3. Reposicionamiento de clases de *targets*

Aún cuando un proceso celular no ha sido completamente esclarecido, si existe evidencia directa o indirecta de expresión de cierta clase de blancos, se puede reposicionar lo que se conoce de esa familia en otras especies. Así, aunque el blanco no haya sido validado o siquiera identificado fehacientemente, pueden surgir hipótesis de trabajo para búsqueda de nuevos *leads*. Dada la naturaleza de estas hipótesis, la mayoría de los ensayos que surgen del reposicionamiento de clases de targets son fenotípicos. Esto supone, por un lado, la ventaja de poder medir directamente el efecto del compuesto sobre el patógeno. Por el otro, limita la capacidad de optimizar racionalmente los *leads* obtenidos dado que no se conoce con exactitud el mecanismo de acción [22].

Las kinasas han sido el ejemplo de reposicionamiento de clase de target por antonomasia. Esto se debe no solo a que se sabe mucho de ellas [30] (en organismos modelo), sino fundamentalmente al amplio arsenal de inhibidores de kinasas en el mercado [31] y en la literatura en general (+1000) [32]. En tripanosmátidos, en particular, aunque algunas kinasas han sido caracterizadas bioquímicamente, el solo indicio de conservación de algunas de ellas en estos parástitos [33] ha motivado múltiples *screenings* fenotípicos con inhibidores de kinasas humanas [34].

Posiblemente, el ejemplo más emblemático de esto haya sido la identificación Lapatinib, un inhibidor de tirosin-kinasas, como agente tripanocida [35]. Su posterior optimización dio lugar al compuesto NEU-617, con potencia sub-micromolar contra *Trypanosoma brucei*. Desafortunadamente, este compuesto tuvo problemas de biodisponibilidad y toxicidad en ensayos ulteriores.

2.3.4. Reposicionamiento de *leads*

A diferencia de los casos de reposicionamiento mencionados más arriba, cuando se trata de *leads*, lo que se busca probar no son moléculas prácticamente listas para la clínica, sino aprovechar los resultados de *screenings* de alto rendimiento disponibles en literatura. Esto significa que se parte, casi siempre, de bibliotecas sesgadas a cierto blanco (o familia de blancos, según el caso) y con características fisicoquímicas símil-droga; lo que supone una ventaja importante respecto a la búsqueda de fármacos a ciegas (o no sesgadas). Como con el reposicionamiento de familias de blancos, se puede inferir poco sobre el mecanismo de acción posible de estas moléculas, por lo que suelen probarse directamente en ensayos fenotípicos [22].

El reposicionamiento de *leads* es el más laxo y presenta el mayor número de oportunidades para hallar *hits* con capacidad antibiótica pero, como contraparte, precisa de un mayor esfuerzo de optimización y validación en comparación con el resto de casos de reposicionamiento.

Durante el transcurso de este trabajo usaremos "*reposicioamiento de drogas*" para referirnos indistintamente a cualquiera de los enfoques arriba mencionados, salvo expresa necesidad de ofrecer una distinción; aunque por la naturaleza del trabajo aquí expuesto, casi siempre nos estaremos refiriendo a este último tipo de reposicionamiento.

2.4. Enfermedad de Chagas

La enfermedad de Chagas es una enfermedad infecciosa que constituye un problema sanitario, social, y económico de suma importancia en América Latina, reconocida por la Organización Mundial de la Salud como una de las enfermedades tropicales desatendidas más comunes en las zonas de mayor pobreza [36].

Trypanosoma cruzi es el agente etiológico de la enfermedad de Chagas. Esta enfermedad es endémica en la mayor parte de América Central y Sur afectando aproximadamente a 8 millones de personas [37], con un número creciente de casos en Norteamérica [38]. La enfermedad tiene varias consecuencias clínicas y en su forma aguda puede conducir a la muerte (mayormente en niños). En su forma crónica produce patologías asociadas como megacolon, megaesófago, y cardiopatías entre otras. Debido a la falta de un conocimiento más detallado, se asume que estas patologías diferentes son producto de la conjunción de un número de variables: factores ambientales, características genéticas del hospedador y la variabilidad genética presente en las poblaciones de parásitos [39].

No existen vacunas para prevenir la infección y las drogas actualmente utilizadas para el tratamiento (nifurtimox y benznidazol) son tóxicas para el paciente y no son muy efectivas cuando la enfermedad se encuentra establecida. Es por esto que el desarrollo de nuevas drogas efectivas y de baja toxicidad es de gran importancia. Para esto es crucial tener en cuenta que *Trypanosoma cruzi* presenta una alta variación genética, la que podría ser responsable de la sensibilidad y/o resistencia diferencial observada en algunas cepas de la especie a un mismo tratamiento quimioterapéutico [40]. Se ha demostrado también que la resistencia a benznidazol puede ser adquirida por mutaciones en un gen [41].

2.4.1. Agente etiológico

El genoma de *Trypanosoma cruzi*, secuenciado a partir de la cepa CL-Brener fue publicado en 2005 [42], junto con los de otros dos tripanosomátidos de importancia médica: *Trypanosoma brucei* (patógeno causante de la enfermedad del sueño) [43], y *Leishmania major* (patógeno causante de la leishmaniasis) [44].

Aunque la cepa CL-Brener se usa comunmente como referencia, existen en la naturaleza otras cepas de importancia clínica. La tipificación en unidades discretas de estudio (DTUs) es, en la práctica, una clasificación más útil desde el punto de vista eco-epidemiológico y clínico [45]. Aunque la historia evolutiva de estas DTUs está aún en disputa, se reconocen al menos 6 de ellas (TcI-VI); con las Tc I, II, V y VI de presencia eminentemente doméstica en américa latina, y de mayor relevancia clínica en general. Aun con el modelo evolutivo en disputa, hay acuerdo general en que las Tc I y II serían las DTUs ancestrales, mientras que las IV-VI surgieron en uno o más eventos de hibridación entre éstas [46]. La cepa comunmente utilizada como referencia, CL-Brener, surge hipotéticamente de la hibridación entre las DTU TcII y TcIII, afiliándose como TcVI.

El parásito tiene un ciclo de vida que alterna entre hospedadores invertebrados (insectos hematófagos), y vertebrados (mamíferos). Los insectos (de la familia *Triatominae*) funcionan como vector de la enfermedad, transmitiéndola a los humanos. Estos triatominos hematófagos habitan predominantemente en regiones selváticas. Sin embargo, la colonización de regiones originalmente selváticas han provocado la urbanización de vectores naturales, aumentando signicativamente la probabilidad de contagio. En Argentina, la enfermedad es endémica en la

región del Gran Chaco, la cual incluye a las provincias de Chaco, Formosa y Santiago del Estero, y sectores de las provincias de Córdoba, La Rioja, San Juan, Salta, Jujuy y Catamarca.

El ciclo de vida del parásito es complejo (figura 2.3). El parásito prolifera en el tracto gastrointestinal del vector. Este estadío no infectivo se denomina epimastigote. En respuesta a estrés nutritivo se diferencia en el recto a tripomastigotes metacíclicos, que constituyen la forma infectiva. Los tripomastigotes metacíclicos son liberados con las heces cuando el vector se alimenta de la sangre de un mamífero y alcanzan el torrente sanguíneo a través de lesiones en la piel, por ejemplo al rascarse.

Una vez en el torrente sanguíneo, el parásito invade distintos tipos celulares, como células musculares y nerviosas del corazón y del tracto intestinal, al igual que células del sistema retículo-endotelial. Luego de la invasión de la célula blanco el parásito escapa rápidamente de los lisosomas y se diferencia, dentro del citosol, al estadío de amastigote, la forma proliferativa intracelular. Los amastigotes se replican por mitosis y eventualmente se diferencian a tripomastigotes que destruyen la célula huésped y alcanzan el torrente sanguíneo. Los tripomastigotes sanguíneos pueden invadir otras células o ser ingeridos por un vector hematófago apropiado, en donde se diferenciaran nuevamente a epimastigotes, completando así el ciclo.

2.4.2. Vectores

Los triatominos que transmiten al protozoario *Trypanosoma cruzi* pertenecen a la familia *Reduviidae*, orden *Hemiptera*. *Reduviidae* tiene 22 subfamilias, incluyendo *Triatominae*.

Todas las especies de esta subfamilia son capaces de transmitir el parásito a humanos, pero sólo unas pocas especies tienen un ciclo domiciliario extendido y por lo tanto son relevantes en la transmisión de *Trypanosoma cruzi* a humanos. Las tres especies más importantes de vectores en este sentido son *Triatoma infestans*, *Rhodnius prolixus* y *Triatoma dimidiata*. *Triatoma infestans* es el vector primario en las regiones endémicas subamazónicas, *R. prolixus* se encuentra típicamente en Centroamérica y las naciones andinas, mientras que *T. dimidiata* se distribuye en una región similar a la anterior pero también se extiende de forma extensiva en México.

2.4.3. Hospedadores

Más de 100 especies de mamíferos son susceptibles a la infección por *Trypanosoma cruzi* [47, 48] a lo largo de la región geográfica que abarca desde el norte de Argentina hasta el sur de Estados Unidos. Los mamíferos involucrados típicamente en el ciclo selvático de transmisión incluyen armadillos (familia *Dasypodidae*), monos (orden *Primates*), y comadrejas (*Didelphis* spp.) entre otros. Las mascotas, como los perros y gatos, pueden infectarse al ser picados así como al comer presas infectadas o al ingerir insectos infectados. Se ha demostrado que los perros pueden ser un nexo importante en el mantenimiento del ciclo domiciliario y la consecuente transmisión a humanos. Esto sugiere que en zonas endémicas, el descubrimiento de drogas podría tener importancia no sólo para el uso humano, sino también para el uso veterinario, ayudando a controlar la aparición de la enfermedad en el contexto domiciliario.

2.4.4. Tratamiento de la enfermedad de Chagas

El tratamiento de la infección por *Trypanosoma cruzi* se basa en dos agentes quimioterápicos: Nifurtimox (Lampit; Bayer 2502) y Benznidazol (Abarax; ELEA) (figura 2.4), los cuales son

efectivos durante la fase aguda de la enfermedad, pero no durante la fase crónica de la misma. Además resultan inconvenientes debido a su alta toxicidad y la consecuente baja adherencia a los programas de tratamiento por parte de los pacientes [49]. A pesar de su menor efectividad durante fase crónica, hay tendencia a indicar el tratamiento también en estos pacientes [50, 51].

Se cree que el efecto del nifurtimox se debe a que la droga sufre procesos de óxido-reducción que generan especies reactivas tales como radicales superóxido, peróxido de hidrógeno y radicales hidroxilo. Estas especies producen estrés oxidativo que puede dañar el DNA o lípidos de las membranas celulares [52]. Además se reportó que el nifurtimox inhibe a la enzima tripanotiona reductasa, o cual resulta en la inhibición de la formación de tripanotiona. Esta molécula es un tiol

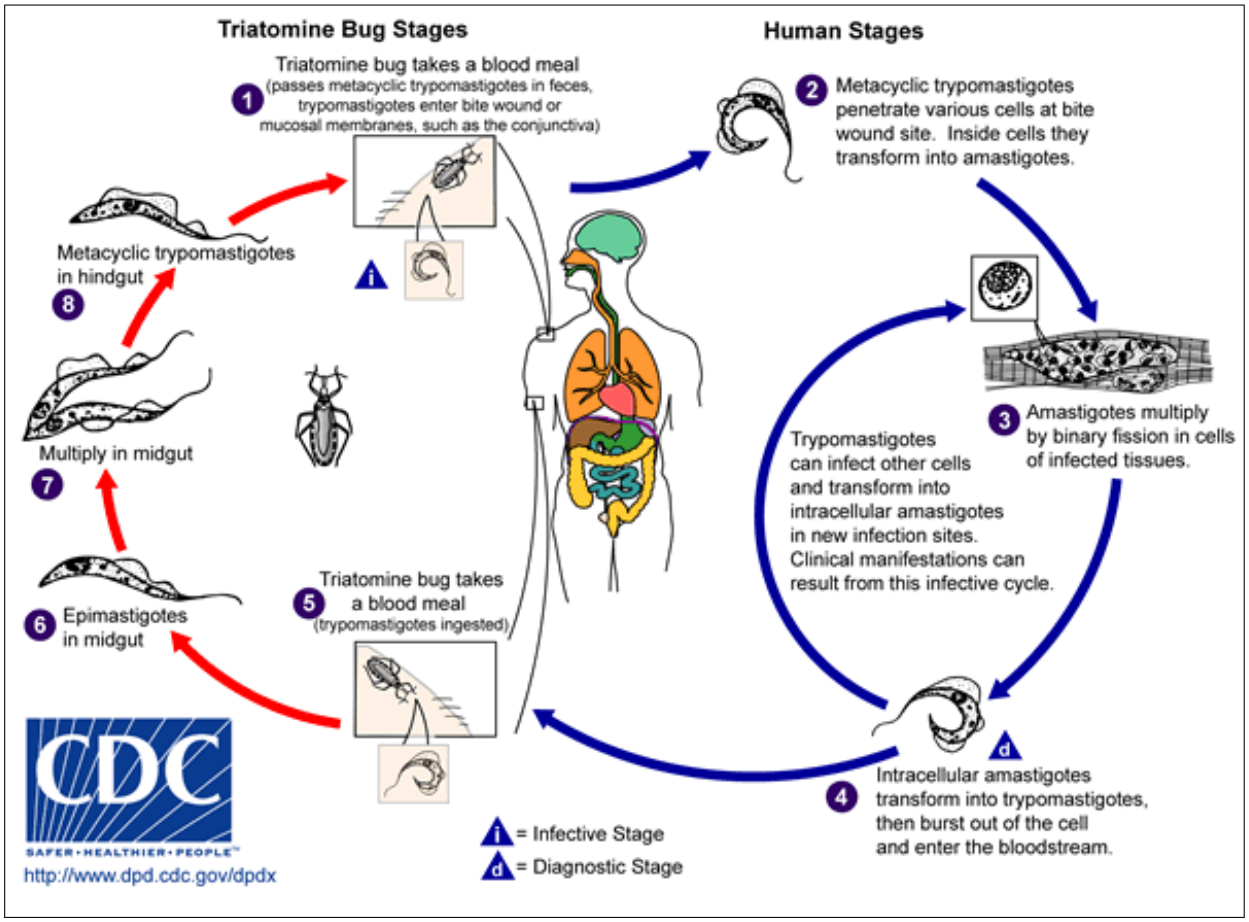


Figura 2.3 – Ciclo de vida de *Trypanosoma cruzi*. Un insecto vector triatomino se alimenta de sangre liberando tripomastigotes en sus heces cerca del sitio de la herida. Los tripomastigotes entran al hospedador a través de la herida o a través del contacto con mucosas, tales como la conjuntiva (1). Las especies de vectores triatominos normalmente pertenecen a los géneros *Triatoma*, *Rhodnius*, y *Panstrongylus*. Dentro del hospedador, los tripomastigotes invaden células cerca del sitio de inoculación, donde se diferencian a amastigotes intracelulares (2). Los amastigotes se multiplican por fisión binaria (3) y se diferencian a tripomastigotes, y luego son liberados a la circulación sanguínea como tripomastigotes (4). Los tripomastigotes infectan células de diversos tejidos y se transforman en amastigotes intracelulares en los nuevos sitios de infección. Las manifestaciones clínicas pueden resultar de este ciclo de infección. Los tripomastigotes sanguíneos no se replican (a diferencia de lo que ocurre en la tripanosomiasis africana). La replicación ocurre sólo cuando el parásito entra en otra célula o cuando son ingeridos por otro vector. El insecto vector se infecta al alimentarse de la sangre de humanos o animales que contienen parásitos circulantes (5). Los tripomastigotes se transforman a epimastigotes en el tracto intestinal del vector (6). Los parásitos se multiplican y se diferencian en el intestino medio (7) y se diferencian a tripomastigotes metacíclicos infectivos en el intestino posterior (8). El parásito *Trypanosoma cruzi* puede ser transmitido también por transfusiones sanguíneas, transplante de órganos, a través de la placenta, o en accidentes en el laboratorio

de bajo peso molecular, exclusivo de tripanosomátidos, implicado en la defensa contra oxidantes, xenobióticos y proteínas regulatorias, y es esencial para la supervivencia del parásito [53].

En cuanto al benznidazol hay estudios que sugieren que el el tratamiento con esta droga no produce la generación de radicales libres, sino que afecta al parásito por otros medios [54]. Estudios más recientes muestran que el benznidazol actúa como una pro-droga, que es metabolizada por una nitroreductasa específica [55], produciendo derivados tóxicos. Estos derivados podrían ser los responsables de la depleción de tripanotiona observada como consecuencia del tratamiento del parásito con esta droga [52]. Aunque el mecanismo no está elucidado completamente y es posible que también se afecten otros componentes celulares.

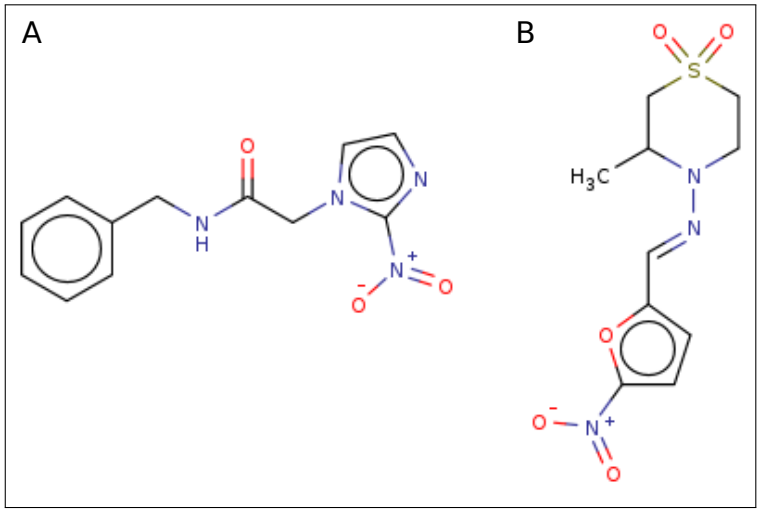


Figura 2.4 – Estructuras de las drogas utilizadas actualmente para el tratamiento de la Enfermedad de Chagas: (A) Benznidazol (Rochagan; Radanil; Abarax) y (B) Nifurtimox (Lampit; Bayer 2502).

2.5. Información biológica: Bioinformática

Aunque podría creerse que la bioinformática es una respuesta a la necesidad de colectar y analizar los datos generados por las nuevas tecnologías de secuenciación, lo cierto es que la disciplina precede largamente al descubrimiento y auge de éstas. Los inicios de la bioinformática ocurrieron hace más de 50 años, cuando las computadoras de escritorio todavía eran una hipótesis y el ADN aún no se podía secuenciar. Los cimientos de la bioinformática se establecieron a principios de los años '60 con la aplicación de métodos computacionales para el análisis de secuencias de proteínas (el ensamblaje de secuencias de novo, las bases de datos de secuencias biológicas y los modelos de sustitución). Más tarde, el estudio del ADN tuvo avances paralelos en (i) los métodos de biología molecular, que permitieron una manipulación más fácil y reproducible del mismo, así como su secuenciación; y (ii) la informática, que vio el surgimiento de computadoras cada vez más miniaturizadas y más potentes, así como el desarrollo de software especializado para manejar tareas de bioinformática. En los años '90 y 2000, las mejoras significativas en la tecnología de secuenciación, junto con la reducción de costos, dieron lugar al aumento exponencial de datos y al surgimiento de lo que conocemos hoy como *Big Data* en Biología [56].

Para alcanzar este nivel de interés y desarrollo, la piedra angular fue la representación de secuencias proteicas como cadenas de caracteres. Para principios de la década del '50, ya se sabía que las proteínas estaban compuestas por secuencias ordenadas de aminoácidos [57, 58], pero no fue sino hasta fines de los '60 que se propuso la notación actual, que asigna a cada aminoácido una letra del alfabeto y representa a las proteínas como una secuencia ordenada de éstos [59, 60]. Margaret Dayhoff, quien ocupara el rol protagónico en la creación de esta notación y en su impulso, desde el desarrollo de una herramienta para ensamblaje de secuencias proteicas; creó también la primera base de datos biológica de la historia [61].

Si bien la forma en la que representamos la proteínas en bioinformática no ha cambiado sensiblemente desde entonces, sí se han desarrollado múltiples métodos que permiten inferir su estructura, ubicación dentro de la célula, su relación filogenética con otras proteínas, función bioquímica e incluso su rol como parte del complejo entramado del metabolismo celular. En esta breve introducción a la bioinformática de proteínas se pondrá foco en los modelos y algoritmos que permiten obtener este tipo de información.

2.5.1. Dominios funcionales y Ontologías

Los dominios funcionales son regiones estructurales y funcionales de las proteínas que se repiten a lo largo de muchas proteínas diferentes y que se consideran como unidades estructurales y funcionales independientes. La identificación de los dominios funcionales es crucial para comprender la función de las proteínas y su papel en la biología celular. Hay varias bases de datos que recopilan información sobre los dominios funcionales de las proteínas, como Pfam, InterPro y PROSITE [62–64], entre otras. Estas bases de datos utilizan diferentes métodos, como alineamiento de secuencias, análisis de estructuras tridimensionales y predicción de dominios basados en la estructura, para identificar y clasificar los dominios funcionales de las proteínas. En la actualidad, la tendencia parece inclinarse hacia meta-predictores que usan múltiples algortimos y ofrecen una respuesta unificada a través de un único portal, como la última versión de InterPro (v88.1) [65]. La detección del mismo dominio en dos o más proteínas permite trazar conexiones que las relacionan funcional o estructuralmente.

Independientemente de su centralización en único recurso, cada algoritmo o modelo de

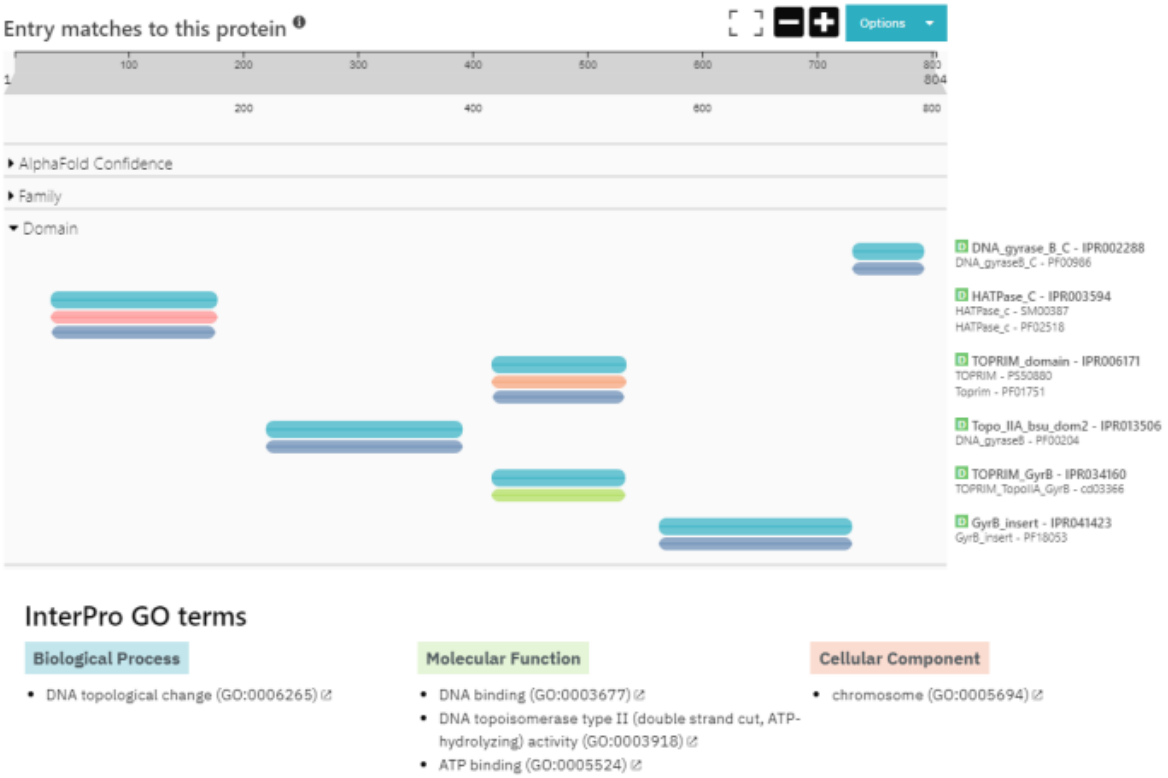


Figura 2.5 – Meta-predictor InterPro v88.1. Vista de ejemplo para la DNA Girasa de *Psuedomona stutzeri* A4VFG1.

predicción utiliza una metodología puntual para realizar sus inferencias. Como puede verse en el ejemplo presentado en la figura 2.5, suele haber acuerdo en la existencia, ubicación y extensión de los dominios funcionales entre distintos predictores. En el ejemplo pueden verse 5 dominios, todos identificados por Pfam (denotados con `PF00986`, `PF02518`, `PF01751`, `PF00204`, `PF18053`) y algunos detectados además por otros predictores, como Prosite [63], SMART [66] y CDD [67]. A su vez, aunque la base de datos InterProSite recopile predicciones para algunas proteínas o proteomas completos solamente, si el organismo de interés no estuviera presente en ésta, es posible realizar predicciones *ad hoc* instalando el software provisto por InterPro.

Otra anotación de gran utilidad en la caracterización de un secuencia proteíca es la identificación de sus ontologías. La ontología de genes es un vocabulario estandarizado que se utiliza para describir los genes y sus relaciones en los organismos vivos. Esta ontología proporciona una forma sistemática de clasificar los genes en función de su rol biológico, su ubicación, su expresión y su interacción con otros genes. El objetivo principal de la ontología de genes es crear una base de datos integrada y completa de información sobre los genes que pueda ser utilizada por los investigadores en biología molecular y genómica. La definición aplica también para proteínas.

La ontología de genes se basa en una ontología que recopila términos del conocimiento de la biología molecular y celular [68] (*Gene Ontology Consortium*). Esta ontología ha sido adoptada por muchas bases de datos biológicas y se utiliza ampliamente en la investigación en biología molecular y genómica. Como puede verse en la figura 2.5, InterProSite incorpora *GO terms* (del inglés, *Gene Ontology Terms*) [69] para referirse a estos vocablos. En el ejemplo provisto, puede verse que la DNA Girasa está involucrada en `GO:0006265` (*DNA Topological change*), a través de distintas funciones denotadas por `GO:0003677` (*DNA Biding*), `GO:0003918` (*DNA topoisomerase type II*), `GO:0005524` (*ATP binding*), fundamentalmente ubicada en `GO:0005694` (*Chromosome*)

2.5.2. Ortología entre proteínas

Como las afiliaciones funcionales, la ortología permite trazar relaciones entre proteínas. La diferencia subyace en que, en lugar de buscar la existencia de dominios compartidos e inferir posibles funciones compartidas entre dos proteínas, la ortología utiliza la conservación o diversificación de una secuencia para establecer si dos proteínas se han conservado a lo largo de la evolución en dos (o más) especies diferentes como resultado de una divergencia o especiación. Es decir, la ortología se produce cuando dos genes homólogos se encuentran en dos especies diferentes que se han separado evolutivamente, y ambos genes han evolucionado independientemente, desde un ancestro común [70]. Es importante notar que dos proteínas pueden ser ortólogas sin compartir todas sus afiliaciones funcionales o dominios estructurales, por lo que estas anotaciones pueden ser complementarias [71]

El proceso de detección de ortólogos es un campo de estudio extremadamente importante para ayudar a mejorar la anotación funcional de varios organismos [72] y sigue siendo muy importante para dilucidar los procesos que dieron lugar a la aparición de especies [73]. El reconocimiento preciso de la ortología es un paso esencial para las investigaciones de genómica comparativa [74].

El análisis de ortología se vuelve complejo cuando se necesitan incluir grandes cantidades de secuencias a ser analizadas [75]. Algunas herramientas consolidadas como BLAST all-vs-all [76], RBBH [77] y OrthoMCL [78] demandan una alta carga computacional, que excede las capacidades del hardware normalmente disponible. Por fortuna, algunos recursos de acceso público ya disponibilizan *clusters* de ortología listos para usar, o proveen de algún servicio de solicitud de cálculo automatizado que permite cargar un proteoma propio, reduciendo así el costo computacional de asignación de *clusters* cuando se busca anotar varios proteomas completos [79].

En este trabajo se utilizó OrthoMCL como fuente de determinación de *clusters* de ortología. Este algoritmo consiste en un cálculo de similitud de tipo *todos contra todos* usando BLASTP, seguido de la identificación de posibles ortólogos a través de la búsqueda de pares de proteínas que son recíprocamente un *best-hit*. Esto significa que, dada una proteína p de un organismo P y una proteína q en un organismo Q, las proteínas p y q son posibles ortólogas si realizando una búsqueda con la secuencia p sobre una base de datos constituída con todas las proteínas de Q, el *mejor hit* obtenido es q, y recíprocamente cuando realizo la búsqueda con q sobre una base de datos que contiene todas las proteínas de P.

2.6. Información química: Quimioinformática

La quimioinformática es un campo de estudio que aprovecha las herramientas informáticas para facilitar el almacenamiento, análisis y manejo de datos relacionados con la química, como las estructuras moleculares, fórmulas, propiedades químicas y metadatos asociados a las moléculas, entre otros aspectos relevantes [80]. Aunque el término quimioinformática se acuñó en 1998, los conceptos centrales que conforman esta disciplina, como el análisis computacional de la relación entre la estructura y la actividad () y la predicción de las propiedades de los compuestos, datan de épocas anteriores [81].

Con el aumento en la escala de los ensayos, surgió la necesidad de contar con bibliotecas de compuestos significativamente mayores, incluso en el rango de millones. Esto llevó a que la disciplina de la quimioinformática se convirtiera en un área de gran importancia en el descubrimiento de fármacos [81]. Para los químicos, los dibujos de las estructuras moleculares son un lenguaje natural. Sin embargo, para procesar esta información en una computadora, se

han desarrollado diversas representaciones y formatos. Algunos identificadores, como SMILES o InChIs, pueden codificar una gran cantidad de información de una molécula química en una cadena de texto corta. Hay otras representaciones que incluyen la información dentro de archivos de texto con un formato definido. Actualmente, la mayoría de los repositorios públicos, como PubChem o ChEMBL, utilizan el formato MOL/SDF para el intercambio de datos. Este formato es análogo al formato FASTA para secuencias, mientras que el formato SDF es similar al formato GenBank/EMBL, ya que incluye anotaciones [82]). Estos formatos se describen con más detalle en las secciones siguientes pero, en lineas generales, consisten en una tabla que describe cada átomo, su conectividad con otros átomos y el tipo de enlace que los une, entre otros datos.

Los identificadores químicos se dividen en dos grandes categorías según su método de generación. La primera categoría incluye identificadores sistemáticos, que se generan algorítmicamente a partir de la estructura química (MOL) y tienen una correspondencia ideal de uno a uno (aunque esto no siempre se cumple). Ejemplos de estos identificadores son SMILES, InChI y la nomenclatura IUPAC. La segunda categoría de identificadores son los no sistemáticos, como los números CAS, que no tienen un significado químico y son asignados por la American Chemical Society para identificar de manera unívoca las moléculas. Otros ejemplos de identificadores no sistemáticos son los identificadores internos de bases de datos y los nombres genéricos o comerciales de las drogas. [82].

2.6.1. Formatos de representación e identificación de moléculas

Para la representación en sistemas computacionales se han desarrollado formatos que permiten la representación de las moléculas en forma de texto, como el formato SMILES, que es una cadena de caracteres que representa la estructura química de una molécula. Además, existe el formato InChI, que es una cadena de texto que representa de manera única una estructura química y su estereoquímica. Otra forma de representación es el formato MOL, el cual consiste en un archivo de texto que contiene información sobre la estructura química, el peso molecular, la fórmula molecular, entre otras cosas. A continuación se describen estos formatos en mayor detalle.

Tablas químicas

Hay diversos formatos de tablas químicas capaces de representar distinta información. En la figura 2.6 se muestran algunos de los formatos posibles. En este trabajo las formas utilizadas son molfile (MOL) y SDF. El archivo MOL puede verse esquematizado en la figura 2.7.

- *Header.* Es el encabezado del archivo, identifica al archivo MOL. Contiene el nombre de la molécula, nombre del usuario, programa, fecha, comentarios e informaciones varias. Ocupa tres líneas en total, las cuales pueden ser líneas en blanco.
- *Ctab.* Es la tabla de conexión, a su vez formada por:
  - *Counts line:* contiene el número de átomos, número de uniones, indicación de quiralidad (0= no quiral, 1=quiral).
  - *Atom block:* contiene el símbolo atómico, diferencias de masa, de carga, estereoquímica, H asociados a cada átomo, y coordenadas espaciales xyz para cada átomo.

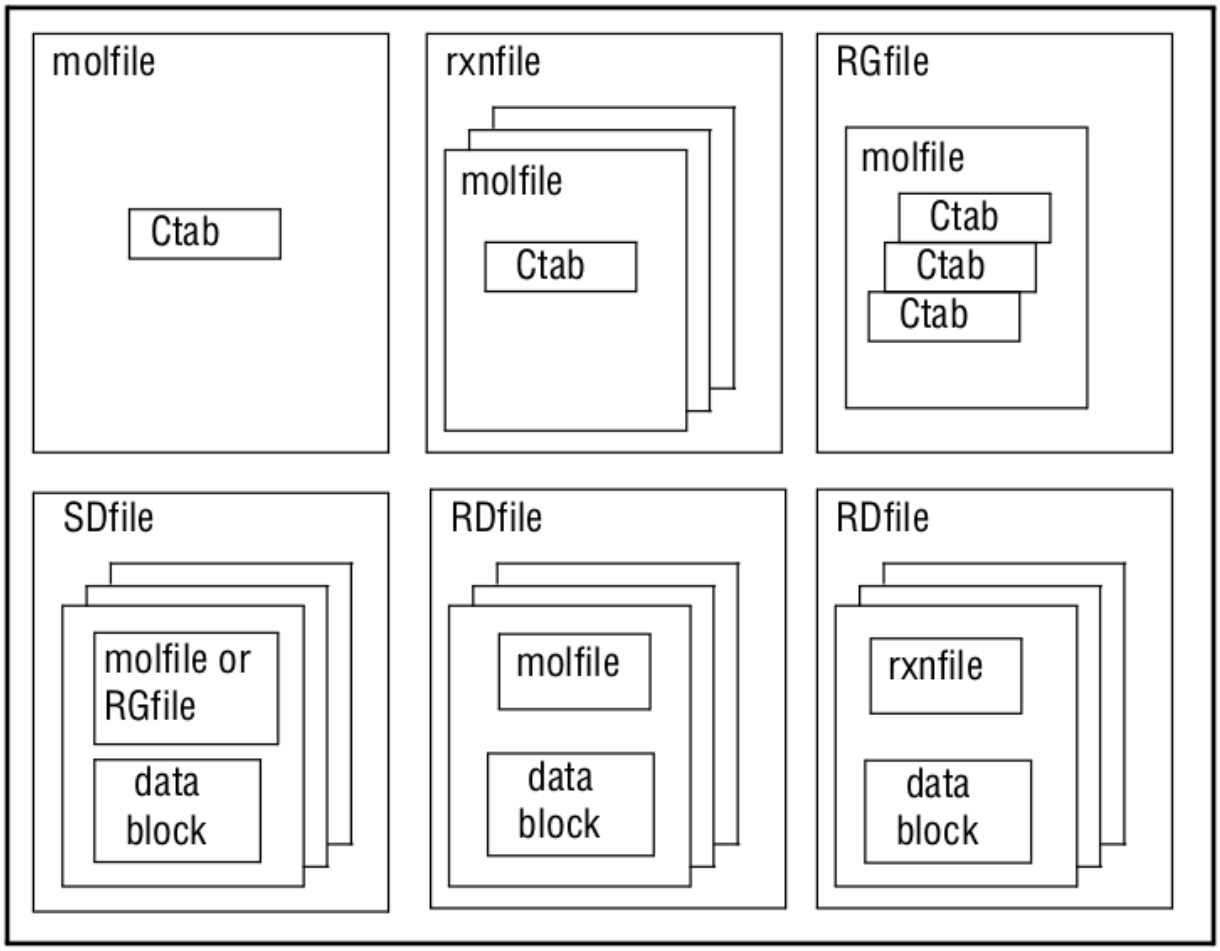


Figura 2.6 – Formatos de representación de moléculas desarrollados por Symyx. Ctab = tabla de conexión, donde se especifica cómo están conectados los átomos, tipo de unión, coordenadas y propiedades [83]

- *Bond block:* contiene el número de los átomos que forman el enlace, el tipo de enlace (simple, doble, etc), y la configuración espacial del enlace.
- *Properties block:* líneas donde se especifican propiedades adicionales, tales como por ejemplo la carga. Cada línea comienza con M y a continuación tres letras, por ejemplo para especificar la carga: **M CHG**. Este bloque termina con **M END**.

El formato SDF contiene, a continuación del MOL, un bloque de datos adicional, que puede contener información de la molécula, por ejemplo, peso molecular, sinónimos, etc. Luego del bloque de datos hay una línea en blanco, y luego los caracteres "$$$$", los cuales indican el final del registro de datos de esa molécula. Un archivo SDF puede contener múltiples registros (ver figura 2.8).

SMILES

SMILES (SMILES) es un formato de representación química que utiliza una cadena de texto para describir la estructura de una molécula de forma simplificada. Este formato fue desarrollado por David Weininger en el *Environmental Research Laboratory* de la Agencia de Protección Ambiental de Estados Unidos (USEPA) en 1987, y luego continuado en Pomona College. La implementación de SMILES fue completada en Daylight Chemical Information Systems (CIS). Una de las ventajas

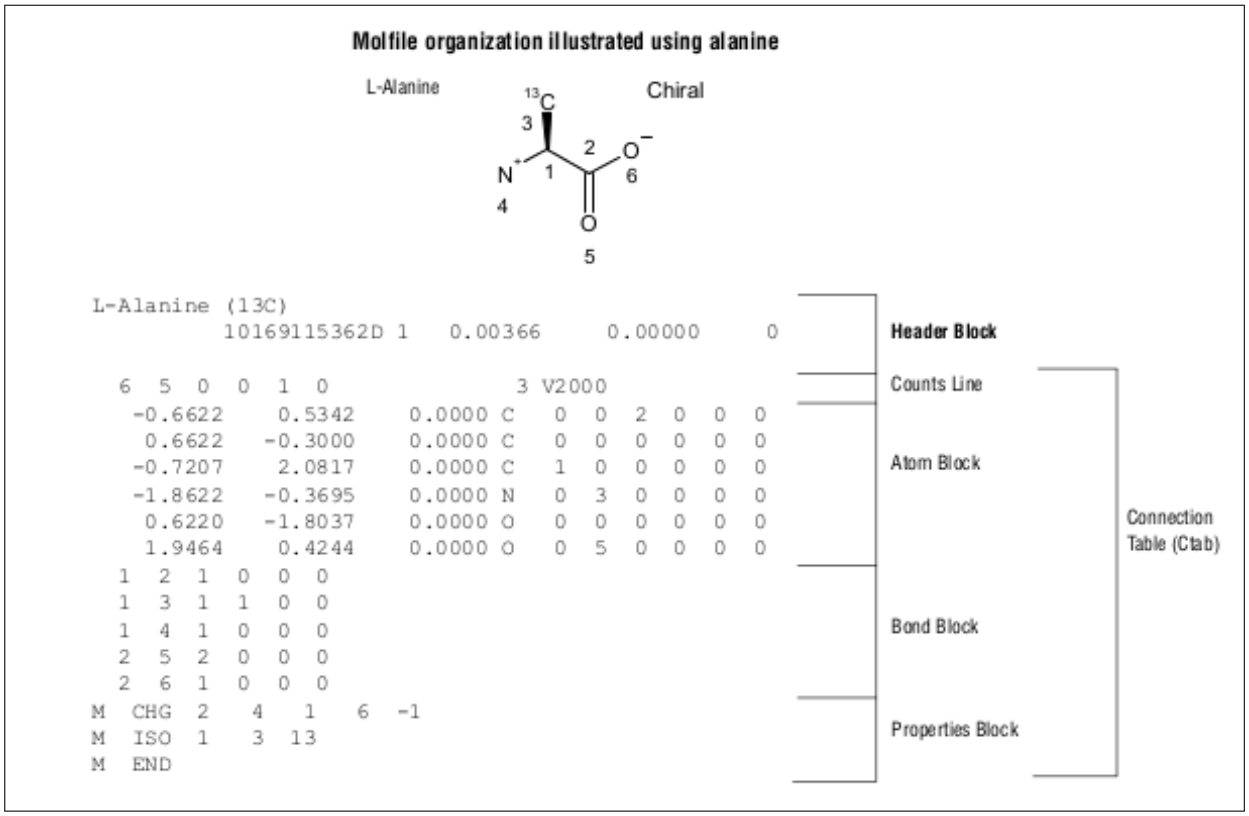


Figura 2.7 – Formato MOL correspondiente a la molécula L-alanina. Adaptado de Figura obtenida de [83])

de SMILES es que permite representar estructuras químicas de manera compacta y fácil de leer, lo que lo hace adecuado para el intercambio de información química en línea. Además, la notación SMILES puede ser convertida de forma eficiente en estructuras moleculares en formato 2D o 3D para su visualización y análisis computacional [84].

La notación de SMILES consiste en una serie de caracteres sin espacios. Los átomos de H pueden estar incluidos (explícitos) o no. Existen 5 reglas genéricas que corresponden a especificaciones de átomos, enlaces, ramificaciones, anillos, y desconexiones; y reglas para especificar distintos tipos de isomería [84].

Átomos. La notación de SMILES representa los átomos por su símbolo atómico encerrados por corchetes ([]). Para los átomos del subconjunto orgánico (**B**, **C**, **N**, **O**, **P**, **S**, **F**, **Cl**, **Br**, **I**), se pueden omitir los corchetes si el número de átomos de hidrógeno (H) se corresponde con el menor valor de valencia consistente con los enlaces explícitos. Los valores normales de valencia para estos átomos son: B (3), C (4), N (3,5), O (2), P (3,5), S (2,4,6), Br (1), F (1), I (1) [84]. Los átomos pertenecientes a anillos aromáticos se especifican en minúscula. ej. carbono alifático: **C**, carbono aromático: **c**).

Enlaces. Los enlaces simples, dobles y triples son representados por los símbolos -, = y # respectivamente. Las uniones simples o aromáticas pueden omitirse; se asume que los átomos adyacentes están conectados por un enlace simple o aromático. Ejemplos:

- Etano: `CC`, `CH3CH3`
- Eteno: `C=C`, `CH2=CH2`
- Etino: `C#C`, `HC\#CH`

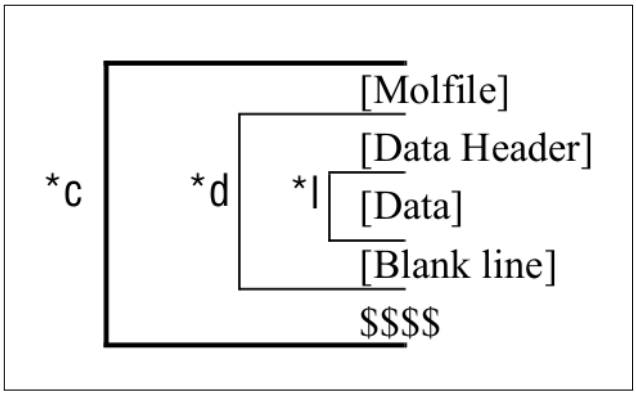
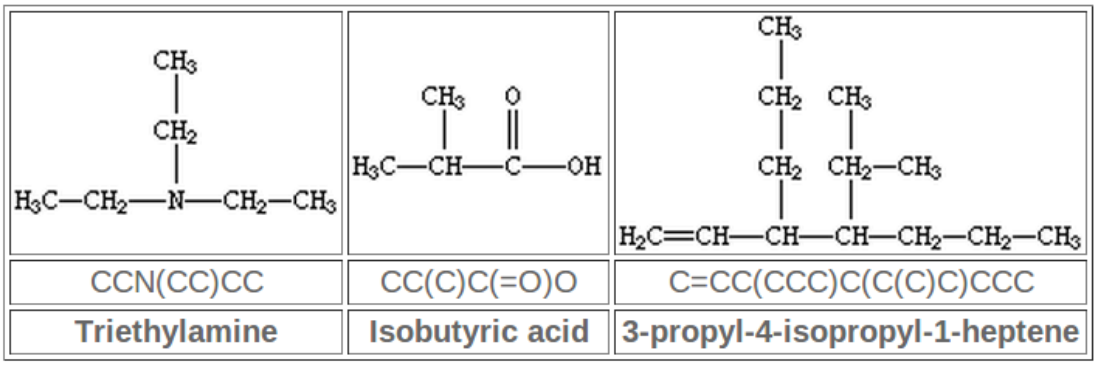


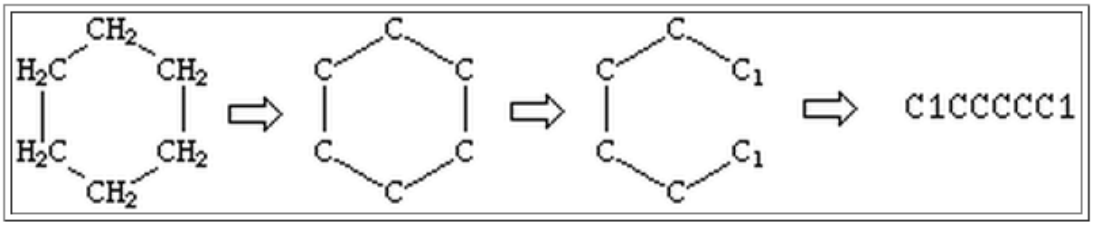
Figura 2.8 – Estructura del formato SDF. *l se repite para cada línea de datos, *d se repite para cada item, y *c se repite para cada compuesto. Molfile: bloque en formato MOL; *Data header*: línea que precede cada item. Comienza con el símbolo >" seguido por un título, por ejemplo: ><melting.point>" [83]

- Fenol: `c1ccccc10`
- Ácido benzoico: `O=C(O)c1ccccc1`

Ramificaciones. En SMILES, se especifican encerrando la estructura ramificada entre paréntesis, con la conexión a la estructura principal hacia la izquierda. Por ejemplo, la estructura de isobutano, un isómero del butano, se puede representar en SMILES como `CC(C)C`. La estructura principal es el grupo de tres átomos de carbono que forman una cadena lineal, y la ramificación se indica con el átomo de carbono entre paréntesis, que está unido al segundo átomo de carbono de la estructura principal mediante un enlace simple.



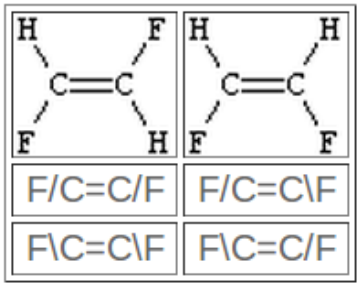
Estructuras cíclicas. Las estructuras cíclicas se representan rompiendo una unión en el anillo y numerando las uniones en cualquier orden. Las uniones que abren el anillo se designan con un dígito inmediatamente después del símbolo atómico correspondiente.



Estructuras desconectadas. Las estructuras desconectadas se escriben por separado y se las conecta con un punto ".". El orden en el cual se listan es arbitrario. Por ejemplo, el acetato de sodio puede expresarse como `[Na+].CC([O-])=O`

Especificación de isótopos. Los isótopos se especifican precediendo el átomo con el número de masa atómico. Siempre se especifica entre corchetes. Por ejemplo, el ^{13}C se escribiría como `[13C]`.

Configuración de dobles enlaces. Para los dobles enlaces, la notación SMILES utiliza los símbolos "\" y "/", que indican la dirección relativa entre los átomos conectados por el doble enlace. El símbolo "\" se utiliza para indicar una orientación hacia la izquierda y hacia abajo, mientras que el símbolo "/" indica una orientación hacia la derecha y hacia arriba. Estos símbolos tienen sentido solo si ocurren en ambos átomos conectados por el doble enlace. La configuración específica se determina por la regla de Cahn-Ingold-Prelog, la cual asigna una prioridad numérica a los átomos conectados en los extremos del doble enlace.



Configuración de centros tetraédricos. La configuración de centros tetraédricos se indica utilizando el símbolo "@" o "@@" inmediatamente después del átomo quiral. El símbolo "@" indica que los átomos siguientes se enumeran en sentido antihorario, mientras que la notación "@@" indica sentido horario [84].

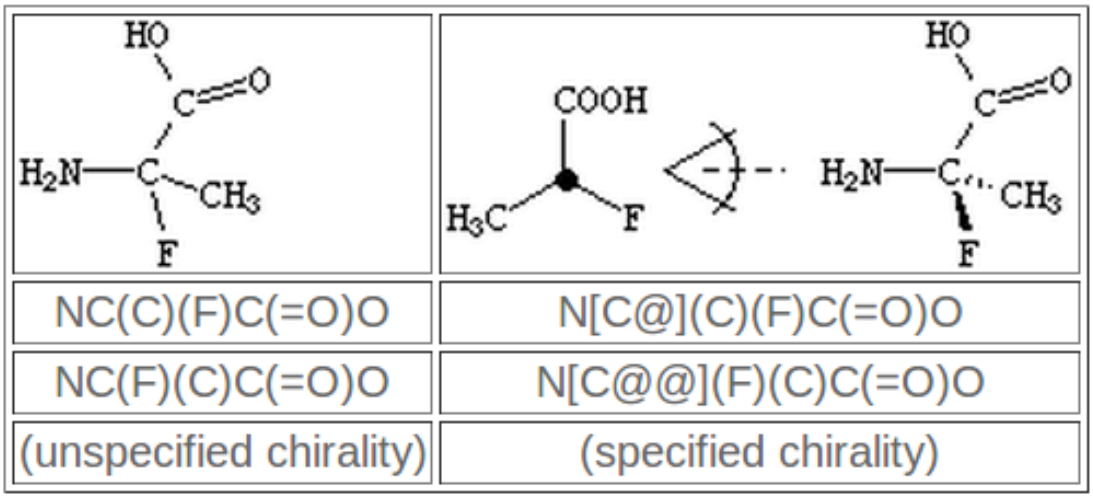


Figura 2.9 – Ejemplos de moléculas y sus representaciones SMILES. Este ejemplo y los mostrados en estas páginas (Adaptado de Daylight [84])

Existen distintos programas que generan el SMILES de formas distintas, con lo cual pueden existir distintos SMILES para una misma molécula [85].

InChI

Debido a la existencia de diversas versiones de SMILES y a la propiedad de los programas que los generan por parte de Daylight, la Unión Internacional de Química Pura y Aplicada (IUPAC) decidió en el año 2000 desarrollar un identificador de compuestos químicos que fuera de uso libre y no propietario para ser utilizado en medios digitales. Este identificador se llamó InChI (International Chemical Identifier). Poco más tarde, no obstante, Blue Obelisk desarrolló en 2007 OpenSMILES, como parte de una iniciativa *open access* con las mismas prestaciones que el SMILES de Daylight. Sin embargo, el InChI encontró asidero en otro tipo de usos.

El InChI es un identificador único de sustancias químicas que se genera a partir de la estructura química de la molécula, representada mediante su tabla de conexión. Este consiste en una serie de caracteres que identifican una sustancia química de manera unívoca, independientemente de cómo se dibuje la molécula.

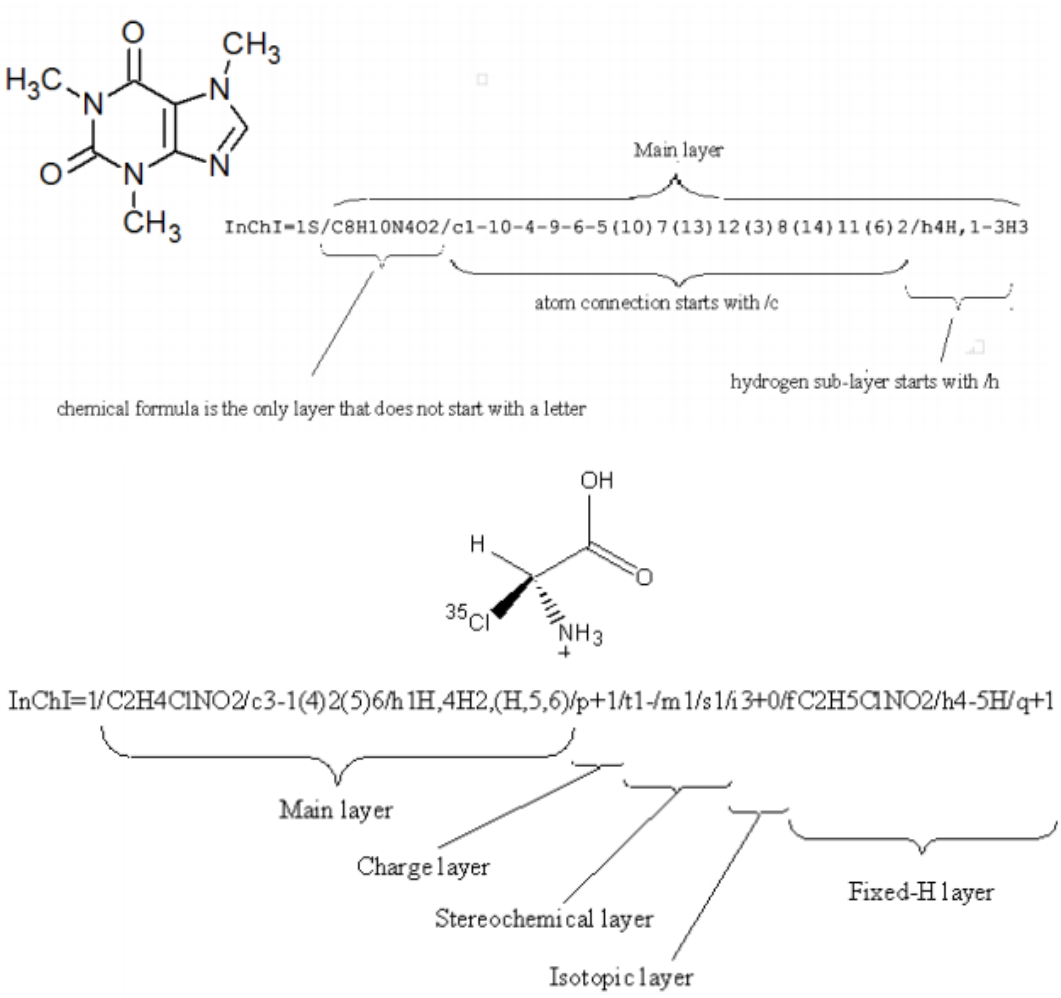


Figura 2.10 – Las distintas capas de información contenidas en el identificador InChI. (Figura tomada de [86].)

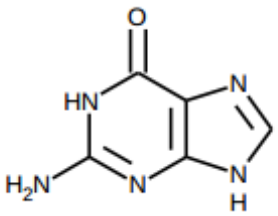
El identificador InChI está compuesto por varias capas que describen diferentes aspectos de la estructura química (figura 2.10). La primera capa, siempre generada, representa la **fórmula química** convencional. Las siguientes capas se generan sólo si la estructura de partida contiene la información correspondiente. En la segunda capa, llamada **capa de conexiones**, se definen las uniones covalentes entre los átomos de la molécula y se subdivide en tres posibles capas

que representan las uniones que no involucran átomos de hidrógeno, las uniones de átomos inmóviles de hidrógeno y las uniones de hidrógeno móviles. En la tercera capa, llamada **capa de cargas**, se especifica la carga neta de la molécula. La cuarta capa, **capa estereoquímica**, permite generar identificadores diferentes para estereoisómeros y se compone de dos subcapas: la primera corresponde a **uniones dobles con hibridación sp^2** y la segunda a **estereoquímica tetraédrica de sp^3 y alenos**. La quinta capa, **capa isotópica**, permite especificar información isotópica para átomos de la molécula. Por último, la sexta capa, llamada **capa de H fijos**, está presente en InChIs no estándar cuando se detectan hidrógenos potencialmente móviles y se requiere fijarlos. El InChI comienza con un prefijo que indica la versión del software utilizado para generarla, seguido de la letra S si se trata del InChI estándar, y cada capa se separa por un carácter "/".

A partir de un InChI se puede obtener un identificador más corto, conocido como InChIKey, utilizando la función criptográfica no reversible SHA-256 del *National Institute of Standards and Technology* de los Estados Unidos. El InChIKey consta de una cadena de 27 caracteres de longitud fija, lo que lo hace mucho más corto que el InChI, que varía en longitud y tiene un promedio estimado de alrededor de 146 caracteres para una colección representativa de 10,000 moléculas.

La capacidad de condensar una cadena de longitud variable, como el InChI, en una de longitud fija es extremadamente útil en aplicaciones de búsqueda en bases de datos o páginas web. Actualmente, todos los repositorios de información química, incluyendo Wikipedia, utilizan InChIs e InChIKeys. Recientemente, Google ha comenzado a indexar moléculas químicas utilizando InChIKeys, lo que permite realizar búsquedas altamente específicas de páginas web que contienen referencias a una molécula específica. Como ejemplo, una búsqueda en Google de la palabra "atorvastatin" puede identificar más de un millón de páginas web. Sin embargo, una búsqueda utilizando el InChIKey correspondiente a esta droga ("`XUKUURHRXDUEBC-KAYWLYCHSA-N`") puede reducir la búsqueda a aproximadamente mil páginas, lo que simplifica enormemente la tarea del usuario de encontrar información relevante [87].

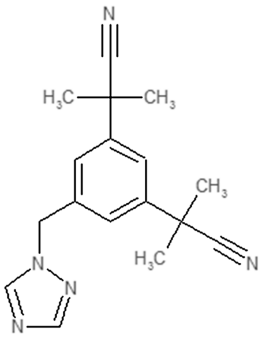
El InChIKey consta de dos partes: el primer bloque es siempre el mismo para un mismo esqueleto molecular, mientras que el segundo bloque contiene información sobre sustituciones isotópicas, cambios en la configuración estereoquímica, tautomería y protonación. Debido a la existencia de InChIs estándar y no estándar, los InChIKeys generados a partir de diferentes InChIs pueden ser diferentes (ver figura 2.11) [86, 88, 89].




```
InChI=1/C5H5N5O/c6-5-9-3-2(4(11)10-5)7-1-8-
3/h1H,(H4,6,7,8,9,10,11)/f/h8,10H,6H2

InChIKey=UYTPUPDQBNUYGX-GSQBSFCVNA-N

InChI=1S/C5H5N5O/c6-5-9-3-2(4(11)10-5)7-1-8-3/h1H,(H4,6,7,8,9,10,11)

InChIKey=UYTPUPDQBNUYGX-UHFFFAOYSA-N
```

Figura 2.11 – Ejemplos de InChIKey estándar y no estándar generados a partir de la misma molécula. Arriba: dibujo de la molécula. Abajo: dos InChIs con sus correspondientes InChIKeys. El primer InChI es el no estándar (Figura tomada de la documentación del InChI [86].)

Anastrozole



SMILES
CC(C)(C#N)c1cc(cc(c1)C(C)(C)C#N)Cn2cncn2
CC(C)(C#N)c1cc(Cn2cncn2)cc(c1)C(C)(C)C#N
CC(C)(C#N)c(cc(cc1C[n]([n]c[n]2)c2)C(C)(C)C#N)c1

IUPAC
2-[3-(1-cyano-1-methyl-ethyl)-5-(1,2,4-triazol-1-ylmethyl)phenyl]-2-methyl-propanenitrile
2,2'-[5-(1H-1,2,4-triazol-1-ylmethyl)benzene-1,3-diyl]bis(2-methylpropanenitrile)
2-[3-(1-cyano-1-methylethyl)-5-(1H-1,2,4-triazol-1-ylmethyl)phenyl]-2-methylpropanenitrile

InChI
InChI=1S/C17H19N5/c1-16(2,9-18)14-5-13(8-22-12-20-11-21-22)6-15(7-14)17(3,4)10-19/h5-7,11-12H,8H2,1-4H3

Figura 2.12 – Distintas formas de identificar una misma molécula. Dado que los identificadores SMILES y IUPAC pueden ser generados por distintos programas, el identificador resultante puede no ser siempre el mismo para una misma molécula en distintas bases de datos. Esto no ocurre con el InChI. Figura tomada de [85].

2.6.2. Algoritmos convencionales y modelos predictivos

Predicción de propiedades químicas

Las drogas son moléculas pequeñas que suelen tener un peso molecular menor a 1,000 Da, lo que comparten con metabolitos naturales y toxinas. Por lo general, actúan en blancos específicos en las células al unirse a receptores, enzimas o transportadores, y modificar su funcionamiento. Para que esto suceda, la droga debe viajar a través del organismo hasta llegar al tejido donde ejerce su efecto. Los estudios de farmacocinética de una droga buscan determinar su destino y las modificaciones que sufre desde que entra al organismo. Este proceso se puede dividir en cuatro fases: absorción, distribución, metabolismo y excreción (ADME).

En el caso de una droga administrada por vía oral, el primer paso es la absorción adecuada desde las paredes intestinales hacia el sistema circulatorio y su transporte al hígado, donde puede ser modificada por enzimas hepáticas. Algunas moléculas son metabolizadas y otras son excretadas. Si la molécula supera este paso metabólico, entrará en la circulación arterial y será distribuida por todo el organismo, incluyendo el tejido donde ejerce su acción. Una vez que la droga cumple su función, debería ser eliminada del organismo para evitar problemas de bioacumulación. Además, una droga no debería causar efectos tóxicos colaterales [90].

Algunas de estas propiedades fisicoquímicas que pueden afectar la biodisponibilidad oral de una droga incluyen su solubilidad en agua y su lipofilicidad. Una baja solubilidad en agua puede dificultar la disolución y la absorción de la droga en el tracto gastrointestinal, mientras que una alta lipofilicidad puede aumentar la afinidad de la droga por los tejidos grasos y disminuir su disponibilidad en la circulación sanguínea. Otros factores importantes pueden ser la estabilidad química de la droga en el tracto gastrointestinal y su capacidad para atravesar las barreras biológicas, como la membrana celular.

Para superar estos desafíos, se deben optimizar las propiedades fisicoquímicas de las drogas, por ejemplo mediante la modificación de grupos funcionales, la conjugación con otros compuestos, o la incorporación de sistemas de administración especiales como nanopartículas o formulaciones en microemulsión. Estos enfoques pueden mejorar la biodisponibilidad oral de las drogas y aumentar su eficacia terapéutica.

Se han podido establecer correlatos entre las propiedades químicas de las drogas y su farmacocinética y biodisponibilidad. Las propiedades más utilizadas para caracterizar compuestos son la solubilidad, hidrofilicidad, hidrofobicidad, lipofilicidad, volumen de la molécula, la extensión de superficie polar de una molécula y la presencia de grupos reactivos. Estas propiedades se pueden medir experimentalmente por métodos directos (e.g. solubilidad), o indirectos – por ejemplo la lipofilicidad se estima en base a la medición del coeficiente de partición de una molécula entre octanol y agua (denominado "logP"). Para muchas de estas propiedades existen predictores computacionales. Incluso características más complejas, como el metabolismo hepático de una droga (que puede variar, afectado por la diversidad genética natural de los genes de los distintos citocromos P450 humanos) pueden ser modeladas por métodos como las redes neuronales, basados en aprendizaje automático [91–93].

Tradicionalmente, la predicción de propiedades químicas ha sido utilizada en la etapa de optimización de moléculas *leads* en el diseño de fármacos (ver Figura 2.2). Sin embargo, cada vez se utilizan más en etapas previas y exploratorias debido a los altos costos de los estudios pre-clínicos. En estas etapas, la predicción de propiedades ADMET es crucial para priorizar un pequeño número de compuestos que luego son sintetizados y evaluados *in vivo* [90].

El estudio de estas propiedades ha llevado a diferentes autores a definir reglas empíricas para evaluar la biodisponibilidad de una molécula. Entre ellas, destacan las reglas de Lipinski, que surgieron de la observación de similitudes en las propiedades fisicoquímicas de moléculas con buena biodisponibilidad oral. Christopher Lipinski las describió en 1997 y posteriormente las publicó en su obra [94], tras analizar 2,200 drogas del *World Drug Index*. Estas reglas son generales y válidas para la mayoría de las drogas consideradas en el estudio, y se conocen popularmente como *rule of five* (RO5), debido a la frecuencia con la que aparece este número o sus múltiplos en las reglas. En términos generales, las reglas sugieren que la absorción o facilidad para permear de una droga administrada por vía oral se ve obstaculizada cuando:

- el coeficiente de partición octanol/agua (LogP) >5
- el peso molecular >500
- el número de grupos dadores de H >5
- el número de aceptores de H >10

Las reglas de Lipinski implican que una droga debe tener un equilibrio adecuado entre las propiedades hidrofílicas e hidrofóbicas. La solubilidad en agua y la hidrofobicidad son dos características fisicoquímicas que tienen una gran influencia en las propiedades de una molécula como droga [90]. Estudios posteriores han ampliado o modificado las reglas de Lipinski, con el objetivo de mejorar las predicciones del potencial de diversos compuestos químicos. Por ejemplo, en un estudio realizado por Veber y colaboradores [95], se proponen los siguientes criterios mejorados para predecir si un compuesto tiene propiedades similares a las de las drogas (conocido como *drug-likeness*):

- logP en el rango -0.4 a +5.6
- refractividad molar entre 40 y 130
- peso molecular entre 180 y 500
- número de átomos entre 20 y 70
- número de enlaces rotables ≤ 10
- extensión de superficie polar ≤ 140 $Å^2$

En la práctica, sin embargo, las reglas empíricas duras como la de Lipinski o sus variantes, se reconocen útiles en algún estadío del desarrollo pero de desaconsejan efusivamente para *early-stage drug discovery*, dado que reducen enormemente el espacio de búsqueda y sesgan el desarrollo de *leads* a un espacio químico prácticamente agotado. Los compuestos derivados de extractos naturales, por ejemplo, suelen violar este tipo reglas, así como buena parte del arsenal de antibióticos de origen fúngico.

Es común que las sustancias tengan efectos nocivos en el cuerpo humano, por lo que, además de evaluar su biodisponibilidad, es importante contar con métodos confiables para medir su toxicidad [90]. Sin embargo, los estudios de toxicidad a gran escala suelen ser costosos y lentos, lo que hace impracticable realizarlos para todos los compuestos candidatos a convertirse en drogas. Una alternativa es establecer ensayos estándar *in vitro* o *in vivo* en sistemas modelo relevantes para evaluar la seguridad de la droga, o utilizar modelos computacionales que intenten predecir la toxicidad a partir de la estructura molecular de la sustancia [90, 96].

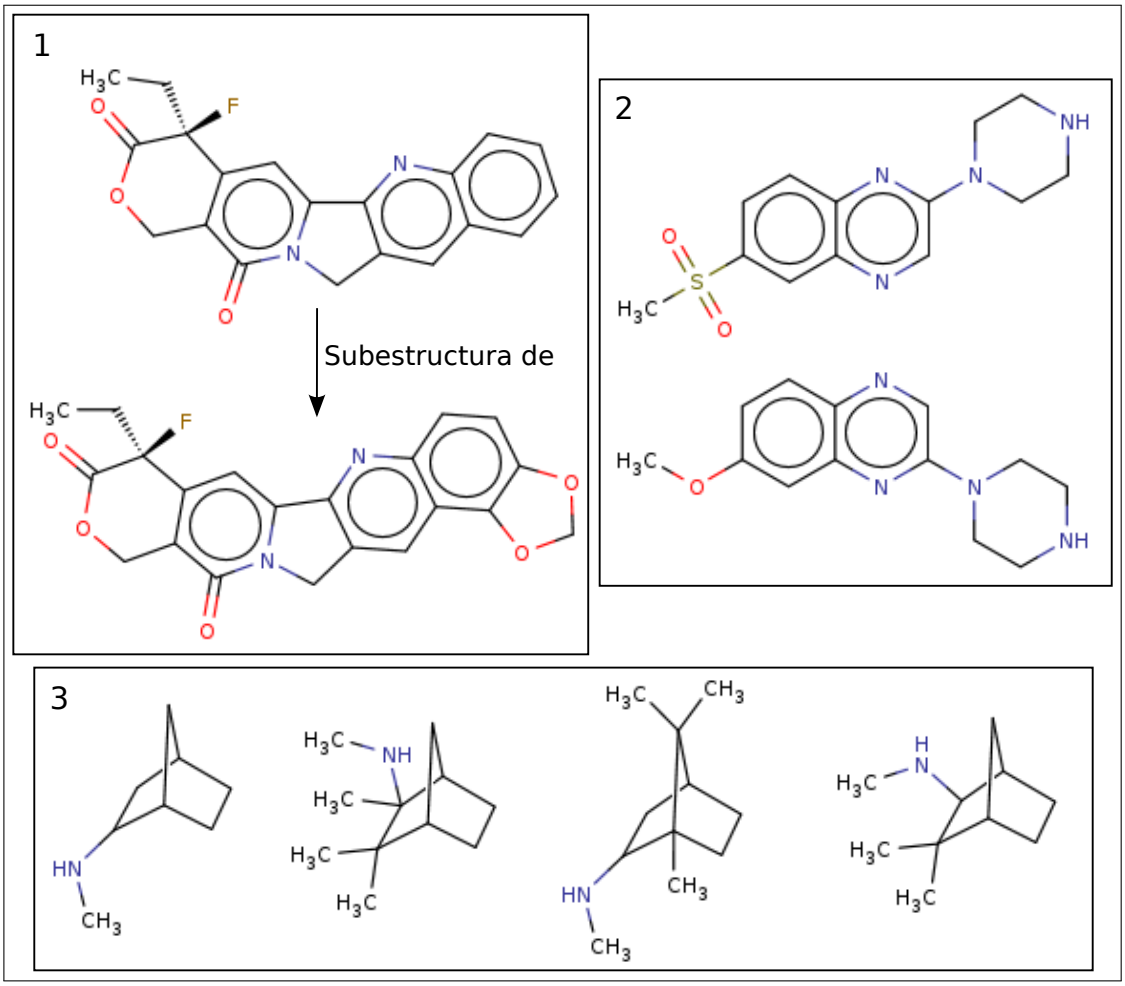


Figura 2.13 – Ejemplos de moléculas similares entre sí. (1) Dos moléculas con alta similitud, donde además se cumple que la molécula de arriba es una subestructura (está totalmente contenida) en la de abajo. (2) Dos moléculas con alta similitud que comparten una subestructura en común, pero ninguna está completamente contenida en la otra. (3) Moléculas con distintos grados de similitud estructural.

Similitud estructural entre moléculas

Los métodos de diseño racional de drogas y la priorización de compuestos se basan a menudo en el principio de que los compuestos con estructuras similares son probablemente similares en cuanto a sus propiedades. Este concepto de similitud molecular es ampliamente utilizado en la literatura química y es especialmente importante en la química medicinal, donde el bioisosterismo permite el intercambio de subestructuras similares para mantener cierto grado de actividad. Los métodos de similitud también son utilizados en la predicción de propiedades relacionadas con la toxicidad o la seguridad de las drogas [97]. La figura 2.13 muestra diferentes ejemplos de similitud estructural.

El aumento de la cantidad de compuestos en las bases de datos, gracias a los avances tecnológicos en la síntesis química y en los ensayos a gran escala, ha incrementado la popularidad de los métodos basados en similitud química para predecir propiedades fisicoquímicas o posibles interacciones. Esto se ha visto necesariamente acompañado por el desarrollo de métodos computacionales para buscar y seleccionar compuestos en bases de datos usando la similitud como punto de entrada [97].

Cuantificación de similitud entre moléculas

La precisión necesaria para cuantificar la similitud entre moléculas requiere una definición más exacta del concepto de similitud estructural. Existen diversas definiciones posibles, como la similitud en la forma y/o volumen de la molécula, o la similitud de los grafos utilizados para representar las moléculas, entre otras [98].

En este trabajo, se considerará la similitud estructural como la similitud entre los grafos. Es decir, la similitud entre el número, tipo y conectividad de los átomos que componen la molécula. Para calcular la similitud y buscar subestructuras en las bases de datos, se utilizan "fingerprints", que son una forma abstracta de representar una molécula mediante un vector binario [84, 99].

Encontrar subestructuras en una base de datos es un problema que pertenece a la clase NP-completo, lo que significa que no existen algoritmos deterministas que puedan solucionarlo en tiempo polinómico (razonable). Por lo tanto, todas las soluciones disponibles tienen que utilizar estrategias de fuerza bruta, lo que hace que la búsqueda sea computacionalmente costosa y no apropiada para bibliotecas grandes de compuestos o bases de datos. Para superar esta limitación, se emplean algoritmos que pueden detectar la ausencia de una subestructura con una confianza del 100 %, lo que requiere mucho menos esfuerzo computacional, y luego buscan la presencia de la subestructura con una confianza menor en un subconjunto reducido de moléculas. De esta forma, se reduce significativamente el tiempo total requerido para resolver el problema [84].

Para buscar una subestructura específica en una base de datos de moléculas, se debe comparar un patrón con cada una de ellas, lo que puede ser un proceso costoso en términos computacionales. Para acelerar este proceso, se pueden precalcular algunas búsquedas que responden a preguntas específicas. Por ejemplo, se puede precalcular la fórmula molecular de cada compuesto y luego compararla con la fórmula molecular de la subestructura que se está buscando, descartando de esta manera cualquier compuesto que no contenga algún elemento de la subestructura buscada [84].

Existe otra opción para descartar de forma rápida moléculas que no contengan la subestructura buscada, que implica el uso de un tipo de "fingerprint" llamados"claves estructurales" o "structural keys". Estas *keys* se representan mediante vectores binarios en los que cada posición representa una característica estructural específica, y la presencia o ausencia de dicha característica se representa mediante `1` o `0`, respectivamente (véase la Figura 2.14). Algunas de las características que podrían estar representadas en una *key* estructural son:

- Presencia/ausencia de cada elemento o, si un elemento es muy común (por ejemplo nitrógeno), podrían usarse distintos *bits* para representar "al menos 1 N", "al menos 2 N", etc.
- Presencia de grupos funcionales de importancia, como alcoholes, aminas, carboxilos, etc.
- Grupos funcionales de interés para una aplicación o base de datos particular. Por ejemplo en una base de datos de moléculas organo-metálicas puede haber *bits* asignados para grupos funcionales que contengan metales; en otros casos podrían haber *bits* designados para esqueletos carbonados como los de esteroides o barbitúricos, etc.

En los primeros sistemas de quimioinformática, los vectores solían ser pequeños y listaban solo unas pocas decenas de patrones estructurales seleccionados por expertos. Sin embargo, en la actualidad, la tendencia es construir vectores más largos. Dado que el espacio químico es enorme (alrededor de 10^{60} moléculas orgánicas pequeñas [100, 101]), anotar la presencia o ausencia de

todos los posibles patrones estructurales resultaría en vectores binarios excesivamente largos. Además, en términos prácticos, aunque el número total de patrones sea grande, el número de patrones efectivamente presentes en cualquier molécula pequeña está acotado por definición.

Una limitación adicional de este sistema es que se requiere tomar una decisión previa sobre cuáles características estructurales son importantes y, por lo tanto, deben incluirse en el vector. Esta selección puede carecer de generalidad, ya que la elección de los patrones a representar varía según la naturaleza de la base de datos en cuestión.

Los *hashed fingerprints* o *fingerprints* cifrados son una alternativa para representar moléculas en la cual se eliminan los patrones predefinidos (ver figura 2.15). Un *hashed fingerprint* también es una cadena de *bits*, pero a diferencia de las *structural keys*, cada *bit* no tiene asignado un significado específico. Para generar un *fingerprint* se examina la molécula y se genera un patrón para cada átomo, así como para cada uno de sus vecinos inmediatos (incluyendo el enlace entre ellos), y para cada grupo de átomos y enlaces conectados por caminos de hasta una determinada longitud. Por ejemplo, la molécula OC = CN generaría los siguientes patrones:

- Caminos de 0 enlaces: C, O, N
- Caminos de 1 enlace: O – C, C = C, C#N
- Caminos de 2 enlaces: O – C = C, C = C – N
- Caminos de 3 enlaces: O – C = C – N

Se producen todos los patrones posibles (de hasta la longitud especificada) exhaustivamente. Dado que no hay patrones predefinidos y la cantidad de patrones posibles es muy grande, no es posible asignar un *bit* para cada patrón. En su lugar, se utiliza cada patrón como semilla para generar un número pseudoaleatorio que generalmente resulta en un conjunto de 4 o 5 *bits* (en la figura 2.15 se muestran 3 *bits*) en una cadena de longitud fija. Estos conjuntos de *bits* se agregan al *fingerprint* de la molécula utilizando una operación de OR lógica. Debido a que cada conjunto de *bits* se produce a partir de un generador de números pseudoaleatorios, es probable que estos conjuntos de *bits* se solapen. Al igual que con las *structural keys*, cada *bit* presente en una subestructura también estará presente en la molécula que la contiene. Todo esto implica que un *fingerprint* puede garantizar con un 100 % de certeza que una subestructura determinada está

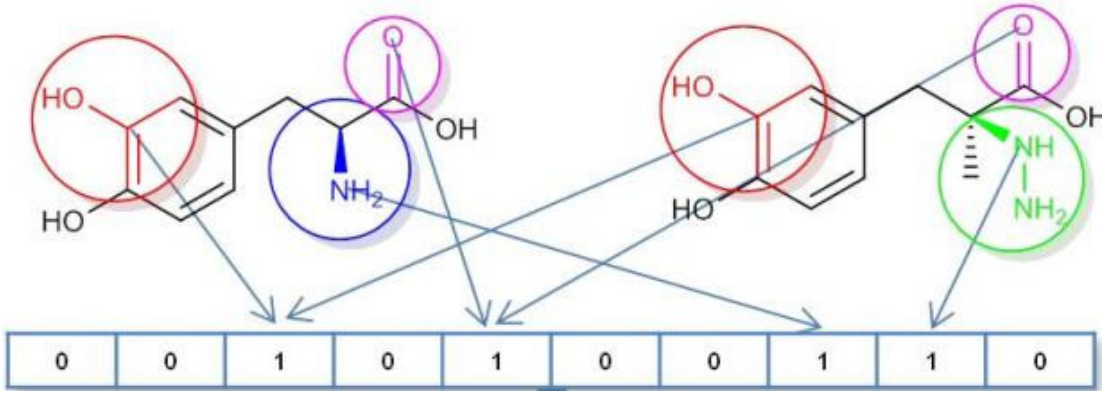


Figura 2.14 – Ejemplo esquemático de un vector binario conteniendo información sobre la presencia (1) o ausencia (0) de fragmentos químicos (subgrafos) definidos. Los patrones o fragmentos seleccionados se marcan con círculos de distintos colores (Tomado de ChemAxon).

ausente en una molécula. Sin embargo, la presencia de una subestructura se asegura con cierta probabilidad [84]

Fingerprints comprimidos

Una vez que se ha generado el *fingerprint* para una molécula, éste puede ser comprimido en un vector binario de tamaño fijo mediante una técnica de plegado o *folding* en la que se realiza una operación de "plegado" del vector sobre sí mismo, como se muestra en la Figura 2.16. El resultado del plegado es un vector binario de tamaño predefinido (usualmente de 512 o 1024 bits) que contiene la unión (operador OR lógico) de todos los subvectores de tamaño del módulo.

La técnica de plegado se basa en la aritmética modular, y su objetivo es distribuir la información contenida en el *fingerprint* de forma equitativa en todo el vector binario comprimido. Para ello, se divide el vector binario en subvectores de tamaño igual al módulo (que suele ser un número primo), y se aplica una función de plegado que consiste en sumar los subvectores módulo el módulo, de manera que los elementos del vector resultante contienen la unión de los elementos de los subvectores.

El vector resultante de la operación de plegado se utiliza como representación comprimida de

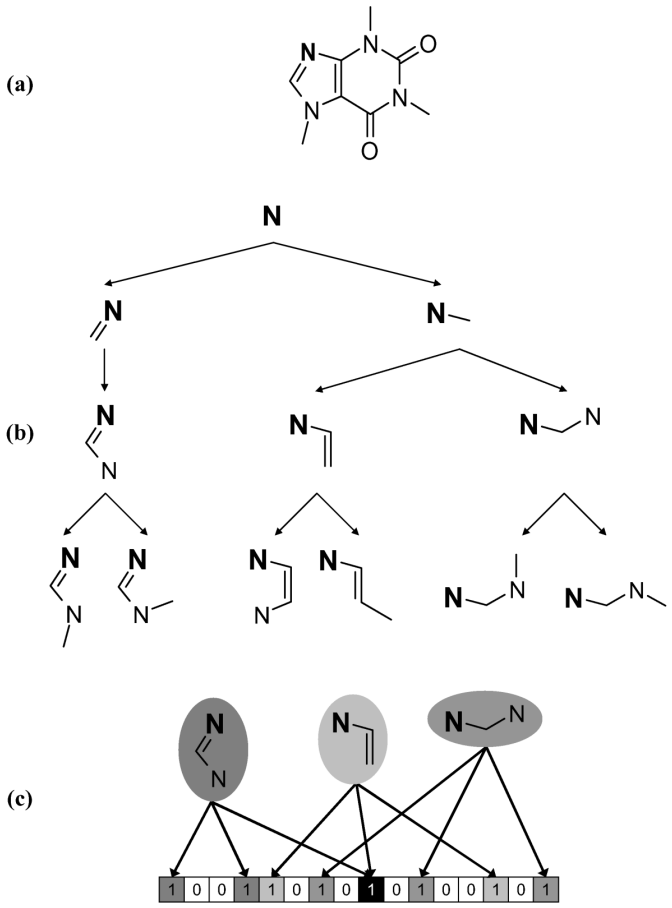


Figura 2.15 – Un ejemplo parcial del codificado de la estructura de la cafeína en un fingerprint binario. La estructura original de la cafeína se muestra en (a) con el átomo de nitrógeno, que constituye la raíz, resaltado. En (b) se muestra la enumeración de caminos posibles de longitud <=3. La presencia de cada uno de estos caminos se codifica mediante *n* posiciones del vector (3 en este caso). Esto se muestra en (c) para caminos de 2 enlaces (Tomado de Brown, 2009 [99]).

la molécula en las bases de datos químicas, y permite realizar búsquedas eficientes de moléculas similares. Algunos ejemplos de algoritmos de plegado utilizados en la generación de *fingerprints* son el algoritmo de plegado de Weisfeiler-Lehman, el algoritmo de plegado de Murcko y el algoritmo de plegado de Daylight [102].

La compresión de *fingerprints* mediante la técnica de plegado o folding es fácil de implementar, pero tiene la desventaja de ser un método de compresión con pérdida, lo que significa que no se puede recuperar el vector binario original a partir del vector comprimido. Además, las colisiones de bits son más frecuentes a medida que la densidad de bits del vector comprimido aumenta, lo que puede conducir a resultados inesperados y poco intuitivos en búsquedas de similitud. A pesar de estas limitaciones, esta técnica de compresión se utiliza en sistemas comerciales como Daylight, Avalon y Unity [103, 104], y es parte también de implementaciones *open source* como las vistas en RDKit [105] o ChemFP [106]

En los últimos tiempos, se han desarrollado algoritmos de compresión sin pérdida para comprimir los vectores de *fingerprints*, según se indica en el estudio de Baldi y colaboradores [104]. Utilizando modelos estadísticos de *fingerprints* binarias, estos métodos permiten reducir el tamaño de las *fingerprints* a aproximadamente 300 bits por molécula con un costo computacional bajo. Al comprimir de esta manera, se pueden realizar cálculos de similitud directamente con los vectores originales (es decir, descomprimidos), lo que mejora el rendimiento y evita la aparición de valores de similitud poco intuitivos que Flower detectó en un estudio anterior [103].

Medidas de similitud

En resumen, la similitud entre moléculas puede ser calculada mediante el conteo de características compartidas, también conocidas como patrones o bits. Esta medida de "distancia química" se puede convertir en un coeficiente de similitud utilizando diferentes métodos, como los coeficientes de asociación o de correlación, por ejemplo el coeficiente de Jaccard/Tanimoto. Este coeficiente se calcula a partir de la presencia o ausencia de características compartidas entre dos moléculas, tal como se ilustra en la tabla 2.1. De esta forma, se pueden comparar diferentes moléculas entre sí y calcular su grado de similitud. Dada esta matriz, el coeficiente de Jaccard/Tanimoto, se define como:

$$T(A, B) = \frac{c}{(a + b - c)} = \frac{|A \cap B|}{|A \cup B|} \tag{2.1}$$

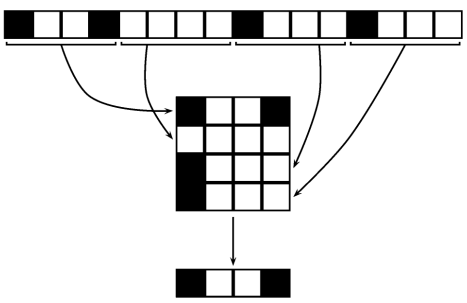


Figura 2.16 – Ilustración del proceso de compresión o "plegado" de un vector binario. En el ejemplo, un vector de longitud N=16 (`1001000010001000`) es plegado en un vector binario de longitud N=4 (`1001`), modulo 4. Notar como la información en la primera posición del vector resultante se pierde debido a colisiones. Tomado de Baldi, *et al.* (2007) [104].

		MOL B		
		0	1	Total
MOL A	0	d	b	b+d
	1	a	c	a+c
	Total	d+a	b+c	n

Tabla 2.1 – Matriz de comparación de características entre dos moléculas A y B. En la tabla se indica la presencia (1) o ausencia (0) de características compartidas entre A y B. a es el número de *bits* presentes en A y ausentes en B. b es el número de *bits* presentes en B y ausentes A. c es el número de *bits* presentes en A y B d es el número de *bits* ausentes en A y B. n es el número total de *bits* (a + b + c + d)

Dado que los bits varían según el vector con el que se representan las moléculas, se pueden obtener medidas de similitud de Tanimoto que atiendan a problemas específicos, independientemente de si las moléculas se parecen o no. También se pueden crear métricas ponderadas en las que se usen varias representaciones vectoriales [107]. En todos los casos, no obstante, este índice tiene un rango entre `0` y `1` y, aunque otros coeficientes para medir similitud entre moléculas, es el más adoptado por la comunidad [84].

Agrupamiento de compuestos químicos

Los compuestos químicos pueden ser clasificados según sus descriptores, lo que resulta útil para diversos propósitos [108]. Estos incluyen agrupar compuestos en familias relacionadas estructuralmente, determinar la estadística del agrupamiento de estructuras en un conjunto grande de datos, seleccionar compuestos representativos de las clases estructurales del conjunto total en relación a una determinada bioactividad, identificar compuestos inusuales en el conjunto de datos y descubrir relaciones entre objetos en un espacio particular de descriptores.

Para llevar a cabo el agrupamiento o *clustering* de compuestos químicos, es necesario utilizar una medida de similitud. Una de las medidas comúnmente utilizadas es la similitud estructural, que puede ser evaluada mediante el índice de Tanimoto [108].

Métodos de agrupamiento de compuestos químicos

Se han realizado estudios sobre diversos algoritmos de agrupamiento para compuestos químicos, y se ha encontrado que el más apropiado para este problema es el algoritmo de Jarvis-Patrick [109, 110]. Este método utiliza una estrategia basada en vecinos más cercanos y produce agrupamientos no jerárquicos con grupos no superpuestos. A diferencia de otros algoritmos, como K-means, el número de grupos no está predefinido, pero depende de dos parámetros: J, el máximo número de vecinos más cercanos a considerar para cada elemento; y K, el número de vecinos comunes que deben tener dos compuestos para ser puestos en el mismo grupo. Sin embargo, este método es computacionalmente prohibitivo para grandes conjuntos de datos [108].

También existen casos de uso para *clustering* jerarquico en la generación de agrupamientos químicos. En este enfoque, las moléculas se agrupan en *clústers* a medida que se van comparando entre sí mediante medidas de similitud, tales como la distancia euclidiana o el coeficiente de Tanimoto. El proceso de agrupamiento se realiza en un árbol jerárquico, donde cada *cluster* es un subconjunto de moléculas que comparten cierto grado de similitud. Este enfoque permite visualizar la similitud estructural entre las moléculas en diferentes niveles de granularidad, lo

que puede ayudar en la identificación de patrones y relaciones entre las moléculas. Además, el *clustering* jerárquico puede utilizarse como una herramienta previa para reducir la complejidad de un conjunto de datos de moléculas antes de aplicar técnicas más avanzadas de minería de datos o aprendizaje automático [111].

El método de K-means es más adecuado para conjuntos de datos grandes. Este algoritmo requiere que se defina de antemano el número de clusters deseados y cuáles serán los representantes iniciales de cada uno de ellos (**K**). Cada elemento a agrupar es asignado al cluster más cercano, y el centroide del cluster se recalcula, incluyendo el nuevo elemento. Este proceso se repite, y en cada iteración es posible que haya reubicaciones de elementos en otros clusters. El proceso termina cuando ya no hay más reubicaciones de elementos [108].

Aunque no son métodos de *clustering*, amerita destacar en este inciso que estos métodos suelen estar acompañados por alguna forma de reducción de la dimensionalidad que permita visualizar los grupos generados en el espacio químico. Algoritmos como PCA, tSNE o UMAP se usan en quimioinformática para reducir la complejidad de los datasets a algo fácilmente observable, y acompañanan los análisis exploratorios. No debe perderse de vista que cualquiera de estos procesos redunda en la pérdida de información y, por tanto, debe usarse con fines meramente exploratorios [112].

2.7. Uso de quimiogenómica para *target* y *drug discovery*

La quimiogenómica (o "genética química") es una disciplina emergente que aprovecha la combinación de la manipulación genética de organismos a escala de genomas completos con la química [113]. Mediante la medición de los efectos de compuestos químicos en células completas (con genotipos conocidos), la quimiogenómica puede ayudar a identificar y validar blancos terapéuticos. Un número de estudios en levaduras han establecido el escenario de medir fenotipos asociados con disruptión genética [114, 115] así como perturbaciones químicas [116]. Los ensayos más informativos son aquellos donde la perturbación química se evalúa contra una colección de mutantes codificados (por ejemplo, una biblioteca de deleción o sobreexpresión). La resistencia o sensibilidad al compuesto en estos ensayos se asocia con un tipo genético específico y puede proporcionar pistas esenciales sobre el modo de acción de las sustancias químicas [117, 118]. De manera simplificada, los mutantes de pérdida de función pueden ayudar a identificar transportadores o enzimas implicados en la absorción o activación de fármacos, los mutantes con función reducida (mutantes de haploinsuficiencia o aquellos con *knockdowns* parciales) pueden hacer que las células sean más sensibles a un fármaco y esto puede ayudar a identificar su blanco; mientras que los mutantes con ganancia de función pueden ayudar a identificar fármacos con otros modos de acción 2.17.

En *T. brucei*, hubo una serie de estudios quimiogenómicos pioneros, basados en la aplicación de bibliotecas de ARNis de todo el genoma para realizar *knockdown* de genes, junto con perturbaciones químicas, seguidas de secuenciación. Estos *screenings* de secuenciación de blancos de interferencia de ARN (RIT-seq) definen mapas del genoma al vincular firmas de densidad de lectura de fragmentos de ADN que producen dsRNAs largos, con la susceptibilidad a los fármacos [119]. Los cinco medicamentos que se usan actualmente para tratar la tripanosomiasis africana humana se probaron en estos *screenings* quimiogenómicos, vinculando tres de estos medicamentos a proteínas individuales: el transportador 1 de adenosina P2 (AT1) para la captación de melarsoprol; el transportador de aminoácidos AAT6 para la captación de eflornitina [119, 120]; y la nitroreductasa I a la activación de nifurtimox [119, 120]. Estos casos fueron ejemplos

de una mayor resistencia a los fármacos cuando los genes que codifican para estas proteínas fueron eliminados y resultaron compatibles con funciones en el transporte o la activación de los fármacos [121]. Para las otras dos drogas (suramina y pentamidina), las pruebas de RIT-seq revelaron vínculos con varias proteínas, además de los transportadores, lo que sugiere una farmacología más compleja.

Screenings quimiogenómicos similares también han esclarecido el mecanismo de acción de los fármacos antileishmaniales utilizando *T. brucei* como modelo [122]. Esto fue posible porque *T. brucei* también es susceptible de ser eliminado in vitro por cuatro fármacos antileishmaniales actuales: antimonio (SbIII), paromomicina (PMM), miltefosina (MTF) y anfotericina B (AMB). El *screening* quimiogenómico RIT-seq condujo a la identificación de ortólogos de *T. brucei* de transportadores de membrana plasmática de fármacos antimoniales y miltefosina; así como un transportador lisosomal para la acción de la paromomicina, y una proteína de membrana asociada

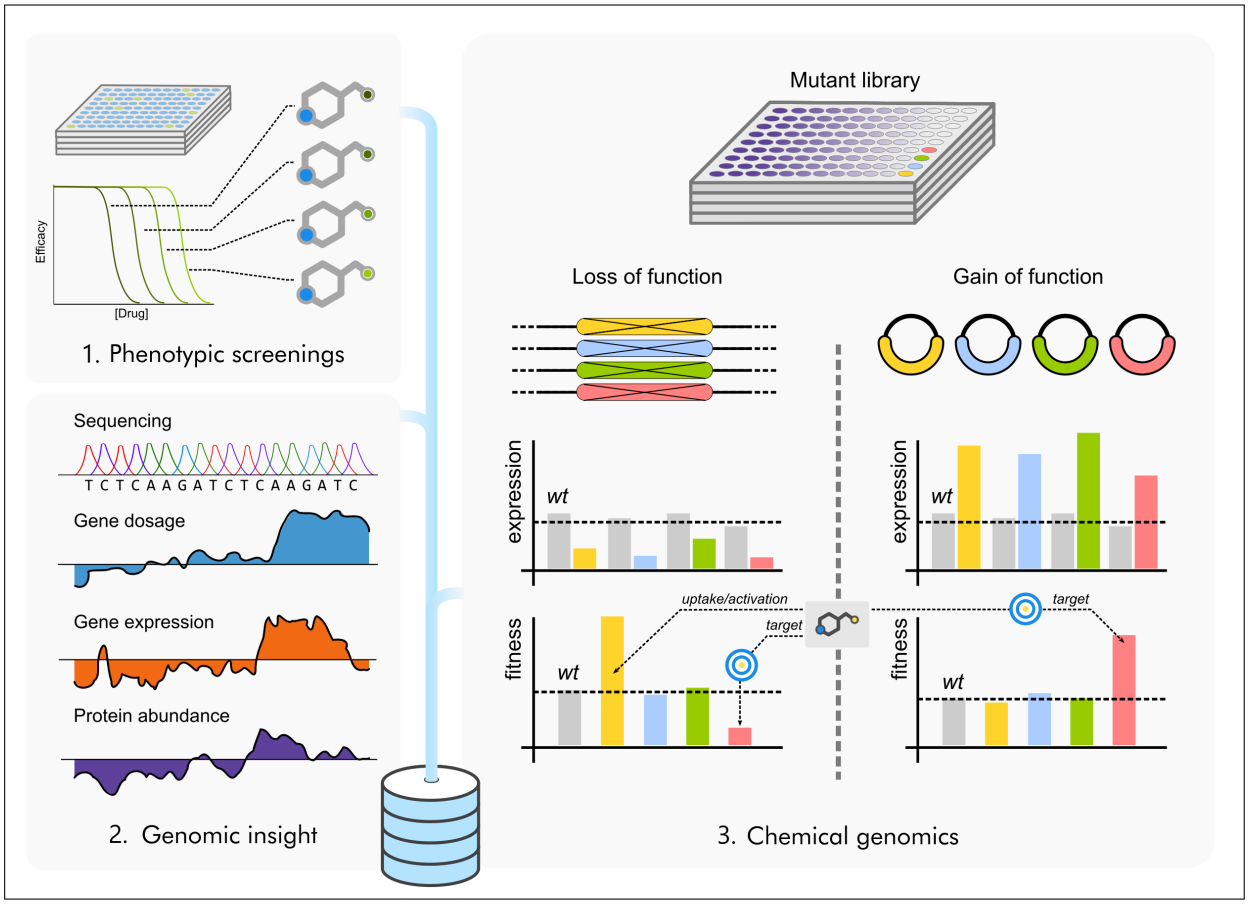


Figura 2.17 – Ilustración de los puntos de partida para el descubrimiento de fármacos por quimiogenómica. En las evaluaciones fenotípicas, los compuestos con efecto deletéreo pueden establecer un compuesto principal que puede optimizarse más tarde (panel 1). La validación de blancos es un proceso desafiante en los tripanosomátidos, debido a la variación del número de copias de genes y las aneuploidías. Para validar con éxito un blanco, el dosaje de genes, la expresión génica y la abundancia de proteínas deben evaluarse a lo largo de todo el proceso (panel 2). Se puede validar un blanco analizando el efecto del compuesto frente a mutantes con pérdida o ganancia de función para genes específicos y comparando el *fitness* frente a parásitos de *wild type* (panel 3). El aumento del *fitness* en un ensayo de pérdida de función puede deberse a la alteración de un transportador o de una enzima activadora (gen amarillo), lo que genera fenotipos similares a los resistentes, mientras que la disminución de la aptitud puede deberse a la alteración de la proteína misma (gen rojo). En un ensayo de ganancia de función, la sobreexpresión de la proteína (gen rojo) puede causar una mayor aptitud, lo que hace que la titulación del blanco putativo sea más difícil. En ambos casos, el gen podría estar involucrado en contrarrestar el efecto del compuesto químico.

a la vesícula y una flipasa que contribuye a la acción de la anfotericina B y la miltefosina [122]. El silenciamiento génico inducido por ARNi se basa en la eficacia del ARNsi para dirigirse a un gen específico [123, 124]. Por lo tanto, diferentes niveles de silenciamiento pueden producir diferentes dosis génicas efectivas tras la inducción de ARNi. Así, la eliminación de un gen esencial puede, en teoría, dar lugar a dos escenarios: si los parásitos mueren porque la dosis del gen es demasiado baja para garantizar la supervivencia, no se producirán lecturas de secuenciación del fragmento objetivo de ARNi. Alternativamente, si la célula sobrevive con una dosis génica reducida, la reducción adicional de la función de la proteína por la acción de un fármaco resultará en un *hit*, vinculando la sustancia química con el fragmento diana de ARNi. Sin embargo, a la fecha, los ensayos de RIT-seq han llevado a la identificación de proteinas indirectos (transportadores, activadores) o letales sintéticos (por ejemplo, proteínas que reparan el daño inducido por fármacos) [125], lo que sugiere que el primer escenario es el resultado más probable en ensayos RIT-seq.

La validación genética de blancos con otros mecanismos de acción puede lograrse mediante la sobreexpresión del blanco potencial en parásitos transgénicos y la demostración de resistencia adquirida. Una de estas pruebas de detección de ganancia de función en *Leishmania infantum* se realizó utilizando cósmidos y analizando la dinámica del enriquecimiento de éstos mediante secuenciación después de la presión selectiva del fármaco [126, 127]. La validación de esta técnica con metotrexato (MTX), reveló un enriquecimiento en los cósmidos que codifican la dihidrofolato reductasa-timidilato sintasa (DHFR-TS) y la pteridina reductasa (PTR1), que son las proteínas de unión a ligandos primaria y secundaria de MTX, al tiempo que reveló nuevos candidatos responsables de resistencia [126]. Tras la validación de la técnica Cos-Seq, se analizaron los cinco principales fármacos antileishmaniales (los enumerados anteriormente más la pentamidina (PTD)), lo que condujo al aislamiento de un número sin precedentes de proteínas conocidas y no identificadas previamente, todas capaces de unirse a estos fármacos. Una lista corta e incompleta de éstas incluye una proteína fosfatasa 2A como el blanco plausible de SbIII, dos genes que codifican una enzima de biosíntesis de ergosterol y una ATPasa translocadora de lípidos como blancos putativos adicionales de MTF, una nueva proteína de unión a ligandos de resistencia a AMB (proteína de membrana hipotética de función desconocida), y otra proteína hipotética de función desconocida que contiene dominios de repeticiones de leucina como una nueva proteína de unión a ligando vinculada a la resistencia tanto a PTD como a PMM [126, 128]. En *T. brucei*, se utilizó una biblioteca de sobreexpresión de alta cobertura para revelar los blancos potenciales de los benzoxaboroles SCYX-7158 y AN11736, nuevos fármacos prometedores para el tratamiento de la tripanosomiasis humana y animal en África [129] y en desarrollo clínico activo [130].

También se ha propuesto la utilización de perfilado térmico de proteínas (TPPs), que se basa en el cambio en la estabilidad térmica de las proteínas tras la estabilización de un interacción con moléculas pequeñas [131, 132]. Sin embargo, en la práctica, el TPP genera muchos posibles candidatos, ya sean blancos de fármacos de genuinos o falsos positivos [133, 134], por lo que los resultados deben confirmarse o validarse utilizando criterios ortogonales. Por el contrario, cuando se utiliza con blancos puntuales *in vitro*, el perfil térmico de dicho blanco permite para validar exitosamente la unión de un fármaco o incluso ser usado para *screenear* una biblioteca de drogas contra un único blanco potencial. Un ejemplo de ello es el estudio reciente sobre la unión de la N-miristoiltransferasa (NMT) de *L. donovani* por el derivado de pirazolil sulfonamida DDD100097, un compuesto con actividad contra los promastigotes [135].

La genómica clásica también puede ayudar a validar blancos, sin el uso de bibliotecas de mutantes, mediante la secuenciación de parásitos naturales o seleccionados *in vitro* aislados después

de haber inducido resistencia a un determinado fármaco. Este tipo de análisis de secuencias genómicas se centró en la variación del número de copias (CNV), somía [136–138], segmentos del genoma amplificados [139], pérdida de genes [140–142], tanto así como mutaciones puntuales [41, 137, 141, 143, 144]. Estos estudios han arrojado luz sobre mecanismos de acción nuevos (o adicionales) de varios fármacos. En *Leishmania donovani*, por ejemplo, los mutantes inducidos químicamente se desafiaron con miltefosina y paromomicina para seleccionar clones resistentes, que se secuenciaron para identificar SNPs condujo a la identificación de la kinasa CDPK1 como el gen mutado más abundante en 14 clones independientes con el fenotipo resistente para paromomicina [145].

La resistencia a los medicamentos, no obstante, puede ser compleja. Además de la mutación del blanco natural de un fármaco, la resistencia puede implicar proteínas adicionales en las vías de desintoxicación, captación alterada, aumento del flujo de salida y secuestro intracelular. No todos estos pueden ser descubiertos por ensayos de ganancia de función. Los tripanosomátidos hacen que este enfoque sea particularmente desafiante debido a la plasticidad de su genoma [41, 146], con la posibilidad de que la sustancia química pueda causar la selección de clones específicos de parásitos (p. ej., resistentes a los medicamentos) enriqueciendo el cultivo con haplotipos particulares [147].

3. Democratización de datos en *drug discovery*

3.1. Introducción

Las herramientas computacionales se están volviendo cada vez más esenciales en el descubrimiento de fármacos traslacionales, tanto en la academia como en la industria farmacéutica. La integración inteligente e intensiva de los crecientes volúmenes de datos generados durante todas las fases del descubrimiento de fármacos ya permite abordar algunos de los desafíos clave del proceso [148]. Desde su introducción, la base de datos TDR Targets ha sido un recurso confiable para que los investigadores que trabajan en enfermedades desatendidas accedan a datos de quimiogenómica para la priorización de blancos terapéuticos y el reposicionamiento de fármacos contra los patógenos que las causan. Presentado en 2008 [29], este recurso de acceso abierto permitió a los investigadores buscar blancos terapéuticos (y más tarde también inhibidores químicos [149]), y priorizarlos para ayudar al desarrollo de fármacos para los patógenos causantes de NTDs. TDR Targets hace uso de conjuntos de datos funcionales de todo el genoma, disponibles públicamente, para permitir a los usuarios encontrar y priorizar proteínas en función del conocimiento de la biología de su patógeno de interés y la naturaleza de la enfermedad [18, 150]. Esto se implementa mediante una selección flexible de proteínas y centrada en el usuario (usando criterios de filtrado) y una clasificación (usando una ponderación específica de criterios) [149, 151].

En este capítulo, se describirán las actualizaciones de los conjuntos de datos subyacentes y funcionallidades nuevas en el recurso TDR Targets. El nuevo lanzamiento de TDR Targets (v6.1, abreviado TDR6 a partir de ahora en este documento) integra información genómica específica de patógenos con datos funcionales (por ejemplo, expresión génica, relaciones filogenéticas basadas en ortología, esencialidad) de una selección de organismos, junto con datos de compuestos bioactivos (estructura química, propiedades y bioactividad/información sobre el blanco). Todos estos datos se pueden obtener por consultas intuitivas o simplemente por navegación desde la página web. Los usuarios registrados pueden guardar todas las consultas en un archivo personal y publicarlas a través de la aplicación web para maximizar las oportunidades de colaboración. Las listas priorizadas de blancos se pueden exportar para un análisis adicional off-line. Los detalles completos de todas las características novedosas se pueden encontrar en las notas de la versión (https://tdrtargets.org/releases). Este informe presenta un recorrido completo de la aplicación web al día de la fecha, sus características novedosas y ejemplos varios para ilustrar casos de uso.

3.2. Nuevas funcionalidades incorporadas en TDR6

Como en versiones anteriores de TDR Targets, TDR6 también está organizado en dos secciones principales: *Targets* (proteínas) y *Compounds* (moléculas pequeñas). La sección de *targets* de la base de datos contiene datos de todo el genoma para 20 patógenos humanos y permite a los usuarios realizar consultas y priorizar blancos protéicos en función de una serie de características y datos relevantes para el proceso de *drug discovery* (Tabla 3.1). La sección de *compounds* de la base de datos contiene información sobre >2 millones de compuestos bioactivos y permite consultas

basadas en estructura, propiedades químicas y bioactividades anotadas (3.2).

Table 1.
Available target queries in TDR targets

Query group	Pathogens for which data is available	Data types available for querying
Names & Annotations	All	Gene identifiers and functional annotations (EC numbers, GO terms, Pfam domains, metabolic pathway mappings)
Protein Features	All	MW, isoelectric point, presence of predicted signal peptide, trans-membrane segments and glycosylphosphatidylinositol (GPI) anchors.
Structural Information	All	Availability of 3D structures in PDB; availability of structural models in Modbase
Gene expression	*Plasmodium* spp.; *Leishmania* spp.; *Trypanosoma* spp.; *Mycobacterium tuberculosis; Echinococcus multilocularis; Entamoeba histolytica; Toxoplasma gondii*	Gene expression data from pathogen life cycle stages and/or experimental conditions that are relevant to drug discovery.
Phylogenetic information	All	Filter targets using simplified 'present/absent' in other species criteria, based on ortholog group information. Includes model organisms (human) and other related pathogens.
Essentiality	*C. elegans* (model for helminths); *E. coli* (model for bacteria); *S. cerevisiae* (model for eukaryotic pathogens); *Trypanosoma brucei; Mycobacterium tuberculosis; Toxoplasma gondii; Plasmodium berghei*	Ortholog-based inference of essentiality of genes in life cycle stages and/or experimental conditions relevant to drug discovery. Integrated from selected genome-wide gene disruption (e.g. transposon, CRISPR/Cas) and knockdown (e.g. RNAi) datasets in pathogens and model organisms.
Target Validation Data	*Schistosoma mansoni; Leishmania major; Trypanosoma cruzi; Trypanosoma brucei; Mycobacterium leprae; Mycobacterium tuberculosis; Plasmodium falciparum*	Manually curated data on target validation credentials (genetic, chemical and/or pharmacological, observed phenotypes)
Druggability	All	Precedent for successful chemical modulation of target activity or function. Summarized into a druggability score calculated from the network model (see main text)
Assayability	All	Available biochemical assays for protein targets (mapping based on EC numbers)
Bibliographic references	All	Filter targets based on available publications

Tabla 3.1 – Consultas disponibles para proteínas en TDR Targets

Recientemente, nuestro grupo de trabajo informó un modelo basado en redes complejas multicapa en el que todos los datos a escala del genómica disponibles para blancos TDR (blancos protéicos), información química (compuestos bioactivos) y sus relaciones (bioactividad de compuestos en ensayos basados en proteínas y organismos) se vincularon entre sí [152]. En TDR6, este modelo de red se actualizó mediante la integración de nuevos conjuntos de datos (descritos a lo largo de este capítulo). Este modelo incorpora enlaces entre blancos y compuestos bioactivos derivados de la curación manual de ensayos de bioactividad publicados (es conexiones directas entre blancos y compuestos químicos), así como relaciones virtuales (conexiones blanco-blanco y conexiones compuesto-compuesto) basadas en anotaciones de proteínas (dominios Pfam, grupos de ortología) y similitud química. Un aspecto clave de estos enlaces en el modelo de red multicapa es que permiten la rápida exploración y visualización de la vecindad alrededor de blancos y/o compuestos bioactivos de interés. Esto, a su vez, permite a los usuarios explorar compuestos vinculados a blancos, inspeccionar el vecindario de similitud química alrededor de los compuestos bioactivos y visualizar estos datos de una manera integral y fácil de usar (consulte la Figura 3.1) [153].

Query group	Data types available for querying
Text-based searches	
Names & Annotations	Compound names or synonyms; Database identifiers (e.g. ChEMBL, PubChem); InCHI and InCHI key identifiers
Chemical Properties	Molecular weight; LogP octanol/water partition coefficient; number of H donors and acceptors, number of flexible bonds and number of matching Ro5 (Lipinski)
Compound formula	Search by compounds containing a specific number (e.g. 3) of defined atoms (e.g. Cl, F, Br, N)
Bioactivity	Text search on assay descriptions; numerical search for values in assays (e.g. IC50 < 5 µM)
Orphan compounds	Search for compounds that have bioactivity reports in whole-organism or whole-cell assays but lack target and mechanism information (orphans inhibitor/drugs)
Compounds with targets	Find compounds that have target information and mechanism based assays
Structure-based searches	
Compound similarity	Draw/paste compound or fragment 2D structure and search for similar compounds. Search is based on matching of chemical fingerprints
Compound substructure	Draw/paste compound or fragment 2D structure and search for compounds in the database that contain the query fragment.

Tabla 3.2 – Consultas disponibles para moléculas pequeñas en TDR Targets

Con estas actualizaciones, TDR6 ahora brinda a los usuarios las siguientes funcionalidades: i) Priorización de blancos a partir de proteomas completos usando la red, ii) Exploración de oportunidades de reposicionamiento de fármacos; y iii) la exploración de blancos potenciales para compuestos huérfanos (compuestos cuya actividad antibiótica ha sido comprobada, pero para los que su mecanismo de acción es desconocido). Estos casos de uso son posibles gracias al puntaje de drogabilidad derivado de la red, o *Network Druggability Score* (NDS), que maximiza el puntaje para los blancos (drogables o no) que están internamente conectados con blancos drogables, y lo minimiza para blancos aislados o conectados con otros blancos sin evidencia de modulación química. Al asociar una métrica cuantitativa basada en el enriquecimiento de compuestos bioactivos conectados a proteínas dentro red, esta puntuación facilita la clasificación de las proteínas en grupos de drogabilidad (DG), una característica obtenida para todos las proteínas en la base datos, que puede ser utilizada para realizar búsquedas y filtros.

El modelo de red también es la base de las priorizaciones impulsadas por la red, *Network Driven Prioritizations* (NDP); que pueden ser consultadas por los usuarios y también son utilizadas internamente por TDR6 para seleccionar los blancos y compuestos conectados a la entidad seleccionada al construir el subgrafo con la vecindad de la misma (ver más abajo). Al partir desde un compuesto de interés, TDR6 utiliza las priorizaciones precalculadas de blancos candidatos para ayudar a los usuarios en la navegación de los blancos vecinos alrededor del compuesto (y viceversa cuando comienza desde un blanco de interés). Al proporcionar estas clasificaciones y métricas de enriquecimiento precalculadas, la base de datos ahora facilita el descubrimiento de nuevas asociaciones de drogas y blancos. Además de estos nuevos NDP precalculados, los usuarios pueden priorizar los objetivos utilizando la misma funcionalidad que en las versiones anteriores de TDR Targets.

Esta versión también incluye varias actualizaciones de datos, a saber, la inclusión de 22 nuevos genomas (20 nuevos patógenos y 2 nuevos organismos modelo) y amplias actualizaciones de

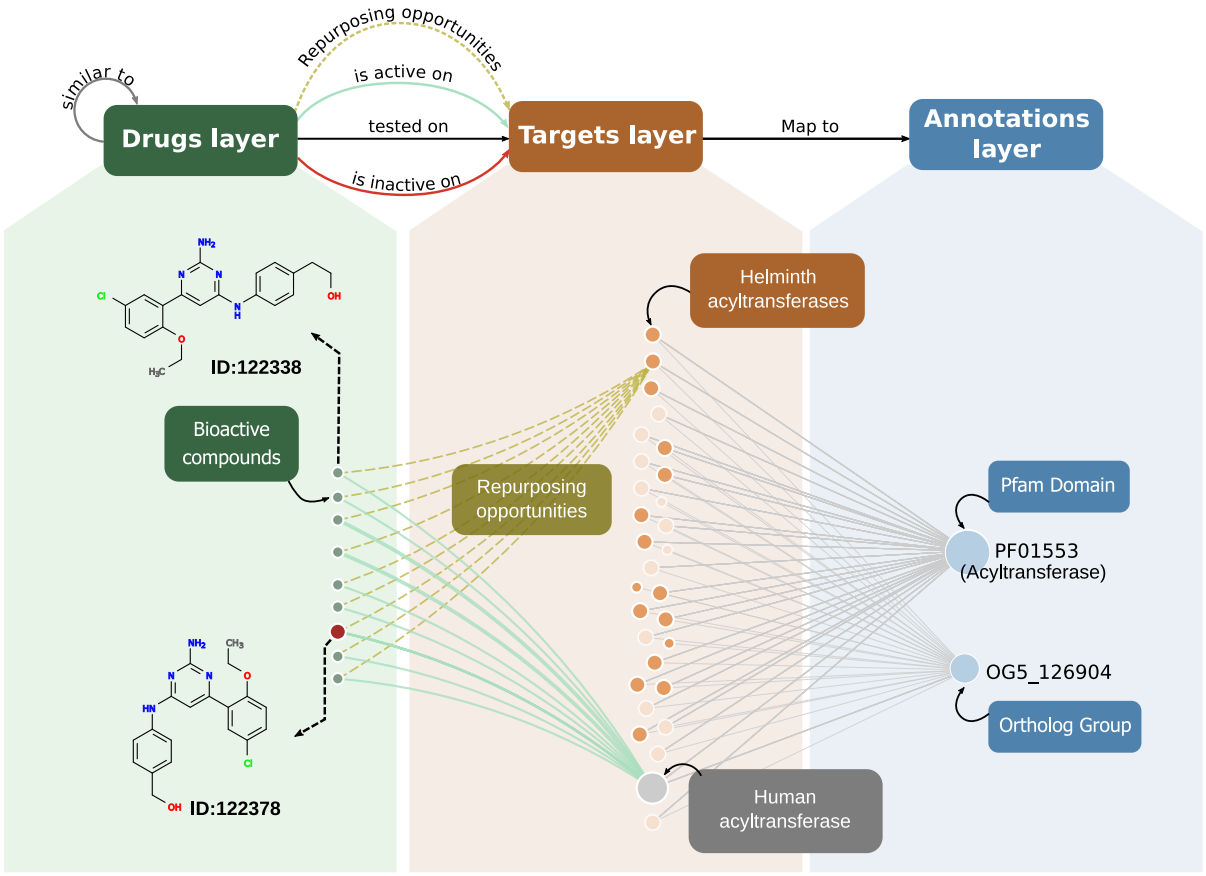


Figura 3.1 – Modelo de red esquemático en TDR6 y visualizaciones de subgrafos. Las páginas de compuestos o de blancos ahora muestran una visualización del subgrafo que contiene todas las entidades en la vecindad de la red del nodo seleccionado. Estos grafos se construyen a partir de un grafo de red complejo de tres capas. Una visión esquemática sobre cómo se conectan las entidades en la red, puede verse en la parte superior de esta figura. Tanto los sub-grafos de blancos como los de compuestos se organizan de la siguiente manera: los nodos de compuestos (verde) están conectados a los blancos (naranja = patógenos; gris = no patógenos) a través de ensayos de bioactividad. Estos enlaces muestran resultados de ensayo positivos (verde), neutros (gris) o negativos (rojo). Finalmente, los blancos se asignan a un conjunto de afiliaciones funcionales (anotaciones, nodos azules). En el ejemplo, el gráfico muestra un conjunto de inhibidores conocidos para una aciltransferasa humana . Estas bioactividades (todas positivas) se dibujan como enlaces verdes entre los compuestos (nodos verdes) y el blanco (nodos grises para no patógenos, nodos naranjas para patógenos). El grafo destaca las oportunidades de reposicionamiento de las compuestos a las aciltransferasas de helmintos (líneas discontinuas, añadidas manualmente para esta figura), según las anotaciones compartidas entre éstas y el blanco original de los compuestos, una aciltransferasa de humano. El nodo rojo en la capa del fármaco indica el compuesto seleccionado. Los tamaños de los nodos están determinados por la cantidad de conexiones en la red (grado), mientras que el ancho de los enlaces de bioactividad están relacionados con la cantidad acumulada de evidencia experimental para un par determinado de fármaco-objetivo (cantidad de ensayos distintos sugiriendo la misma interacción).

datos químicos y de bioactividad, entre otros. Una interfaz de usuario mejorada y versátil, junto con las actualizaciones de datos, renuevan el compromiso de TDR Targets de proporcionar una herramienta integrada y poderosa para explorar datos genómicos y químicos en el contexto de las enfermedades tropicales desatendidas.

3.3. Uso de TDR Targets

3.3.1. Priorización de blancos moleculares drogables

Priorizaciones de genoma completo

El modelo de red [152] es la base para la nueva puntuación de drogabilidad, una métrica derivada de la red que, como se anticipó más arriba, está relacionada con el enriquecimiento en compuestos bioactivos para un blanco determinado conectado directa o indirectamente al blanco de interés. Las puntuaciones NDS están disponibles para todos los organismos de *tier* 1, que son aquellos en los que nuestro grupo pone mayor interés y esfuerzo para la obtención y curación de datos por ser todos patógenos causantes de enfermedades desatendidas. El NDS para estos organismos se se pueden consultar y utilizar para ponderar las consultas para filtrar (dentro o fuera) blancos en experimentos de priorización personalizados definidos por el usuario. Como se explica más detalladamente en la sección de integración de datos de la red, para cada organismo los blancos se clasificaron en cinco Grupos de Drogabilidad (DG), desde el puntaje más bajo (DG1) hasta el puntaje más alto (DG5), según su desempeño en las priorizaciones.

Como en versiones anteriores de TDR Targets, los usuarios pueden combinar diferentes conjuntos de datos simplemente ejecutando consultas individuales en diferentes tipos de datos y combinándolos en la página de historial [18, 29, 149, 151]. Esto es útil cuando, por ejemplo, a los usuarios les gustaría incluir tipos de datos adicionales a la drogabilidad de las proteínas, como aquellos que se basan en la expresión génica en etapas relevantes del ciclo de vida del organismo de interés, o aquellos que brindan información sobre la aptitud/letalidad de los blancos (esencialidad).

Como ejemplo, presentamos aquí un ejemplo de priorización usando *Toxoplasma gondii* como el patógeno de interés. *T. gondii* es un parásito apicomplexa que se utiliza a menudo como modelo para investigar la biología subyacente a varias enfermedades humanas y animales [153]. La estrategia de búsqueda se resume en la Figura 3.2. La consulta se comenzó buscando todas las proteínas de *T. gondii* y filtrando de esa lista todas aquellas con homólogos en humanos (para seleccionar solo blancos específicos de parásitos). A continuación, se seleccionaron genes esenciales en función de los perfiles de *fitness* observados para la infección de fibroblastos humanos mediante secuación masiva de una biblioteca de mutantes *knock-out* generada para todo el genoma del parásito [154]; y también genes altamente expresados en taquizoítos (etapa replicativa de *T. gondii*) recuperando solo los genes en el percentil 80-100 de la abundancia de transcritos de RNAseq [155]. Estas selecciones se combinaron con las clasificaciones de drogabilidad de la red. Para esto, se consideró los genes en los grupos de drobailidad 3, 4 o 5 (DG ≥ 3) (ver Figura 3.2). La figura muestra todas las consultas y sus resultados tal como se ven en la sección Historial de TDR Targets, y las operaciones realizadas al combinar consultas (unión, intersección). La lista final de blancos clasificados en función de estos criterios se ha hecho pública y está disponible en la sección de las listas publicadas.

Reposicionamiento mediante transformaciones de listas de entidades

La consulta de drogabilidad en TDR6 permite a los usuarios seleccionar blancos con inhibidores/fármacos conocidos o previstos. La información sobre blancos con fármacos conocidos proviene de la curación de la literatura, mientras que las asociaciones predichas (indirectas) de blancos con inhibidores/fármacos se obtienen a través de cálculos de similitud de

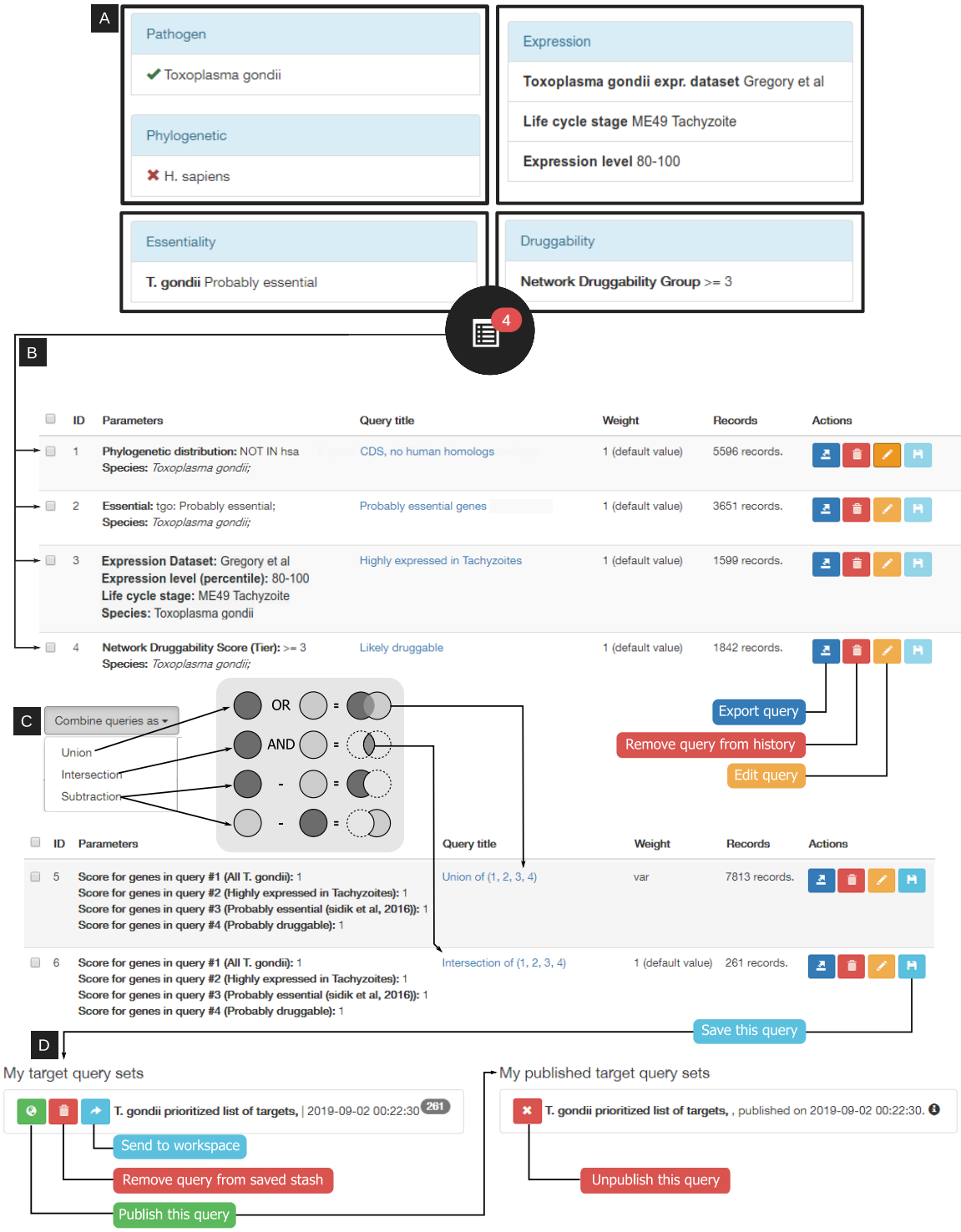


Figura 3.2 – Estrategia de ejemplo de priorización de blancos para *T. gondii*. La imagen compuesta muestra (A) los términos de consulta utilizados para encontrar blancos de *T. gondii* que no tienen homólogos en humanos, que se expresan en gran medida en la etapa de taquizoíto virulento del parásito durante la infección de células humanas, que probablemente son esenciales y que son probablemente drogables. (B) Resumen de las consultas realizadas en la página 'Targets', mostrando cómo aparecen estas consultas en la sección 'Historial', donde se pueden revisar y transformar. Los botones de gestión de consultas en línea permiten acciones adicionales (eliminar, renombrar, exportar). (C) Las combinaciones de consultas permiten a los usuarios ejecutar acciones de unión, intersección o sustracción en consultas entre sí. Se muestran ejemplos de acciones de unión e intersección. (D) Las consultas se pueden guardar en un almacenamiento privado ('*My target query sets*') desde donde se pueden enviar de vuelta al espacio de trabajo (para realizar operaciones de consulta adicionales) o compartirlas públicamente con otros usuarios de TDR Targets (*My published target query sets*).

secuencia u ortología (de un blanco hacia otro para el que sí se conocen interactores químicos), o mediante inferencias respaldadas por redes [152] que aprovechan esta y otras relaciones indirectas. Todos estos métodos están implementados en TDR6. Por lo tanto, cuando los usuarios filtran un conjunto de genes en función de la drogabilidad, limitan la selección a blancos altamente *rankeados*, lo que debería proporcionar una rica fuente de oportunidades de reposicionamiento de fármacos.

Para mostrar la utilidad de TDR6 en esta área, mostramos cómo buscar fármacos candidatos para la reutilización de *Echinococcus multilocularis* (el agente causante de la equinococosis alveolar). Esto se muestra en la Figura 3.3. El proceso es similar al descrito anteriormente para *T. gondii*, pero en esta estrategia de consulta no descartamos los homólogos humanos y hemos utilizado datos de letalidad de ARNi de *C. elegans* como proxy de la esencialidad de genes en estos nematodos. Como resultado, obtuvimos una priorización del genoma completo para *E. multilocularis*. A continuación, al aplicar un filtro basado en drogabilidad para los blancos en a esta consulta, hemos reducido la selección de blancos a un solo puñado de proteínas. El usuario puede inspeccionar manualmente los blancos seleccionados para averiguar qué medicamentos se enumeraron a través de asociaciones indirectas. Las páginas de cada *target* mostrarán todos los compuestos asociados en la sección de drogabilidad, clasificados según la fuente de la inferencia. Para las inferencias impulsadas por redes, la puntuación de cada compuesto propuesto aparecerá como una lista y como un diagrama de clasificación, para identificar rápidamente a los candidatos prometedores. Alternativamente, para minimizar la inspección manual, la lista de genes (es decir, la consulta en sí) se puede convertir fácilmente en una lista de medicamentos asociados haciendo clic en los botones '*Transform this query*' en la parte superior de las páginas de resultados de la consulta. Esta funcionalidad proporciona una forma rápida de comenzar a crear una biblioteca química para *screening* de compuestos. Las transformaciones de consultas pueden basarse en selecciones curadas (medicamentos conocidos para un conjunto de blancos), predichas (asociaciones computarizadas con medicamentos) o ambas. En los tres enfoques, los inhibidores/fármacos asociados con blancos conocidos se asocian, transitivamente, con los genes de la lista. La Figura 3.3 resume la estrategia de priorización, la conversión de consultas de la lista de genes a compuestos y un ejemplo de la visualización de subgrafos disponible en la página de compuesto de una propuesta de reposicionamiento .

3.3.2. Búsqueda de blancos potenciales para compuestos huérfanos

Las actividades de los compuestos extraídos de la literatura mediante curación aparecen en forma de ensayos basados en proteínas (enlace directo al blanco) o en forma de ensayos basados en células o de organismo completo. En ausencia de otra información, esta última clase de ensayos no proporcionan pistas sobre el blanco o el mecanismo de acción de los compuestos. Durante el proceso de actualización de datos químicos en TDR6, se incorporaron a la base de datos compuestos con efectos fenotípicos informados en ensayos de organismos completos o basados en células, según sus clasificaciones ChEMBL. Esta información se usó para identificar compuestos "huérfanos", que son activos contra un patógeno particular en exámenes de detección primarios o secundarios basados en células, pero para los cuales no existe un ensayo contra un blanco puntual.

Los compuestos huérfanos en TDR6 se pueden buscar para cualquier organismos con datos de detección fenotípica disponibles, dentro de la página de búsqueda de compuestos. Esto constituye una forma rápida de aprovechar los datos de los ensayos de alto rendimiento, lo que a su vez habilita a los usuarios a comenzar sus priorizaciones a partir de compuestos con actividad conocida contra un patógeno de interés.

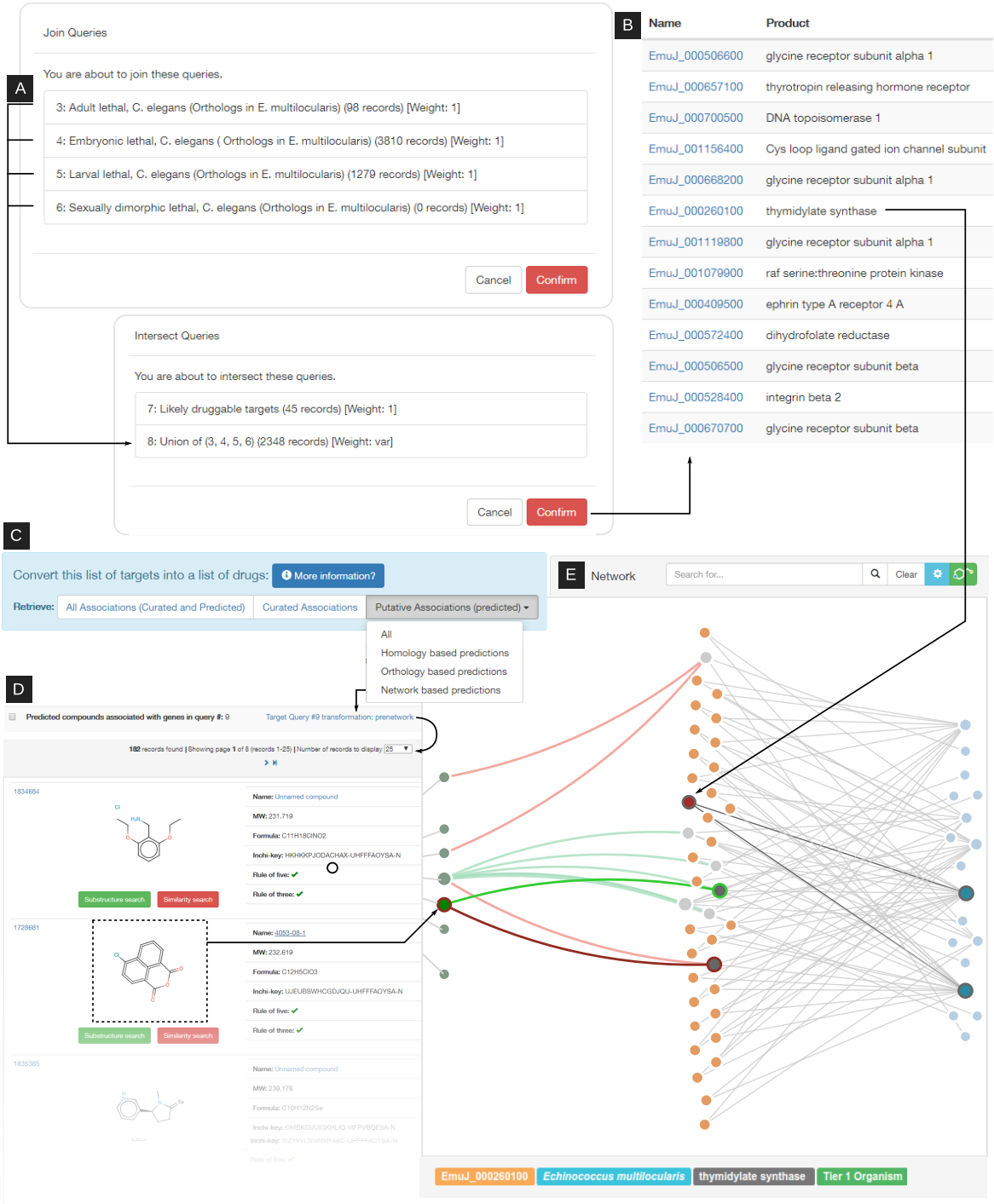


Figura 3.3 – Oportunidades de reutilización de fármacos para *E. multilocularis* mediante transformaciones de consulta: un esquema de priorización de blancos para *E. multilocularis* que se basa en la inferencia por ortología y drogabilidad prevista (DG ≥ 3). (A) consultas combinadas; (B) lista inicial de blancos priorizados. (C) Cualquier lista de blancos se puede "transformar" en una lista de sus moléculas pequeñas asociadas, utilizando cualquiera de los métodos de inferencia de compuestos disponibles (ver texto principal). (D) lista resultante de compuestos bioactivos. (E) Ejemplo de visualización de subgrafos de red de un compuesto seleccionado, que muestra enlaces de bioactividad activos e inactivos. Los compuestos (nodos verdes) están conectados a blancos de patógenos (naranja) de acuerdo con los registros de bioactividad (verde = activo; rojo = inactivo, consulte el texto principal para conocer los umbrales de actividad). Los blancos, a su vez, están conectados (enlaces grises) a afiliaciones funcionales (nodos azules). La representación del subgrafo proporciona sugerencias visuales sobre cómo el blanco inicial de *E. multilocularis* está conectado con el compuesto seleccionado en la red.

El modelo de red integrado en TDR6 también es útil para identificar blancos candidatos para compuestos huérfanos. Como se describe en la publicación original [152], la vecindad de similitud química calculada alrededor de un compuesto huérfano seleccionado puede proporcionar vínculos indirectos con uno o más blancos. Usando esta estrategia, hemos realizado priorizaciones de blancos para todos los compuestos huérfanos en TDR6. Estas priorizaciones precalculadas están disponibles para todos los organismos para los que se dispone de datos de detección fenotípica. Los resúmenes globales que muestran todos los compuestos huérfanos para estos organismos están vinculados desde la sección *data summary*, consultando `https://tdrtargets.org/datasummary`, y haciendo clic en la especie de interés). En la Figura 3.4 se muestra un ejemplo de priorización basada en compuestos huérfanos para *T. cruzi*.

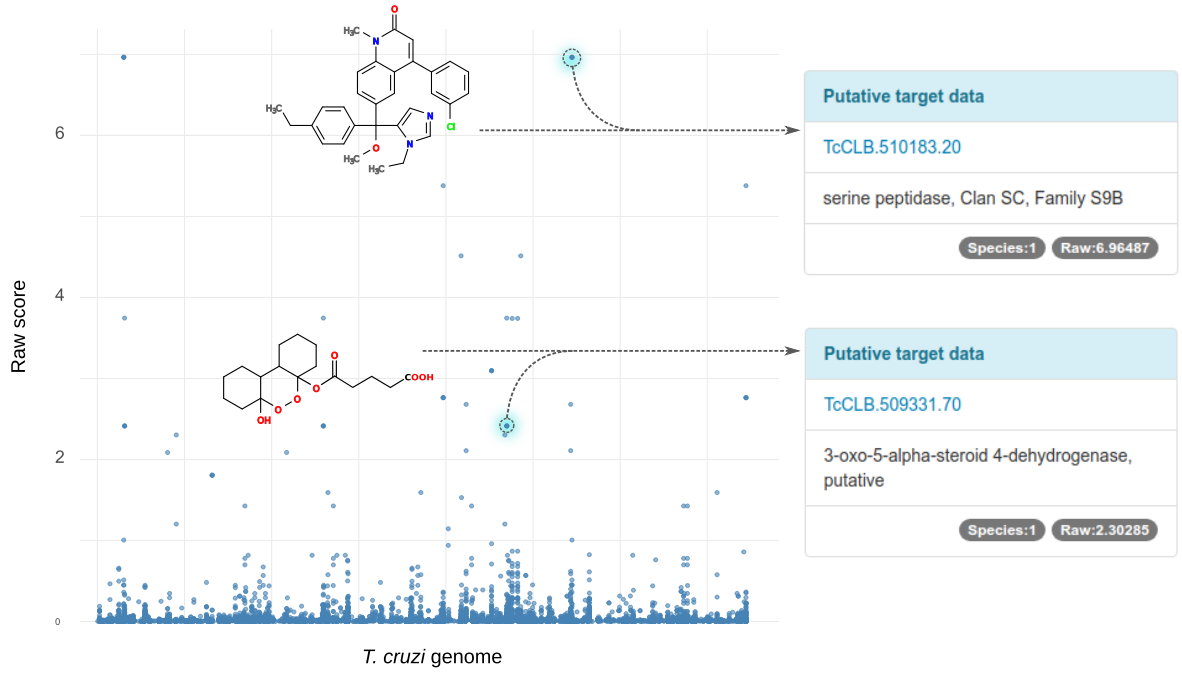


Figura 3.4 – Exploración de blancos candidatos para compuestos huérfanos de *Trypanosoma cruzi*. El gráfico resume la priorización de blancos propuestos por la red para compuestos huérfanos activos contra *T. cruzi*. Todas las proteínas codificadas en el genoma de *T. cruzi* (blancos putativos) están dispuestos en el eje x. Los puntos en la gráfica corresponden a asociaciones blanco-molécula puntuadas por el algoritmo (puntuación trazada en el eje y). Como ejemplo, destacamos dos blancos putativos para dos medicamentos diferentes (tal cual se vería en la página de resumen de datos de *T. cruzi*). Gráficos similares están disponibles en línea para organismos de *Tier* 1 en TDR6, en donde además los datos puede explorarse de forma dinámica.

3.4. Democratización de la información quimiogenómica

3.4.1. Actualización de datos genómicos

Desde la publicación anterior de la base de datos TDR Targets [149], se han agregado varios genomas de patógenos. Se proporciona una lista detallada de éstos en la Tabla 3.3 de este capítulo, así como en línea en la página de resumen de datos de TDR6 `https://tdrtargets.org/datasummary`.

Species	CDS	PFAM	GO	EC	Pathways	Orthologs
Plasmodium falciparum	5349	3322	3551	750	1083	5166
Plasmodium vivax	5344	3264	2631	641	806	5207
Toxoplasma gondii	7946	4025	3795	772	967	6764
Chlamydia trachomatis	887	704	598	269	357	645
Mycobacterium leprae	1630	1236	929	628	611	1473
Mycobacterium tuberculosis	4004	2934	2001	1174	1145	3287
Mycobacterium ulcerans	4232	3602	2578	873	1002	3459
Treponema pallidum	1036	791	634	221	335	733
Wolbachia endosymbiont of B. malayi	805	628	577	308	382	688
Brugia malayi	11316	7042	6368	1278	1787	8424
Echinococcus granulosus	10249	6481	5432	854	1965	7109
Echinococcus multilocularis	10474	6817	5768	878	2079	7539
Loa Loa (eye worm)	16292	8071	6774	1539	2207	10484
Onchocerca volvulus	12224	3248	2178	246	563	4054
Schistosoma mansoni	12692	7818	7384	1218	1649	10386
Leishmania major	8280	4641	4415	1067	1162	8250
Trypanosoma brucei	10270	5665	5482	1019	1264	9259
Trypanosoma cruzi	18639	9908	8572	1495	1735	18140
Entamoeba histolytica	8211	4920	4087	645	1094	7692
Giardia lamblia	9665	2726	2263	326	514	5977
Trichomonas vaginalis	95600	35474	18435	843	1366	87303

Tabla 3.3 – Resumen de disponibilidad de datos para patógenos de *Tier* 1. CDS: número de secuencias codificantes; PFAM: número de proteínas con dominio(s) Pfam mapeado(s); GO: número de proteínas con términos de Gene Ontology mapeados; EC: número de proteínas con números mapeados de la Comisión de Enzimas (EC); *Pathways*: número de proteínas asignadas a mapas de rutas metabólicas KEGG; *Orthologs*: número de secuencias asignadas a grupos de ortología OrthoMCL. Una tabla de resumen de datos más completa está disponible en línea en `https://tdrtargets.org/datasummary`

Dada la diversidad de organismos integrados en TDR Targets y, en consecuencia, la variedad de fuentes de datos necesarias para cubrir todos los genomas; se ha realizado un esfuerzo considerable para estandarizar la recuperación de datos y el análisis de la información del genoma de estos organismos. La mayoría de los genomas completos se obtuvieron de EupathDB [156], GenBank [157], GeneDB [158], Wormbase Parasite [159], GenoList [160] o Mycobrowser [161].

En la Tabla 3.4 se proporciona una descripción completa de las fuentes utilizadas. Para actualizar los datos de los organismos ya presentes la versión anterior de TDR Targets, los genes codificantes de proteínas de la versión actual de los genomas se mapearon a genes existentes en TDR Targets o se ingresaron como nuevos registros. El algoritmo de mapeo usa una combinación de condiciones para rastrear los identificadores de genes en cada versión, y así mantener la identidad de los genes: identidad de los *hashes* criptográficos producidos por las secuencias protéicas (usando valores *hash* de 128 bits generados por el algoritmo MD5), identidad de nombres o identificadores de genes y similitud de secuencia por BLAST [162] cuando se hallan más de una coincidencia (o ninguna) y es preciso resolver una ambigüedad o extender la búsqueda aún cuando no haya coincidencias perfectas. Después de actualizar los registros, el proceso calcula las propiedades fisicoquímicas usando pepstats del paquete EMBOSS [163], busca dominios transmembrana usando TMHMM [164], péptidos señal con SignalP [165] y puntos de anclaje de glicosilfosfatidilinositol usando PredGPI [166]. El algoritmo descarta todas las secuencias no codificantes y pseudogenes, para evitar anotaciones engañosas y minimizar las inferencias falsas durante las priorizaciones ulteriores.

A partir de TDR6, todas las tareas mencionadas anteriormente para la integración y actualización del genoma se han incluido en un flujo de trabajo automatizado para facilitar actualizaciones más rápidas en versiones futuras. En la Figura 3.4 se muestra un esquema del algoritmo que dirige el proceso de actualización. El proceso también automatiza el cálculo de anotaciones utilizando estrategias individuales *ad hoc* para diferentes anotaciones, basándose en servicios web y APIs (como el servicio KAAS ([167] para mapear proteínas en rutas metabólicas y en la clasificación de número EC de enzimas, o la asignación de grupos de ortología por OrthoMCL [79, 168]). La rutina también aprovecha para actualizar anotaciones precalculadas contra bases de datos instaladas localmente como InterPro [169], usando InterProScan [64] para identificar dominios de proteínas (Pfam) y mapear términos a vocabularios controlados y clasificaciones (términos GO). Se recuperaron recursos adicionales como estructuras cristalográficas 3D del *Protein Data Bank* (PDB) [170] mediante servicios web, y modelos estructurales (obtenidos por homología) descargados del sitio FTP de Modbase [171].

También se integraron varios *datasets* funcionales clave en esta versión, incluidos i) conjuntos de datos transcriptómicos que brindan evidencia de expresión génica en etapas del ciclo de vida o en condiciones experimentales que son relevantes para el descubrimiento de fármacos [79, 172–182]; y ii) *datasets* de esencialidad derivados de dos patógenos apicomplexa (*P. berghei* y *T. gondii*) [154, 183], que brindan información vital para ayudar a las estrategias de priorización.

3.4.2. Actualización de datos químicos

Para los compuestos bioactivos, el flujo de datos para la actualización se ha automatizado. La mayoría de los compuestos bioactivos se recuperaron de la versión 24 de ChEMBL [184], que contiene algunos *datasets* adicionales, incluyendo algunos ya introducidos en esta tesis, como las bibliotecas tripanocidas de GSK [185], y otras del mismo tenor pero para otros organismos, como la *MMV Pathogen box* [186]. El proceso de integración comienza con descripciones de moléculas

Type	Parent Group	Species	Strain	Source
Model	Plant	*Arabidopsis thaliana*		GenBank
Model	Helminth	*Caenorhabditis elegans*		GenBank
Model	Fungi	*Candida albicans*		GenBank
Model	Early Branching Eukaryote	*Dictyostelium discoideum*		GenBank
Model	Arthropoda	*Drosophila melanogaster*		GenBank
Model	Bacteria	*Escherichia coli*	K12	GenBank
Model	Mammal	*Homo sapiens*		GenBank
Model	Mammal	*Mus musculus*		GenBank
Model	Plant	*Oryza sativa*		GenBank
Model	Fungi	*Saccharomyces cerevisiae*	S288C	GenBank
Model	Helminth	*Schmidtea mediterranea*	S2F2	SmedGD
Pathogen	Apicomplexa	*Babesia bovis*	T2Bo	PiroplasmaDB
Pathogen	Apicomplexa	*Cryptosporidium hominis*	TU502	CryptoDB
Pathogen	Apicomplexa	*Cryptosporidium parvum*	Iowa II	CryptoDB
Pathogen	Early Branching Eukaryote	*Entamoeba histolytica*	HM-1:IMSS	AmoebaDB
Pathogen	Helminth	*Giardia lamblia*	isolate WB	GiardiaDB
Pathogen	Trypanosomatid	*Leishmania braziliensis*	MHOM/BR/75/M2903	TriTrypDB
Pathogen	Trypanosomatid	*Leishmania donovani*	BPK28	TriTrypDB
Pathogen	Trypanosomatid	*Leishmania infantum*	JPCM5	TriTrypDB
Pathogen	Trypanosomatid	*Leishmania major*	Friedlin	TriTrypDB
Pathogen	Trypanosomatid	*Leishmania mexicana*	MHOM/GT/2001/U1103	TriTrypDB
Pathogen	Bacteria	*Mycobacterium leprae*	TN	GenoList
Pathogen	Bacteria	*Mycobacterium tuberculosis*	H37Rv	GenoList
Pathogen	Apicomplexa	*Plasmodium berghei*	ANKA	PlasmoDB
Pathogen	Apicomplexa	*Plasmodium falciparum*	3D7	PlasmoDB
Pathogen	Apicomplexa	*Plasmodium knowlesi*	H	PlasmoDB
Pathogen	Apicomplexa	*Plasmodium vivax*	Sal-1	PlasmoDB
Pathogen	Apicomplexa	*Plasmodium yoelii*	17XNL	PlasmoDB
Pathogen	Apicomplexa	*Theileria parva*	Muguga	PiroplasmaDB
Pathogen	Apicomplexa	*Toxoplasma gondii*	ME49	ToxoDB
Pathogen	Early Branching Eukaryote	*Trichomonas vaginalis*	G3	TrichDB
Pathogen	Trypanosomatid	*Trypanosoma brucei*	TREU927	TriTrypDB
Pathogen	Trypanosomatid	*Trypanosoma cruzi*	CL Brener	TriTrypDB
Pathogen	Apicomplexa	*Neospora caninum*	Liverpool	ToxoDB
Pathogen	Trypanosomatid	*T. brucei gambiense*	DAL972	TriTrypDB
Pathogen	Trypanosomatid	*Trypanosoma congolense*	IL3000	TriTrypDB
Pathogen	Bacteria	*Chlamydia trachomatis*	D/UW-3/CX	GenBank
Pathogen	Helminth	*Onchocerca volvulus*	Cameroon	Wormbase Parasite
Pathogen	Helminth	*Loa Loa*	Cameroon	GenBank
Pathogen	Bacteria	*Mycobacterium ulcerans*	Agy99	GenBank
Pathogen	Bacteria	*Wolbachia Brugia malayi*	TRS	GenBank
Pathogen	Helminth	*Echinococcus granulosus*	G1	GeneDB
Pathogen	Helminth	*Echinococcus multilocularis*	Java	GeneDB
Pathogen	Helminth	*Schistosoma japonicum*	China/Anhui	GeneDB
Pathogen	Helminth	*Brugia malayi*	TRS	GenBank
Pathogen	Helminth	*Schistosoma mansoni*	Puerto Rico	GenBank
Pathogen	Bacteria	*Treponema pallidum*	Nichol	

Tabla 3.4 – Lista de especies, cepas y fuentes utilizadas para poblar la base de datos y la red quimiogenómica

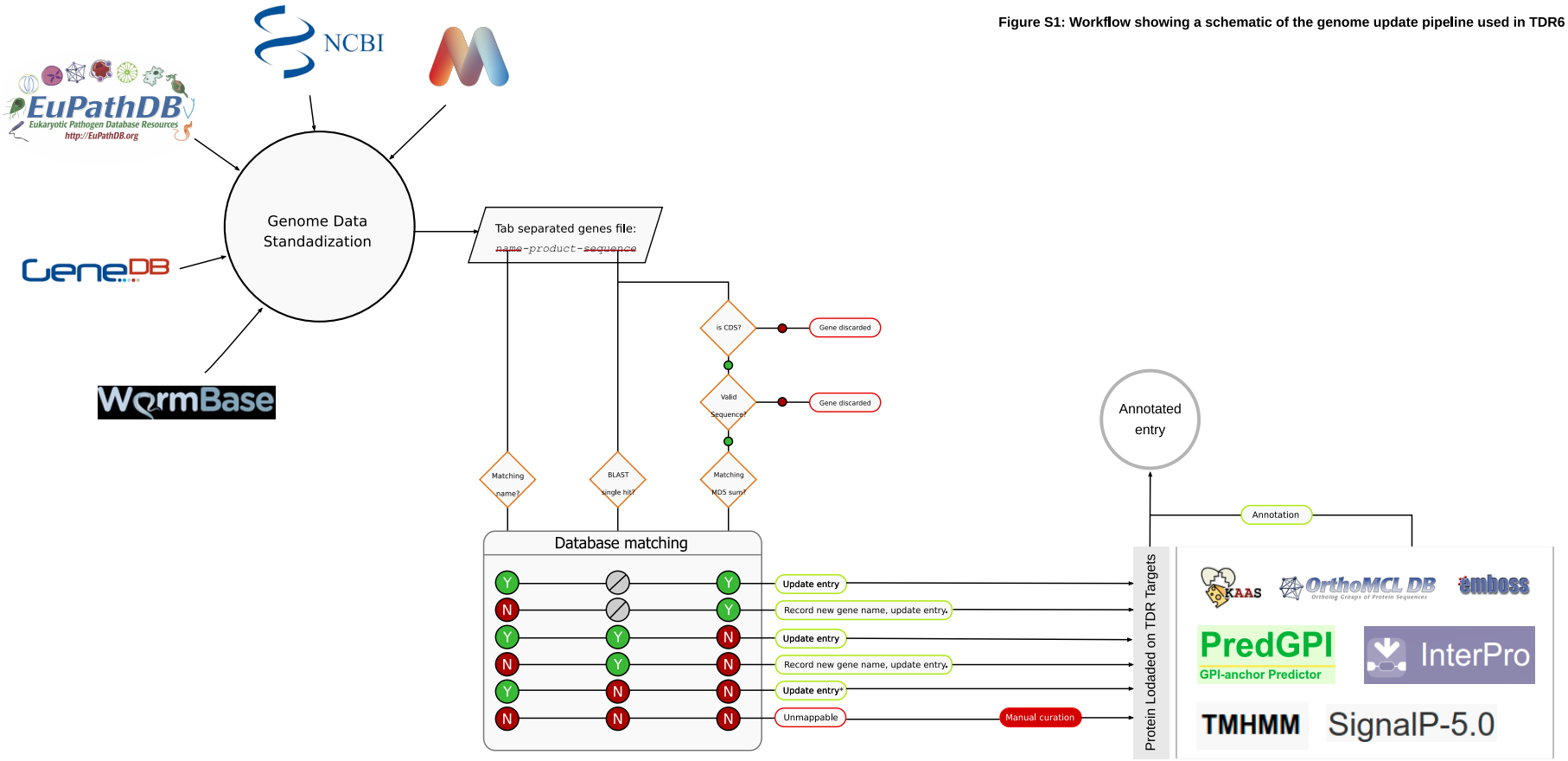


Figura 3.5 – Flujo de trabajo que muestra la rutina de actualización de genomas utilizada en TDR6. Al actualizar la base de datos, las entradas se actualizaron únicamente para almacenar datos adicionales (como fuente, URL externa y información sobre la cepa). Cuando los identificadores de genes entrantes y existentes (nombres de genes) no coinciden, pero las secuencias sí (ya sea a través de la suma de verificación o el *hit* de BLAST), la secuencia entrante es prioritaria y la entrada se actualiza, junto con el nuevo nombre que se le asigna (el antiguo nombre se almacena para propósitos históricos). (*) Cuando los nombres de genes en la base de datos y el conjunto de datos coinciden, la entrada se actualiza solo si la longitud de las secuencias es similar. Los casos que escapan a este procedimiento cuando la suma de verificación MD5 coincide con más de un gen o cuando los *hits* de BLAST son más de 1, se resolvieron usando el nombre como factor decisivo. En los casos en los que todo lo anterior no consiguiera obtener una única entidad para actualizar, la ambigüedad se resolvió por curación manual.

(2D) en formato SDF, a partir de las cuales se calculan todas las *features* estructurales posibles (requeridas para búsquedas de similitudes/subestructuras compuestas) utilizando CheckMol [187]. El flujo de datos también calcula propiedades químicas adicionales, como el coeficiente de partición octanol/agua y otros descriptores estructurales utilizando xLogp3 [188] y las herramientas obprop y obrotamer de Open Babel [125]. De la estructura también se obtuvieron los identificadores InChi e InChIKey [189] utilizados para el seguimiento de compuestos; y otras reglas generales estándar utilizadas en química medicinal y *drug discovery*, como la regla de cinco de Lipinski [190] y la regla de tres relacionada [191].

Todos los compuestos ingresados (y existentes) en TDR6 se sometieron a un cálculo de similitud química de todos contra todos utilizando ChemFP [106], que produce mediciones de similitud por pares basadas en el índice/distancia de Tanimoto [192]. Además, se calculó un mapa global (también de todos contra todos) de relaciones de subestructura entre compuestos en la base de datos (x es una subestructura de y; y es una superestructura de x). Sabiendo que el problema de encontrar subgrafos comunes máximos (MCS) entre moléculas es computacionalmente costoso, se aplicó un enfoque heurístico para encontrar subestructuras: El algoritmo obtiene primero un subconjunto de posibles moléculas candidatas haciendo uso de *fingerprints estructurales* previamente calculadas (al momento de ingresarlas a la base de datos, via Checkmol). Los candidatos a subestructura deben tener *fingerprints* coincidentes con su supuesta superstructura. Una vez que se obtiene una lista de candidatos, se realiza la determinación de la subestructura completa átomo por átomo utilizando MatchMol [187]. Los datos disponibles para los compuestos y las consultas que se pueden ejecutar en cada tipo de datos se resumen en la Tabla 3.2.

3.4.3. Curación e integración de datos de bioactividad

Al igual que con los compuestos químicos, la mayoría de las bioactividades integradas en TDR Targets provienen directamente de fuentes de datos anteriores (p. ej., ChEMBL). Al integrar los datos de bioactividad, conservamos tanto la anotación del ensayo (p. ej., "Ensayo de reducción de la motilidad *in vitro* contra las microfilarias de Brugia malayi a 10 µM") como el valor numérico y las unidades asociadas con las actividades del compuesto (p. ej., "80 % de inhibición", "1,5 µM IC_{50}", "10 nM MIC"), que son todos campos de búsqueda. Además, y para facilitar las consultas de los usuarios, las bioactividades notificadas se utilizaron para agrupar los compuestos ensayados en clases "activas" o "inactivas". Sin embargo, para minimizar el efecto de usar límites estrictos alrededor de umbrales arbitrarios y aumentar la separación entre clasificaciones activas/inactivas, también definimos un área gris "indeterminada", dentro de la cual los ensayos no fueran considerados ni activos ni inactivos, pero sirvieran de constancia de que el ensayo había sido realizado.

Sin embargo, no todos los tipos de actividad resultaron aptos para esta clasificación. A pesar de los esfuerzos en la estandarización de estos datos de actividad, es difícil interpretar las actividades de los compuestos a esta escala, ya que a menudo dependen del tipo de ensayo particular, las unidades reportadas y las condiciones particulares en las que se realizó cada ensayo. Sin embargo, un conjunto significativo de tipos de ensayos podría clasificarse automáticamente en categorías activas/indeterminadas/inactivas en función de los umbrales de actividad. Para esto, todos los tipos de ensayos con más de 100 000 reportes (Ver la Figura 3.6) se consideraron para la auditoría de actividad, aunque solo los ensayos basados en la concentración (como IC_{50}, K_i o Potencia) se consideraron observaron suficientemente robustos para tal determinación. En contraste, los ensayos basados en porcentajes (como % de actividad, % de actividad residual o inhibición) incurrían en ambigüedades en los informes de bioactividad. Los umbrales utilizados

para clasificar las actividades para cada tipo de ensayo se pueden encontrar en la Tabla 3.5, y la distribución de compuestos en estas clases de actividad se resume en la Figura 3.7.

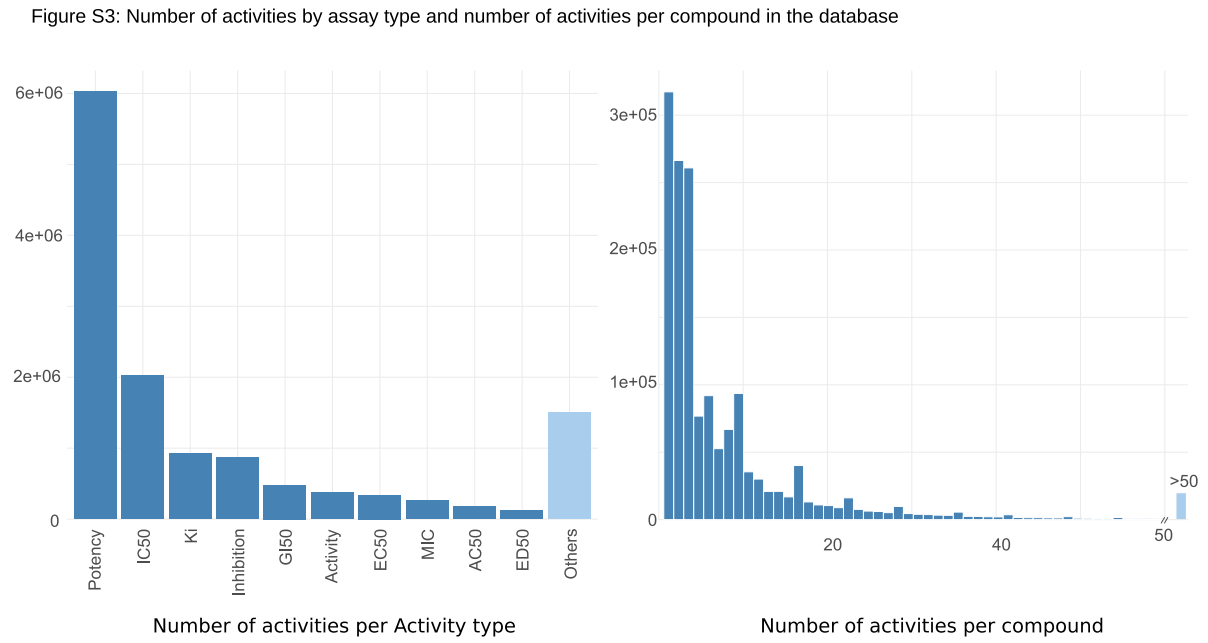


Figura 3.6 – Número de actividades por tipo de ensayo y número de actividades por compuesto en la base de datos: (Izq.) Se utilizaron todos los tipos de ensayo con más de 100000 reportes para ver las frecuencias de reporte entre los tipos de ensayo. Todos los demás tipos de ensayos (con menos de 100 000 informes) se agruparon como "otros". (Der.) Se construyó un histograma utilizando el número de actividades reportadas por cada compuesto con todos los compuestos con menos de 50 reportes. Todos aquellos con más de 50 fueron agrupados y agregados como una serie independiente.

El versión número 24 de ChEMBL cuenta con más de 15,2 millones de bioactividades reportadas, de las cuales solo alrededor de 6 millones corresponden a relaciones que involucran fármacos y blancos de proteínas (ya sean proteínas individuales, familias de proteínas y complejos proteicos, con ~ 93 % de proteínas individuales). Otras bioactividades restantes en la base de datos fueron reportes para una amplia variedad de blancos no proteicos, como células completas (3,6 M), organismos completos (2,2 M), tejidos (83 K) y macromoléculas no peptídicas (85 K) o moléculas pequeñas (menos de 100). Estas bioactividades no se usaron en la construcción de la red, dado que la misma está centrada en proteínas (es decir, blancos); aunque sí fueron integradas a la base de datos. La Figura 3.7 también muestra algunas visualizaciones de red de ejemplo que representan cómo TDR6 muestra estas bioactividades.

3.4.4. Integración de métricas derivadas del análisis de redes: Drogabilidad y Priorizaciones

Como se mencionó anteriormente, los datos genómicos, las anotaciones de genes, los compuestos químicos y las interacciones entre proteínas y fármacos se integraron en una red compleja orientada a al reposicionamiento de fármacos. La red se usó para calcular el *Network Druggability Score* (NDS) para todas las proteínas de patógenos prioritarios (*Tier* 1). El NDS está relacionado con la posibilidad de encontrar compuestos bioactivos en las inmediaciones del grafo de red de un blanco determinado (el rango es de 0 a 1). Brevemente, este algoritmo está basado en una prueba

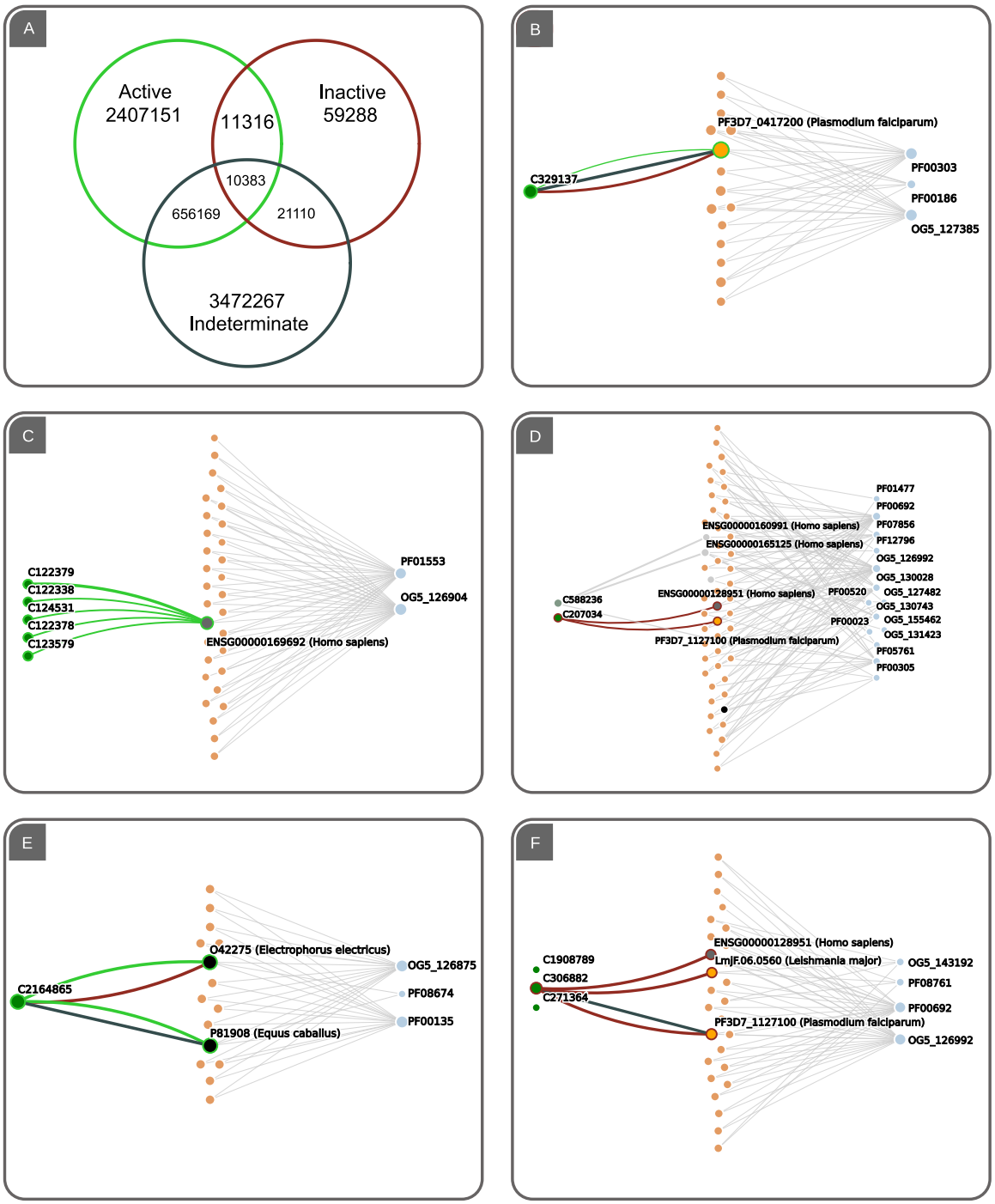


Figura 3.7 – Distribución de etiquetas de actividad y mezclas de evidencia entre los datos. A) Diagrama de Venn que muestra la distribución de valores de bioactividad en las clases activa, inactiva e indeterminada en TDR6. Las intersecciones cuentan los casos en los que la misma especie química tiene diferentes resultados de actividad contra el mismo objetivo. Se proporcionan ejemplos de estos casos en los paneles B a F (los IDs de compuestos representan identificadores TDR6). (B) Actividad de C329137 (una hidroxi-benzamidina) contra la dihidrofolato reductasa-timidilato sintasa bifuncional de *P. falciparum*. (C) Ejemplo de registros positivos para inhibidores de aciltransferasa humana. (D) Ejemplo de actividades negativas y neutras para los compuestos Trifenilcarbinol y Benzohidrol, respectivamente. Finalmente, tanto la evidencia positiva (E) como la negativa (F) pueden mezclarse con evidencia indeterminada, como se muestra para C2164865 probado contra colinesterasa de caballo y C306882 probado contra desoxiuridina 5'-trifosfato nucleótido hidrolasa recombinante de *P. falciparum*, respectivamente.

Assay type	Standard unit	Maximum admitted value for actives	Minimum admitted value for inactives
AC_{50}	nM	20000	100000
EC_{50}	nM	20000	100000
IC_{50}	nM	20000	100000
IC_{50}	$ug\ ml^{-1}$	15	50
K_d	nM	20000	100000
K_i	nM	20000	100000
Potency	nM	20000	100000

Tabla 3.5 – Tipos de ensayos y umbrales de actividad usados para la determinación de etiquetas de actividad: solo se usaron ensayos basados en la concentración para determinar las etiquetas de actividad. Las actividades que informaron menos del valor máximo admitido para positivos se consideraron interacciones activas (+), mientras que las que superaron el valor mínimo admitido para negativos se consideraron inactivas (-). Cualquier actividad reportada entre estos dos valores fue considerada como indeterminada (o)

de sobrerrepresentación de proteínas drogables conocidas anotadas, calculando una puntuación de relevancia (RS) para cada dominio Pfam y grupo de ortología de la red. El NDS para un blanco dado resulta de una suma acumulativa ponderada sobre los RS de todas las contribuciones de afiliación comunes al nodo objetivo y proteínas vecinas unidas a compuestos activos.

En TDR6, para facilitar la interpretación de las puntuaciones NDS, se realizó una evaluación estadística para identificar distintos grupos de drogabilidad (DG) en función de dos tipos de umbrales que ayudan a clasificar las predicciones de drogabilidad en zonas de confianza. Estos se ilustran en la Figura 3.8. Por un lado, mientras que todos los blancos de puntuación distintos de cero tienen cierto grado de conectividad con los blancos drogables conocidos, un NDS bajo sugiere que estas conexiones no constituyen relevancia para la drogabilidad. Por lo tanto, se considera un corte de ruido (una línea de base calculada como 5 veces el valor del percentil 0,25 de la distribución NDS completa) para identificar blancos de puntuación baja. El segundo umbral se deriva del índice máximo J de Youden [193], que se calcula como la puntuación en la que tanto la especificidad (el porcentaje de veces que el modelo acierta al decir que algo es drogable o no drogable para un un valor de umbral dado) como la sensibilidad (el porcentaje de blancos drogables detectados para un umbral determinado) son óptimas. En rigor, este umbral supone un compromiso que permite hallar el valor del umbral en el que se obtiene la mejor sensibilidad sin comprometer la especificidad, y viceversa. Dado que para calcularlo es preciso contar con verdaderos positivos (blancos drogables confirmados), este valor solo se puede calcular para algunos patógenos del *Tier* 1. Un mínimo arbitrario de 10 verdaderos positivos se consideró suficiente para la determinación del límite de Youden. Para los patógenos que no alcanzaron este número de verdaderos positivos, se utilizó un límite global de Youden (calculado utilizando todos los verdaderos positivos en la red). Los Grupos de drogabilidad correspondientes son así: DG1 para blancos con valores de NDS que van desde o hasta el umbral de ruido; DG2 para blancos con

valores NDS que van desde el umbral de ruido hasta el límite de Youden; y DG 3, 4 y 5 con valores NDS que son 1, 10 y 100 veces superiores al límite de Youden. En consecuencia, estos últimos grupos constituyen los blancos susceptibles de tratamiento con fármacos más probables. La figura 6 muestra un ejemplo de una priorización impulsada por la red para *Mycobacterium ulcerans*. Todas las priorizaciones para los organismos prioritarios de TDR Targets se pueden ver en línea en la página de resumen de datos para cada especie (ver https://tdrtargets.org/datasummary, haciendo clic en la especie de interés).

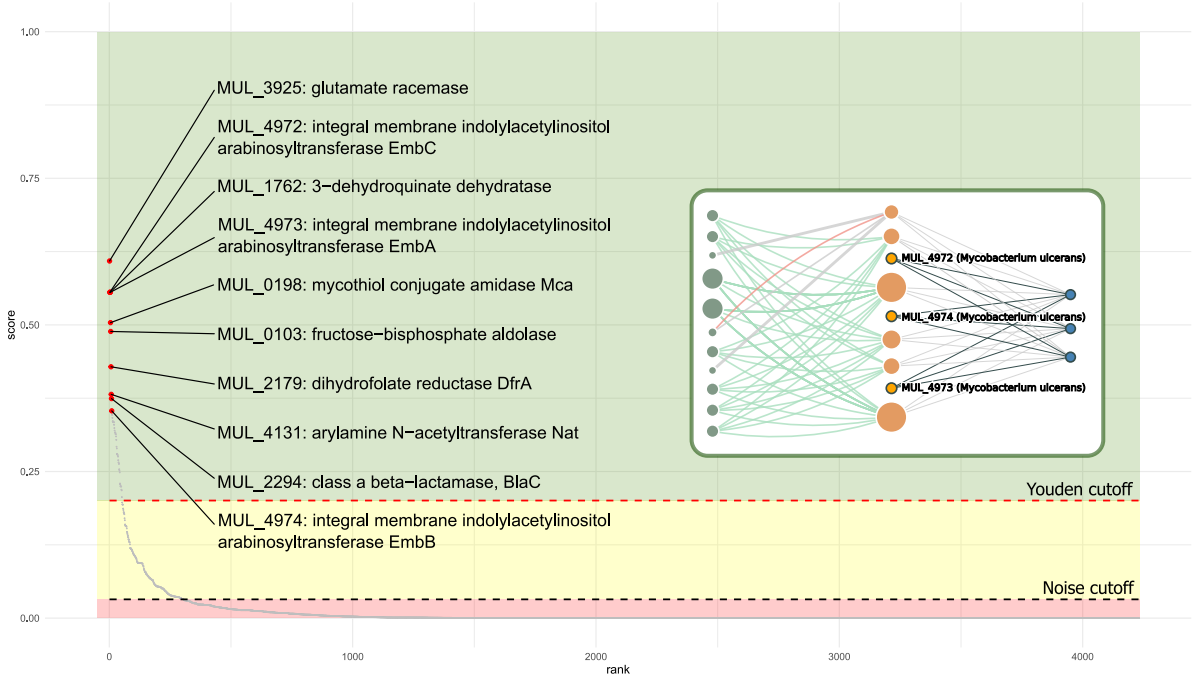


Figura 3.8 – Priorización de blancos del genoma completo por NDS para M. ulcerans: los blancos putativos en el genoma de *M. ulcerans* se clasificaron por su NDS (puntuación de drogabilidad). El gráfico muestra todos los blancos del genoma ordenados de mayor a menor puntuación (en el eje x) junto con el NDS correspondiente (en el eje y). Los puntos rojos corresponden a los 10 blancos mejor ponderados, con etiquetas que indican el nombre del gen y el producto. Al examinar la priorización del genoma completo desde el resumen de datos, el usuario puede acceder a una página de genes haciendo clic en ella en el gráfico de priorización. Se muestra un ejemplo de subgrafo de la familia de genes EmbA/EmbB/EmbC (atal cual se apreciaría en las páginas de genes correspondientes). La figura también muestra zonas de confianza, DG1 (rojo): delimitado por cero y corte de ruido; DG2 (amarillo): entre el ruido y el corte de Youden; y DG3-5: Con puntuaciones superiores al corte de Youden.

Estas priorizaciones pueden funcionar en ambos sentidos. Cuando se parte de un compuesto de interés, el algoritmo puede priorizar los blancos, utilizando la similitud ponderada de los vecinos químicos con los blancos putativos iniciales. Y cuando comienza desde el blanco de interés, puede priorizar los compuestos, utilizando blancos drogables vecinos conectados a compuestos y luego siguiendo enlaces ponderados a inhibidores/fármacos candidatos.

3.4.5. Visualizaciones dinámicas de sub-grafos del vecindario filogenético y químico

Los subgrafos de red para compuestos y proteínas (y sus respectivas puntuaciones NDS) se pueden explorar desde la aplicación web utilizando un fármaco o un blanco como punto de partida para obtener sugerencias sobre fármacos no probados o nuevos blancos terapéuticos, respectivamente.

A través de las visualizaciones interactivas introducidas en TDR6, los usuarios pueden consultar el vecindarios de la red en torno a las entidades de interés en las páginas correspondientes. Las listas de interacciones putativas derivadas de la red también se pueden explorar en formato tabular en las secciones "*Druggability*"(para blancos) y "*Known and predicted targets*"(para fármacos).

Estas visualizaciones están impulsadas por D3.js [194] usando *force fields* para visualizaciones de grafos. Dentro del panel de subgrafo D3, los usuarios pueden realizar búsquedas de nodos dentro del gráfico (identificadores de blancos), así como alternar la visibilidad de los blancos especie por especie y personalizar la opacidad de los nodos. En conjunto, estas nuevas funciones brindan una visualización clara y completa de la subred construída para cada blanco o droga, lo que permite a los usuarios manipular las visualizaciones mientras exploran los datos.

3.4.6. Interfaz Gráfica

La interfaz de usuario (UI) y las herramientas disponibles para el reposicionamiento de fármacos y la priorización de blancos han pasado por una importante actualización. En primer lugar, se ha rediseñado la UI bajo estándares del W3C para lograr una aplicación más saludable y escalable. Se integraron los frameworks Bootstrap (https://getbootstrap.com/) y jQuery (https://jquery.com) en el desarrollo y diseño de la aplicación web TDR6 y en la funcionalidad front-end. Para consultas de estructuras de compuestos, hemos licenciado e implementado la aplicación de dibujo químico Marvin JS de Chemaxon (`https://chemaxon.com/products/marvin-js`). Los registros tabulados dentro de las páginas de proteínas y de moléculas pequeñas ahora usan el complemento jquery de JavaScript de DataTable (https://datatables.net) para crear fácilmente utilidades de paginación, filtrado y clasificación. Finalmente, las representaciones 2D de compuestos ahora se generan automáticamente utilizando una implementación del módulo javascript SmilesDrawer [195], previniendo así la necesidad de generar imágenes con estas representaciones y reduciendo cuantiosamente los datos descargados al mostrar una página.

3.4.7. Arquitectura de la solución

Todas las funcionalidades nuevas fueron extensiones de la arquitectura existente. La misma podría interpretarse como un *Model-View-Controller* (MVC) siriviendo de dominio en un paradigma *Domain Driven Design* (DDD), muy común en aplicativos web por su facilidad para separar/escalar la infraestructura, y la compatibilidad para re-utilizar la capa de dominio para distintas vistas o presentaciones (por ejemplo, una aplicación web y una REST API). La capa de dominio se desarrolló utilizando un framework MVC de Perl, Catalyst, que como su nombre lo indica, separa la lógica del aplicativo en Modelos, Controladores y Vistas. Los modelos son interfaces de conexión con una o más bases de datos, y facilitan la obtención y persistencia de datos mediante el uso de controladores. Éstos, a su vez, nuclean toda la lógica necesaria para procesar los *requests* del cliente y devolver las respuestas acordes. Dicha respuesta es gestionada por una o más vistas que, en última instancia, son las responsables de convertir los eventos de interacción del usuario en *requests HTTP* delegados a los controladores, y de renderizar la respuesta que éstos ofrecen en la interfaz gráfica, explicada más arriba.

En materia de infraestructura, debe destacarse además que el aplicativo utiliza 3 servidores distintos con funciones diferenciales: Un Datastore, encargado de administrar la base de datos (MySQL), un *host* primario, que atiende en forma directa todos los procesos disparados por el aplicativo; y un *host* secundario, encargado de dar respuesta a procesos encolados (de alto costo

computacional), como transformaciones de listas, cuyo tiempo de respuesta es muy lento y no puede servirse en tiempo real. Finalmente, debe incluirse como parte de la arquitectura a todo el hardware y software dedicado a la colección, armonización y persistencia de datos en la base de datos MySQL. La figura 3.9 resume todos estos elementos y esquematiza la forma en la que se contactan entre sí.

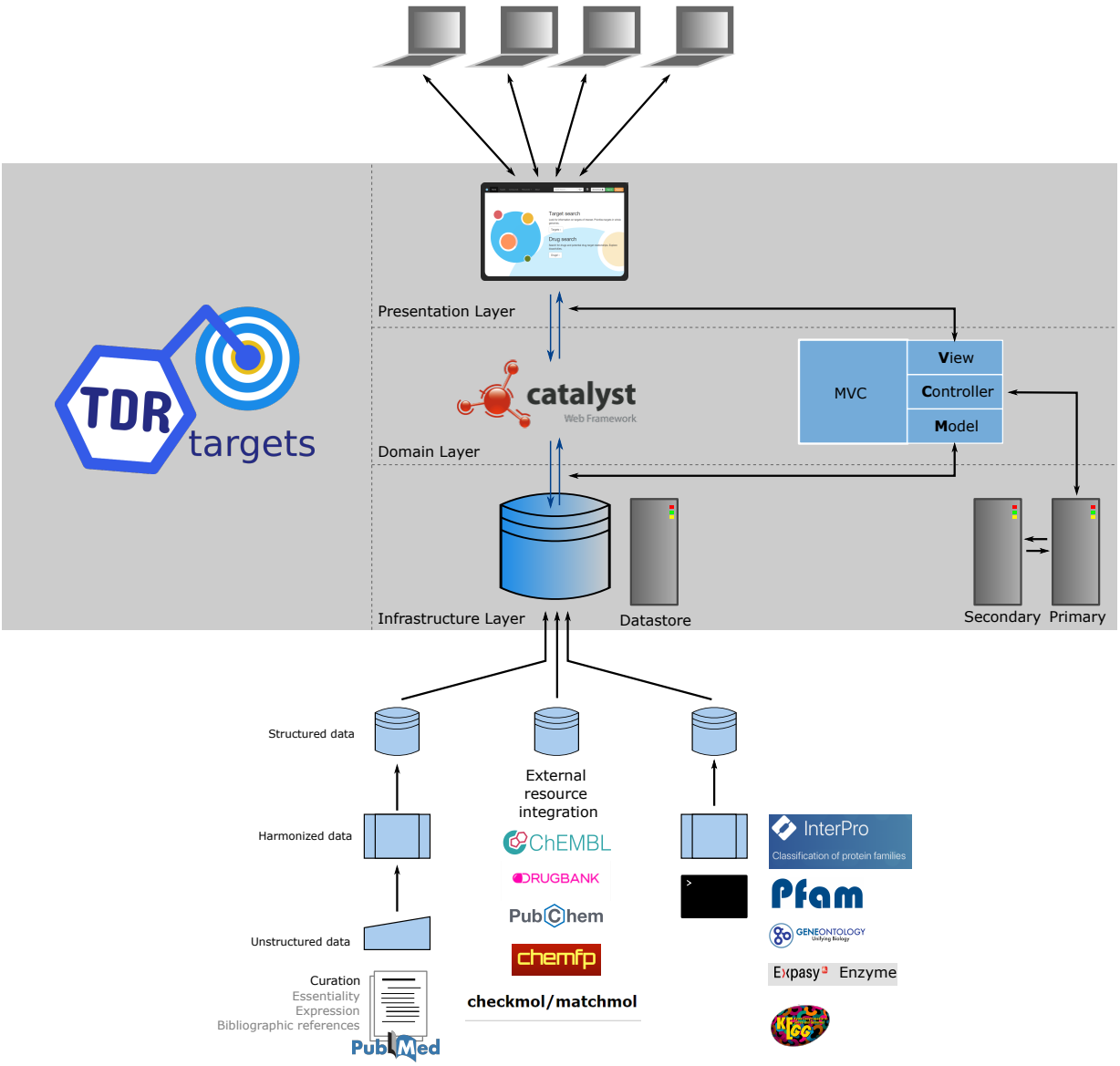


Figura 3.9 – Arquitectura de procesos de ingesta, almacenamiento y servicio de TDR Targets. Se muestra de forma esquemática la forma en la que se preprocesa e ingestan los datos al *Datastore*, que a su vez sirve opera como base de datos transaccional para el software MVC *Catalyst* de TDR Targets a través del *Model*. El mismo provee de datos al *Controller*, que atiende todos los *requests* inmediatos en un servidor *Primary*, pero delega las cargas de trabajo pesado en un servidor *Secondary* para diferir procesos lentos. Finalmente, la lógica de las Vistas (*View*) post-procesa los datos y los muestra al cliente que solicitó los datos

La arquitectura de esta solución tiene ciertas ventajas y desventajas que pueden discutirse largamente, pero fundamentalmente favorece la robustez de la herramienta a costa de su capacidad de cambio o adaptabilidad [196, 197]: La relacionabilidad forzosa de los datos (esto es, la relación pre-existente entre las entidades que componen el repositorio de datos) impone

que éstos deban ser armonizados antes de integrarse al repositorio general. Además, la estructura monolítica del sistema favorece la manutención del servicio pero también limita su escalabilidad horizontal: si una herramienta es mucho más usada que otra, es imposible destinar más poder de cómputo a ésta que a las demás. La incorporación de un servicio secundario de cómputo suple rudimentariamente esta limitación, aunque es incapaz de actuar como un verdadero balanceador de cargas. La estructura monolítica también dificulta la incorporación de nuevas funcionalidades o la modificación de las existentes, dado el alto grado de acoplamiento que existen entre los distintos módulos que componen la herramienta.

3.5. Discusión

Los nuevos datos, la interfaz y la funcionalidad de TDR6 brindan a los usuarios una mejor navegación y visualización de blancos y compuestos.

El modelo de red actual conecta proteínas a través de la afiliación de entidades (proteínas) a anotaciones (dominios Pfam, grupos de ortología). Estos han sido seleccionados en base a su amplia cobertura y relativa facilidad de cálculo u obtención. Los usuarios pueden complementar estos conceptos con otros criterios importantes para la validación de blancos terapéuticos (esencialidad, expresión en etapas relevantes del ciclo de vida) usando las herramientas y la funcionalidad proporcionada por TDR6. En el futuro, pueden construirse otras redes que integren estos otros criterios en el propio modelo de red subyacente, aunque el desafío verdadero para esto es el de disponer con los datos; en especial para organismos que no son pasibles de ser analizados por metodologías masivas o de alto rendimiento.

Un aspecto importante al priorizar compuestos para probar en el laboratorio es su disponibilidad comercial. En TDR6 ahora se implementó temporalmente un servicio que permitía mostrar información sobre la disponibilidad comercial de los compuestos. Esta función fue posible gracias a una vinculación estratégica con Molport (un *market* químico en línea que obtiene compuestos de los principales proveedores del mundo) y mostramos a los usuarios un indicador visual en las páginas de compuestos que brindan una vista rápida sobre si el compuesto está en stock o si se puede hacer a pedido. Sin embargo, la implementación realizada solo permitía ver la disponibilidad comercial de compuestos durante la navegación, y no podía ser utilizada para búsquedas o filtros. Además, el servicio provisto por Molport para obtener la disponibilidad comercial utilizaba límite de consultas por tiempo definido y éstas solían agotarse más bien rápidamente. Esto llevó a que la funcionalidad fuera deprecada y eliminada de la versión que está ahora desplegada.

Se necesitan varias mejoras clave para mantener TDR Targets relevante para la comunidad de científicos que trabajan en enfermedades tropicales. La integración de metabolitos naturales y la conexión de estas pequeñas moléculas con otros compuestos bioactivos a través de subestructuras compartidas o por similitud química será un enfoque importante en el futuro. Esto permitirá navegar por el grafo de compuestos-blancos utilizando también a las reacciones bioquímicas como criterios de conexión, dado naturalmente conectan enzimas no ortólogas a través de sus sustratos/productos y cofactores compartidos.

Finalmente, como ya se mencionó antes [149], todavía hay una gran brecha de curación que debe llenarse. Muchos compuestos bioactivos han sido probados por la comunidad de investigadores que trabajan en Enfermedades Tropicales Desatendidas. Sin embargo, muchos de estos ensayos y resultados se informan en revistas fuera de las principales revistas de Química Medicinal y, por

lo tanto, se pasan por alto en los grandes esfuerzos de curación que llevan adelante consorcios como el ChEMBL [184]. La conservación e integración de estos datos faltantes (¡incluidos los datos negativos!) debería ser una prioridad para la comunidad, ya que ahorraría tiempo y recursos valiosos, y permitiría la construcción de más y mejores modelos predictivos para asistir al proceso de descubrimiento de drogas.

4. Reposicionamiento de fármacos

4.1. Reposicionamiento utilizando TDR Targets

El reposicionamiento de fármacos para *drug discovery* en enfermedades desatendidas ha demostrado ser eficaz para encontrar posibles agentes terapéuticos [198]. Para tripanosomátidos en general, algunas drogas reposicionadas son actualmente tratamientos viables (Eflornitina, Tamoxifeno)[198, 199] mientras que para la enfermedad de Chagas, en particular, algunos de estos agentes han alcanzado ensayos clínicos, aunque sin resultados satisfactorios (Posaconazol)[200]. *In silico*, se ha demostrado que un enfoque de redes complejas para la reutilización de fármacos, utilizando estrategias intensivas de integración de datos quimiogenómicos, permite obtener compuestos bioactivos conocidos y blancos farmacológicos para un patógeno de interés, así como también recomendar nuevos pares de droga-blanco para una ulterior validación experimental [152]. También mostramos en el capítulo anterior, cómo la integración de estas redes en un repositorio de datos de acceso público permite interrogar dichos datos de forma integral.

En este capítulo, presentamos un caso de uso de la base de datos TDR Targets [2] para la priorización de blancos y la recomendación de fármacos, seguida de una estrategia de filtrado *in silico* que considera la novedad de la molécula, la disponibilidad comercial de la misma y la presencia de grupos funcionales para generar listas reducidas de moléculas de interés. Como resultado, este flujo de trabajo culminó con la obtención de 18 bibliotecas, cada una de ellas recopilando compuestos con un *scaffold* conocido diferente. Para la validación experimental, se seleccionaron dos de estas bibliotecas para realizar una curación manual de compuestos culminando con una biblioteca de 28 piperazinas y otra que con 15 compuestos nitro. De un total de 43 compuestos, 21 fueron adquiridos para ensayos *in vitro*. Durante la validación experimental, se obtuvieron 7 compuestos con actividad tripanocida y baja citotoxicidad.

La disponibilidad comercial o la accesibilidad sintética de los compuestos es un aspecto clave en *drug discovery*. De estos dos aspectos la accesibilidad sintética – factibilidad de sintetizar un determinado compuesto – puede ser considerada como validada al usar bases de datos de compuestos bioactivos (como ChEMBL) como punto de inicio de las priorizaciones. Esto es así dado que los compuestos presentes en ChEMBL tienen todos bioensayos asociados, en donde la actividad de la molécula es medida contra organismos enteros, células o blancos moleculares discretos. Sin embargo, la disponibilidad comercial es clave, ya que puede acelerar significativamente los proyectos de reposicionamiento. En este aspecto la industria química evolucionó muy rápidamente en las ultimas décadas [201], logrando cubrir espacios químicos diversos [202]. Entre estos proveedores comerciales se destaca Molport que es un servicio que consolida y acopia compuestos de diversos otros proveedores de pequeños compuestos para screening [1]. En particular, Molport facilita las tareas de priorización y selección quimioinformática mediante la provisión de archivos descargables conteniendo estructuras o mediante el acceso programático utilizando a través de una interfaz de programación de aplicaciones (API).

[1] https://www.molport.com/shop/online-chemical-shop-suppliers.

4.1.1. Introducción

Como mostramos en el capítulo anterior, es posible crear bibliotecas de *screening* utilizando TDR Targets, ya que permite no solo la priorización de blancos, sino también la expansión de series de compuestos mediante relaciones existentes y recomendadas entre entidades dentro de la red. Esto significa que, para cualquier blanco de interés, se pueden encontrar inhibidores potenciales basándose en la evidencia obtenida para proteínas relacionadas a ésta en otros organismos; así como también compuestos estructuralmente similares a un inhibidor conocido, basándose en cálculos de proximidad y relevancia dentro de la red.

En este punto, es posible que existan tantas estrategias como investigadores usando la herramienta. Una de las grandes fortalezas de TDR Targets subyace justamente en la flexibilidad con la que puedan generarse las priorizaciones. Esta ha sido su impronta desde la primera versión, en 2008 [29]; y continúa siéndolo en su versión más reciente. Así, pueden hallarse en la literatura casos de uso que van desde el mapeo de genes ortólogos para la identificación de genes novedosos en *Plasmodium yoelii* [203] (sin afán directo de asociar esta búsqueda al desarrollo o descubrimiento de drogas), a la identificación de posibles blancos de acción en *M. tuberculosis* para el decoquinato [204] una droga de uso veterinario indicada para el tratamiento de la coccidiosis que también posee aptitudes como mycobactericida [205]. También hay casos de uso para este recurso en la priorización de blancos esenciales en organismos más complejos, como *Schistosoma mansoni*, por ejemplo, usando conexiones por ortología para obtener una lista mínima de genes asociados el mantenimiento de la fertilidad en *C. elegans* y al desarrollo en *D. melanogaster* [206]. Para *T. cruzi*, existe un ejemplo de uso para la identificación de blancos potenciales para γ-Lactonas obtenidas de extractos naturales, asistiendo en el proceso de identificación de mecanismos de acción para inhibidores conocidos [207].

Los casos de uso son variados no solo por el organismo de interés, sino porque las estrategias de filtrado y priorización, en cada caso, responden al conocimiento de los investigadores acerca del patógeno, su ciclo de vida, metabolismo, etc; o la existencia de inhibidores para uno o más blancos desde donde partir o al cual dirigirse. Si bien este conocimiento podría estructurarse y servirse como parte de la base de datos, las combinaciones serían infinitas. No obstante, hay varios post-procesamientos problemáticos a la hora de generar bibliotecas utilizando *solo* TDR Targets. Por ejemplo, la imposibilidad de filtrar por disponibilidad comercial, o de quitar moléculas promiscuas (PAINs) de forma automática, agrupar compuestos similares para generar series o bibliotecas ordenadas, entre otras cosas. Estas funcionalidades solo se hacen evidentes al momento de utilizar el recurso de punta a punta, y podrían ser incluidas en futuras versiones de TDR Targets. Asumiendo estas limitaciones, en el presente capítulo se explorará un caso de uso que utiliza priorizaciones básicas de proteínas como punto de partida, pero en donde todos los post-procesamientos y análisis ulteriores de las bibliotecas generadas fueron realizados a partir de los *datasets* armonizados crudos (exportados, eso sí, desde TDR Targets).

4.1.2. Resultados

Exploración de datos y estrategia de filtrado

En un primer análisis exploratorio, obtuvimos una lista de blancos de *T. cruzi* a partir de TDR Targets utilizando un Network Druggability Score ≥ 4. De la lista de blancos priorizados (327 genes), se obtuvo un conjunto de compuestos mediante consultas de TDR Targets de transformación de blancos. Este conjunto contenía 180.023 moléculas. A partir de ese momento,

los datos se exportaron a un archivo de texto estructurado (CSV). De estas 180.023 moléculas 622 ya habían sido ensayadas contra tripanosomátidos; y otras 1.879 estaban a su vez presentes en DrugBank – una base de datos de drogas aprobadas para uso clínico (drogas comercializadas, actuales y/o retiradas). De las 622 ya ensayadas en trypanosomatidos solo 82 estaban también en DrugBank; dejando un total de 1.797 compuestos que serían novedosos en cuanto a no estar representados en DrugBank y no haber sido previamente ensayados en trypanosomatidos. Lamentablemente, estos compuestos no estaban disponibles comercialmente (en forma pura) en Molport. Esto resalta por un lado la dificultad de acceso a compuestos que pueden ser de interés, para el diseño y armado de bibliotecas de *screening* experimental.

La disponibilidad comercial de compuestos es un gran problema cuando se crean bibliotecas químicas, dado que muchos compuestos deben ser sintetizados a pedido o directamente no pueden adquirirse. Para abordar cuantitativamente este problema, se validó la disponibilidad de los compuestos en forma programática utilizando la interfaz de programación de aplicaciones (API) de Molport para el conjunto completo de 180.023 compuestos. Al momento de realizar estas búsquedas, encontramos que solo 28.802 (~15 %) estaban disponibles para adquirir en forma directa o por pedido especial. Todas las bibliotecas mencionadas a continuación se construyeron utilizando los 28.802 compuestos con esta disponibilidad comercial.

Los blancos en el *dataset* exportado se filtraron aún más, seleccionando solo aquellos anotados como parte de las vías de metabolismo energético y metabolismo de aminoácidos, lo que redujo la cantidad de genes totales (327) a aproximadamente el 10 % (44). Esta fue una decisión arbitraria pero informada, basada en la evidencia de que el metabolismo de carbono se ve notoriamente exacerbado en células infectadas con amastigotes [208] y en el papel de algunos aminoácidos (y metabolitos intermedios) en la supervivencia del parásito durante la infección [209]. Luego de esta sub-selección, el número restante de compuestos asociados a la lista reducida de blancos fue de 4.041. El número de blancos proteicos en el dataset pasó de 327 a 44.

Estrategias de clustering para exploración del espacio químico

A partir de la lista completa de compuestos en formato SMILES, se preparó un dataset conteniendo sus identificadores (InChI e InChI-key), descriptores moleculares, propiedades fisicoquímicas y fingerprints (MACC Keys). Con este último, se obtuvo la similitud estructural (índice de Tanimoto) entre todas las especies químicas. Como puede verse en la Figura 4.1-a), las distribuciones de cada descriptor son muy homogéneas, sin ningún sesgo relevante. Más aun, los descriptores moleculares se utilizaron para realizar un análisis de componentes principales (PCA). En este análisis se obtienen combinaciones lineales de las variables originales (descriptores moleculares) que como resultado dan un número final menor de variables (*componentes*) [210]. Las componentes principales son aquellas combinaciones de variables que capturan la mayor parte de la varianza de los datos. Además de servir para reducir la dimensionalidad de los datos originales, el análisis PCA sirve para explorar rápidamente relaciones entre múltiples variables. En la gráfica de componentes principales de la Figura 4.1-b mostramos la distribución de algunos descriptores para las distintas especies químicas. Los autovectores de estos descriptores (flechas de color, *eigenvectors*) dan una intuición de la importancia de cada descriptor en las componentes elegidas (longitud o peso del vector) y la dirección en la que intentan segregar el dataset. No obstante, como puede observarse por la distribución homogénea de los datos (puntos azules) en las dos componentes principales, no se hallaron grupos especialmente separados o interesantes, lo que sugiere que no habría sesgos notables en la composición de la biblioteca.

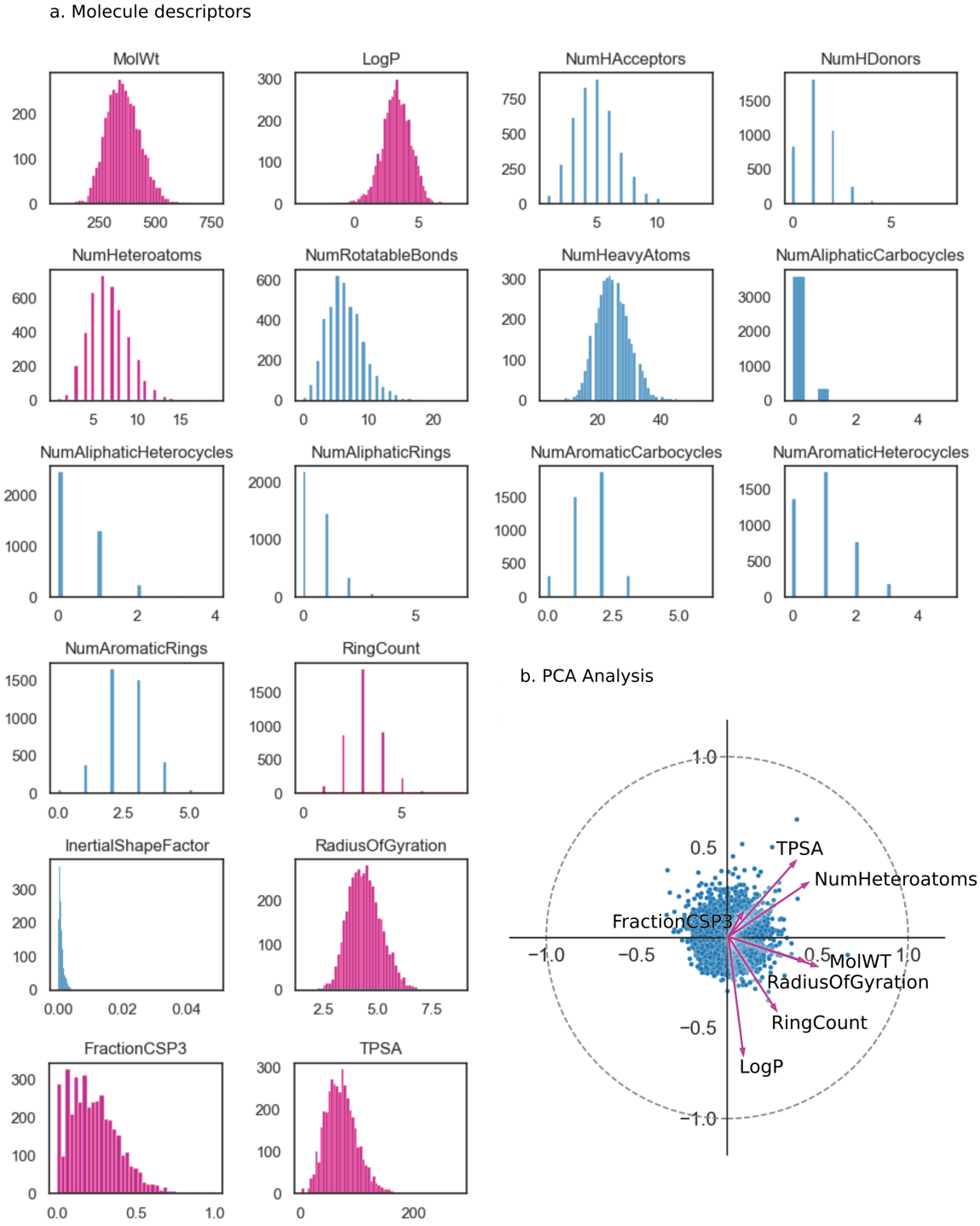


Figura 4.1 – Descriptores moleculares y PCA. (a) Distribución de los valores de distintos descriptores obtenidos para la biblioteca de compuestos: `MolWt`: Peso Molecular, `LogP`: Coeficiente de partición octanol-agua, `NumHAcceptors`: Número de átomos aceptores de puente de hidrógeno, `NumHDonors`: Número de átomos dadores de puente de hidrógeno, `NumHeteroatoms`: Número de heteroátomos, `NumRotatableBonds`: Número de enlaces rotables, `NumHeavyAtoms`: Número de átomos pesados, `NumAliphaticCarbocycles`: Número de Carbociclos alifáticos, `NumAliphaticHeterocycles`: Número de Heterociclos alifáticos, `NumAliphaticRings`: Número de anillos alifáticos, `NumAromaticCarbocycles`: Número de carbociclos aromáticos, `NumAromaticHeterocycles`: Número de heterociclos aromáticos, `NumAromaticRings`: Número de anillos aromáticos, `RingCount`: Número de anillos, `FractionCSP3`: Fracción de carbonos sp3, `TPSA`: Superficie topológica polar, `InertialShapeFactor`: Factor incercial de forma, `RadiusOfGyration`: Radio de giro. (b) Análisis de componentes principales (PCA) a partir de algunos de los descriptores (denotados en fucsia). Las flechas que surgen desde el origen $(0, 0)$ son los autovectores resultantes de la proyección de cada descriptor en las componentes principales 1 y 2.

El análisis de componentes principales (PCA) mostró que el espacio químico explorado por los 4.041 compuestos hasta ahora colectados es amplio, pero no ofrece pistas adicionales sobre cómo particionar la colección para estudiar su actividad *in vitro*. Como alternativa ortogonal al PCA, se evaluaron distintos análisis de agrupamiento (*clustering*). Primero se utilizó tSNE [211] para generar colecciones minoritarias usando la matriz de distancias construída a partir de MACC Keys, seguido de un agrupamiento por k-means. Dado que el resultado del *clustering* es altamente sensible al parámetro de *perplejidad* del algoritmo tSNE, se exploraron varios valores de este hiperparámetro (2, 5, 10, 25, 40, 50) y varios números de *clusters* (con $k \in \{3 \dots 50\}$) para k-means. El resultado de este análisis de *clustering* puede verse en la figura 4.2. Si bien tSNE hace un mejor trabajo separando a la biblioteca en colecciones más pequeñas, debe notarse que el resultado es nuevamente subóptimo: el análisis de Indice de Siluetas muestra índices negativos para la mayoría de los clusters. La biblioteca hasta ahora generada es entonces mayormente indivisible y al menos *a priori*, no parece fácil armar inteligiblemente sub-bibliotecas más pequeñas para *screening in vitro*.

Ante la imposibilidad de obtener clusters por PCA o tSNE que permitieran separar satisfactoriamente las especies según su estructura y propiedades fisicoquímicas, se optó por realizar un *clustering* jerárquico *single linkage* para crear microclusters que contuvieran series de moléculas similares, agrupando así compuestos con altísima similitud que pudieran conformar una serie entre sí. Para obtener el umbral de corte para el *clustering* jerárquico se sondearon 4 umbrales de distancia (t = threshold; $t \in [0{,}8\ 1{,}6\ 2{,}4\ 3{,}2]$). Los clusters se inspeccionaron manualmente para relevar la calidad de los agrupamientos en cada caso. En líneas generales, se observaron buenos resultados con $t = 0{,}8$ que genera un total de 130 microclusters; el más grande de ellos con 5 miembros. Para $t = 1{,}6$ los resultados parecen mejores, con 325 microclusters y el más voluminoso de éstos presentando 9 miembros. Sin embargo, la inspección manual sugiere que, en algunos casos estos clusters más grandes podrían escindirse en al menos 2 sub-grupos. Los clusters generados con $t > 1{,}6$ mostraron alta disimilitud interna. La figura 4.3 muestra cómo evoluciona un cluster al incrementar el umbral de corte. Finalmente, se seleccionó $t = 0{,}8$ como umbral de corte, y se prosiguió con la generación de las bibliotecas.

Remoción de ruido

Dado que TDR Targets no cuenta con una forma integrada para el análisis y remoción de PAINs, la obtención de microclusters mostró que ciertos clusters eran sumamente promiscuos (es decir, presentaban actividad putativa contra múltiples familias de blancos). Para removerlos se eliminaron del dataset todos aquellos con demasiadas interacciones putativas (> 10, es decir, moléculas con muchos blancos de distintas familias). A su vez, los blancos pasaron también por el mismo escrutinio, siendo eliminados aquellos cuyo grupo de ortología presentara asociación con una gran cantidad de micro-clusters de compuestos. Es importante enfatizar que las asociaciones se buscaron entre microclusters y familias de blancos (y no uno a uno entre moléculas y blancos), por lo que si una molécula estuviera potencialmente asociada a más de 10 proteínas de la misma familia, no sería marcada como promiscua y por lo tanto no sería removida del dataset. Lo mismo rige para un blanco con asociación a más de 10 moléculas de la misma serie. Así, después de la eliminación de estas entidades, el número total de compuestos se redujo a 385.

Para saldar esta necesidad es que preferimos utilizar $t = 0{,}8$ por sobre $t = 1{,}6$ a pesar de que ambos son resultados muy buenos: tener *clusters* cuyos integrantes son similares pero que por inspección manual dan cuenta de más de una familia de compuestos, podría incurrir en una limpieza de PAINs demasiado exigente, dado que un microcluster nuclearía actividades de *scaffolds* distintos.

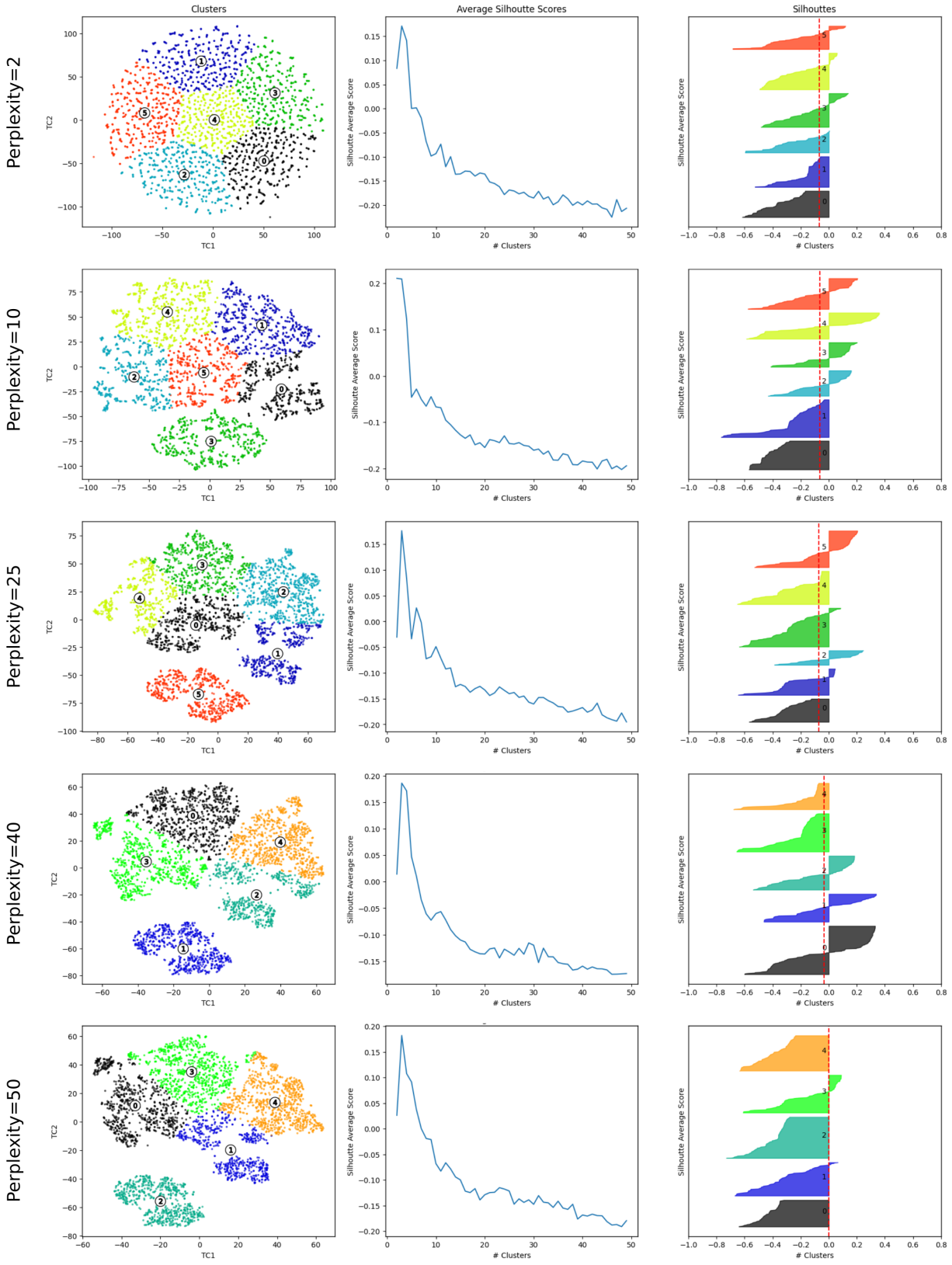


Figura 4.2 – t-Distributed Stochastic Neighbor Embedding (tSNE) y agrupamiento por k-means. Pruebas de reducción de dimensionalidad para identificar patrones estructurales que guien la segmentación del dataset en bibliotecas más pequeñas. Para tSNE, se ensayaron distintos valores de `perplexity` con una cantidad fija de iteraciones (1000) y ángulo también fijo (0.3). Para k-means, se barrió un rango de 3 a 50 clusters, registrando la media del score de silueta obtenido para cada número de clusters. Finalmente, el número de clusters que obtuvo el mejor score de silueta se utlizó para generar los clusters en la gráfica de tSNE (izq) y para construir los gráficos de silueta (derecha) correspondientes. Dada la ineficaz separación de los puntos, se forzó a que el número de clusters fuera mayor a 3 para obtener colecciones más pequeñas.

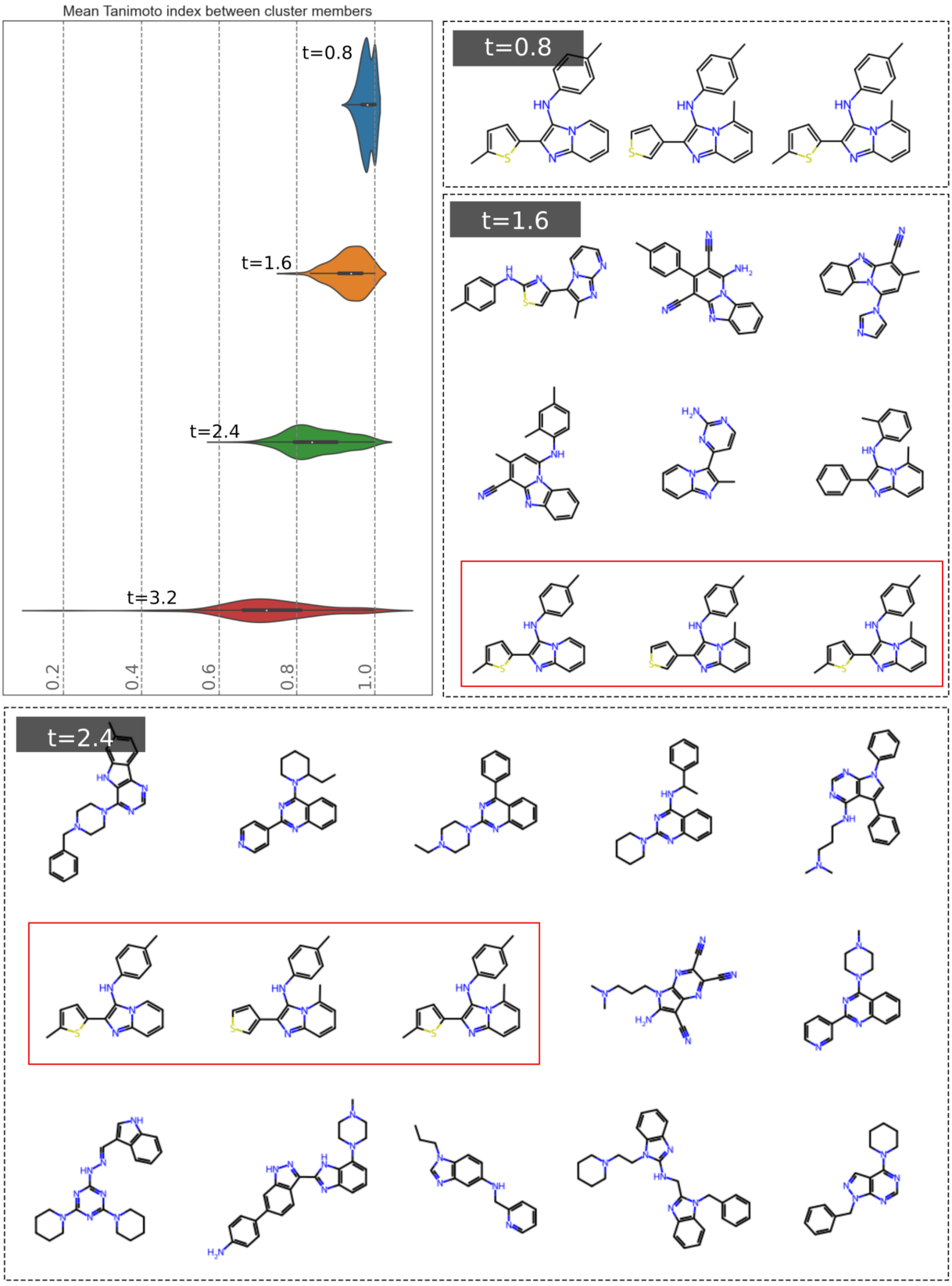


Figura 4.3 – Clustering Jerárquico para obtención de microclusters. Evolución de la homgeneidad de los clusters obtenidos con distintos umbrales de corte para un dendrograma construído a partir de una matriz de distancias usando 1 – $\text{Tanimoto}_{\text{index}}$.Arriba a la izquierda: gráfica de violines con la media de similitud intra-cluster (eje X) para cada cluster de más de 1 integrante y para distintos valores de corte (t). En las cajas de linea punteada: Evolución de la afiliación de 3 moléculas similares al aumentar los valores de corte (0.8, 1.6 y 2.4) del dendrograma. Las tres moléculas en el cluster seleccionado (arriba a la derecha) se marcan en color en los otros clusters.

Conformación de las bibliotecas de screening

Finalmente, sobre el dataset limpio y caracterizado, se realizó una búsqueda por subestructura para un conjunto de grupos funcionales reportados en literatura con posible capacidad tripanocida. La lista de *scaffolds* completa puede visualizarse en la Tabla 4.1. Con esta información particionamos nuestro dataset en bibliotecas enfocadas: cada biblioteca estaba caracterizada por contener compuestos con alguno de estos grupos funcionales. Las bibliotecas no son exclusivas: una molécula puede pertenecer a una o más bibliotecas. Aunque podría esperarse un correlato entre el la pertenencia a una u otra biblioteca y el mecanismo de acción, cabe destacar que, en este punto, es imposible saber si estos *scaffolds* serán o no responsables de la actividad tripanocida observada. Este sub-agrupamiento tuvo como fin facilitar la exploración final de las bibliotecas.

Scaffold	Tamaño biblioteca	*Autor* (Año)	Ref
Benzamidina	37	Vanden Eynde *et al* (2016)	[212]
Sulfonamida	40	Chohan*et al* (2013)	[213]
Indol	7	Menzencev *et al* (2007)	[214]
Piperazina	31	Ciammaichella *et al* (2020)	[215]
Azol	91	Goad *et al* (1989)	[216]
Tiazol	49	De Oliveira filho *et al* (2021)	[217]
Oxazol	19	Rocha *et al* (2022)	[218]
Furano	1	Zuma *et al* (2017)	[219]
Cromeno	9	BatistaJr *et al* (2008)	[220]
Benzotiazol	16	Racane *et al* (2021)	[221]
Tiazoleidina	1	Moreira *et al* (2014)	[222]
Nitro	16	NA*	

Tabla 4.1 – Tabla de bibliotecas de compuestos agrupados por *scaffold*. Los scaffolds aquí presentados se han reportado en compuestos con actividad tripanocida. Indistintamente de si la actividad señalada fuera específicamente atribuible al *scaffold*, este ordenamiento facilita la exploración visual y la validación de los microclusters generados durante la etapa analítica. Se provee una sola referencias bibliográficas representativa en cada caso donde estos *scaffolds* mostraron inhibición de crecimiento de los parásitos, excepto para el grupo funcional Nitro que es omnipresente en Chagas (presente en BNZ y NFX).

Adquisición de los compuestos

Se adquirieron un total de 21 compuestos, 12 de la biblioteca de nitro y 9 de la biblioteca de piperazinas. A pesar de compartir *scaffolds* en común y de estar agrupados de en pequeñas vecindades, según el análisis de tSNE, los grupos funcionales adicionales en cada molécula forman una biblioteca muy diversa, como se puede ver en la representación de la red de la biblioteca. La visualización de las subredes descritas para cada uno de los compuestos (verde) con blancos foráneos (es decir, organismos que no son *T. cruzi*, gris) y putativos para el parásito (naranja) se puede ver en la figura 4.4. Los nodos de moléculas verdes rellenos fueron los adquiridos para la validación experimental, aquellos denotados con una estrella se desempeñaron satisfactoriamente en el *screening* primario.

Validación experimental: screening primario

La actividad tripanocida de estos compuestos se determinó *in vitro* contra amastigotes de *T. cruzi* creciendo en células Vero, utilizando parásitos transgénicos de la cepa Tulahuen modificados para

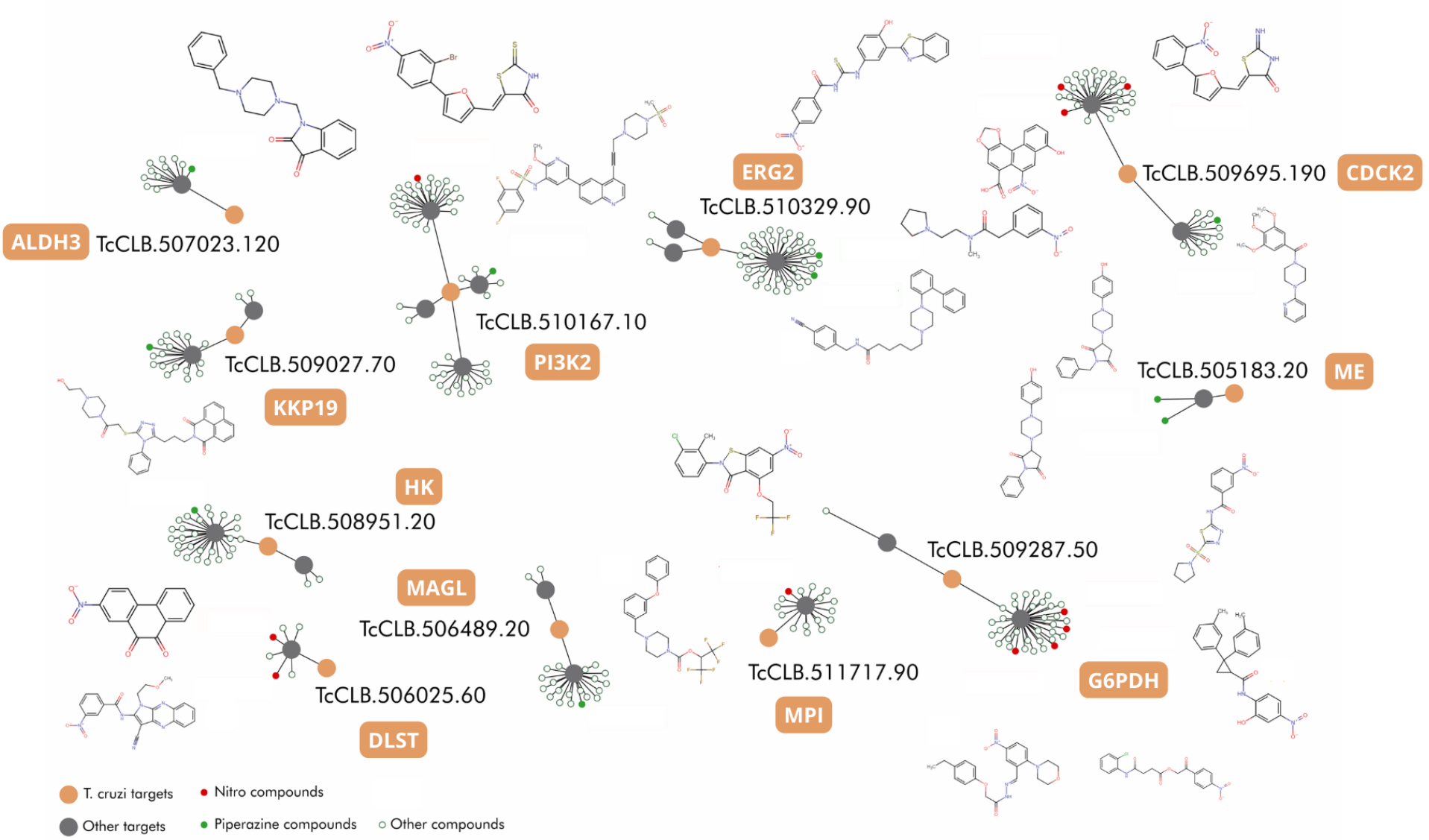


Figura 4.4 – Sub-grafos para cada par de entidades droga-*target* adquiridos para validación experimental. Se presentan de forma esquemática los sub-grafos obtenidos cada una de las drogas en el dataset. Se marcan de forma distintiva las moléculas adquiridas para validación experimental, con los compuestos de la biblioteca nitro (rojo) y piperazina (verde), con su correspondiente identificador numérico en la plataforma TDR Targets. Además, se provee la estructura 2D de éstas para mostrar la diversidad estructural explorada. Cada nodo gris representa uno o más blancos foráneos relacionados a un blanco de *T. cruzi* (naranja) mediante una o más afiliaciones funcionales. Se indican además los blancos putativos asociados en cada caso.

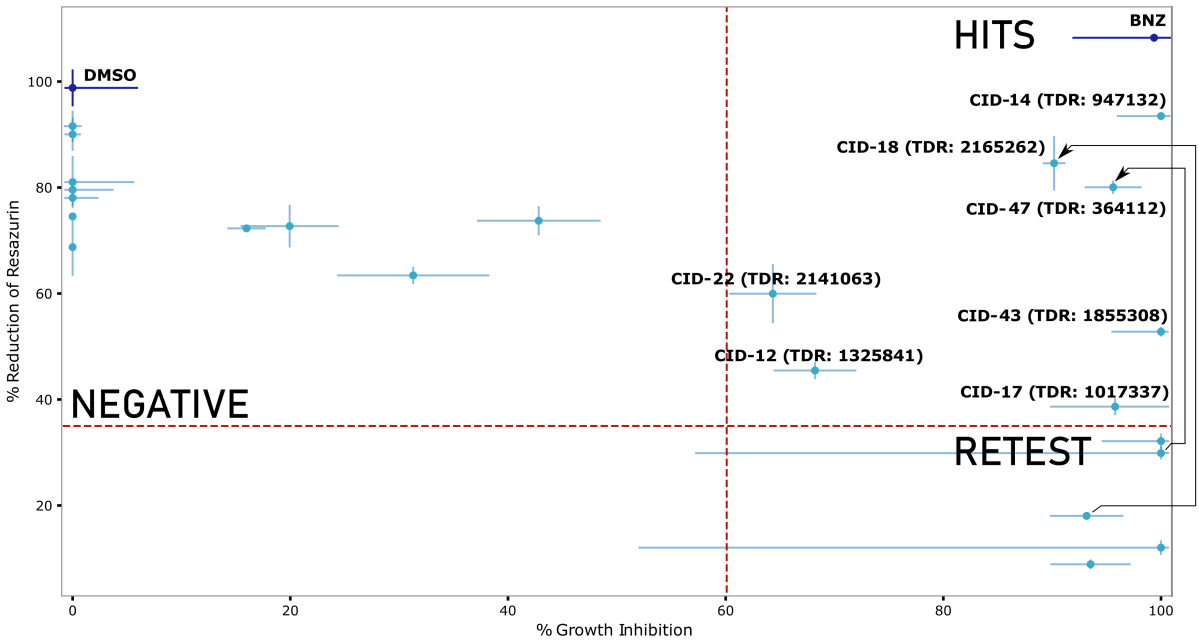


Figura 4.5 – Gráfica de actividad tripanocida vs citotoxicidad. Se muestra el resultado de los ensayos de medición de actividad β-galactosidasa y reducción de resazurina para cada compuesto. Los compuestos se ensayaron a concentración 20μM. En el eje horizontal se expresa el % de inhibición del crecimiento de amastigotes según medición de actividad β-gal contra sustrato CPRG. En el eje vertical se expresa la viabilidad celular en función del porcentaje de reducción de resazurina para un tiempo fijo: cuanto mayor es el porcentaje, mayor es el metabolismo celular y por lo tanto menor es la citotoxicidad del compuesto. Las lineas horizontales y verticales que cruzan cada punto representan el error de medición, expresado como la desviación estándar entre las determinaciones y el promedio. Para este ensayo, el umbral fue fijado en 35 % para el ensayo de citotoxicidad y en 60 % para inhibición del crecimiento. Las moléculas de interés se ubican en el cuadrante superior derecho (alta actividad tripanocida, baja citotoxicidad), y se encuentran etiquetados como 'CID-número', acompañados de BNZ (Benznidazole) que fue utilizado como control positivo; las del cuadrante inferior derecho también fueron seleccionadas para una repetición del ensayo a menores concentraciones. Las compuestos apuntados con flechas indican la recuperación de dos compuestos que tuvieron resultados satisfactorios al ser re-ensayados a menor concentración. Las moléculas en los cuadrantes izquierdos se consideraron negativas, acompañadas de DMSO que se utilizó como control negativo.

expresar el gen bacteriano de β-galactosidasa LacZ [223], incubando cultivos con medio RPMI suplementado con cada compuesto (20 μM) durante 96 hs, y midiendo la actividad de β-gal como indicador del crecimiento del parásito. La citotoxicidad del compuesto se evaluó en células Vero no infectadas mediante un ensayo de resazurina (RZ) con concentraciones de fármaco similares al ensayo de inhibición del crecimiento de los parásitos [224]. En este punto, los compuestos fueron re-etiquetados como 'CID-número' (CID = *compound identifier*) para referencia interna del ensayo. Los resultados del screening pueden verse en la Figura 4.5.

A partir de esta selección preliminar, 3 compuestos mostraron resultados prometedores en el screening primario a 20 μM: CID-14 (TDR-947132), CID-17 (TDR-1017337), y CID-43 (TDR-1855308); con buena actividad tripanocida y citotoxicidad baja o nula. Otros dos compuestos, CID-22 (TDR-2141063) y CID-12 (TDR-1325841), mostraron una capacidad tripanocida cercana al 60 % (por encima del umbral), pero también mostraron una citotoxicidad no despreciable. Un conjunto adicional de 5 compuestos mató tanto a las células hospedadoras como a los parásitos y, por lo tanto, debió ser analizado a concentraciones más bajas para determinar si verdaderamente presentaban actividad tripanocida. Éstos fueron re-analizados a 2 μM: de estos cinco compuestos, CID-18 (TDR-2165262) y CID-47 (TDR-364112) conservaron su actividad tripanocida bajando considerablemente la citotoxicidad; el resto mantuvo el perfil

citotóxico o perdió la capacidad tripanocida, por lo que fueron finalmente descartados. En total, 7 compuestos (CID-12, CID-14, CID-17, CID-22, CID-18, CID-43 y CID-47) resultaron positivos en el screening primario; aunque, ante la existencia de mejores candidatos, CID-12 y CID-22 fueron descartados en los experimentos ulteriores. En adelante, se decidió continuar la caracterización de la actividad tripanocida únicamente con los 5 compuestos que mostraron la mayor capacidad inhibitoria: TDR-947132 (CID-14), TDR-1017337 (CID-17), TDR-2165262 (CID-18), TDR-1855308 (CID-43) y TDR-364112 (CID-47). La lista de hits, su estructura, blancos putativos y librería de origen puede verse en la figura 4.6

Determinación de IC50s

Para la determinación de la concentración inhibitoria 50 (IC50) se realizó una curva de dosis-respuesta, variando la concentración de compuesto en pocillos conteniendo números similares de células Vero y multiplicidad de infección (MOI). Los compuestos de mayor potencia fueron el TDR-2165262 y el TDR-1855308, que mostraron un IC50 submicromolar. El TDR-947132 y el TDR364112 tuvieron IC50s del mismo orden (micromolar), mientras que el TDR-1017337 mostró la potencia más baja de todas, con un IC50 del orden de 1×10^{-5}. Las formas de las curvas también fueron distintas entre los compuestos, lo que se ve reflejado en las pendientes de Hill (*Hill Slope*) disímiles. Las curvas y los respectivos valores pueden visualizarse en la Figura 4.7.

4.1.3. Métodos

Conformación de las bibliotecas de screening

Figura 4.8 – Construcción de bibliotecas de screening.

La conformación de las bibliotecas se realizó en dos partes. En primer lugar se colectaron varios datasets de múltiples fuentes, siendo TDR Targets una de las más relevantes pero no la única. La lista completa de *datasets* puede hallarse en la tabla 4.2.

Una vez colectados, se combinaron secuencialmente los distintos *datasets* para reducir el número de entidades (fundamentalmente de compuestos) a una cifra manejable. En primer lugar, se removió de la lista completa de interacciones putativas (`NDPI`, del inglés, *network derived putative interactions*) a todos los registros cuyos compuestos se hallaran también en la lista de compuestos ensayados (`tested`). El conjunto de datos obtenido se intersecó con el de disponibilidad comercial (`available`), removiendo del dataset todos los registros cuyos compuestos no pudieran ser obtenidos comercialmente y generando un nuevo dataset. Éste se combinó con el dataset de drogas de Drugbank (`drugbank`) con el afan de obtener un listado rápido de drogas reposicionables, pero no se obtuvieron candidatos interesantes.

Los blancos putativos también se usaron para reducir el tamaño del dataset original. Para hacerlo, el dataset obtenido de las operaciones anteriores fue intersecado con una lista de genes de interés de alta drogabilidad putativa (`cherry-D`) y presuntamente involucrados en el metabolismo energético y de aminoácidos

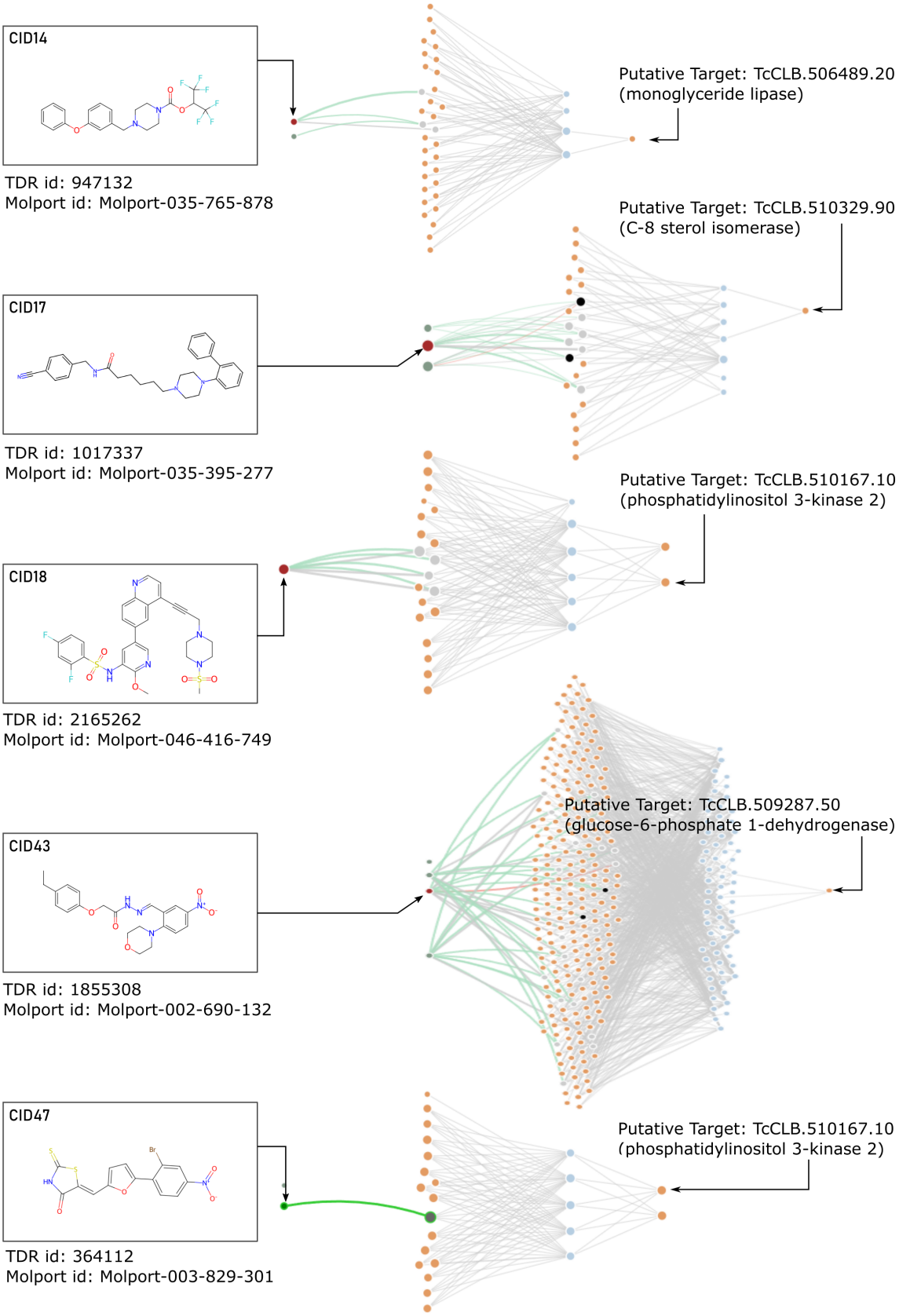
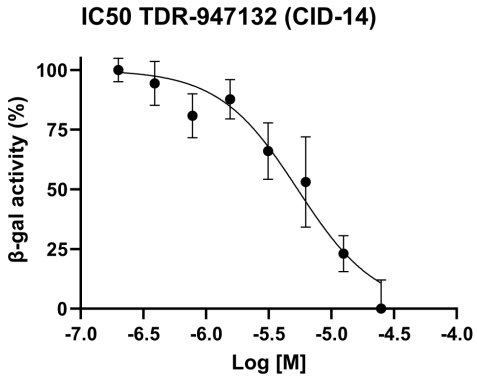
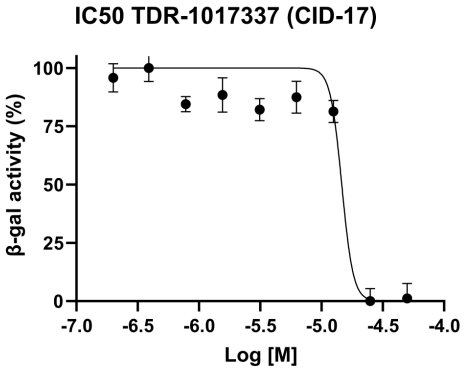
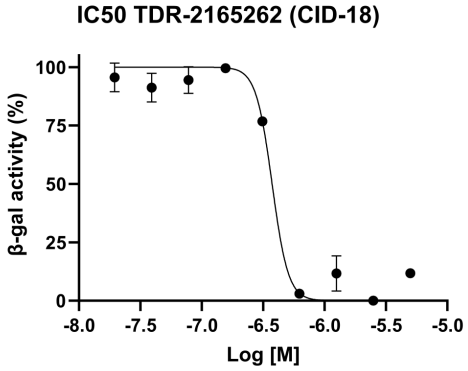
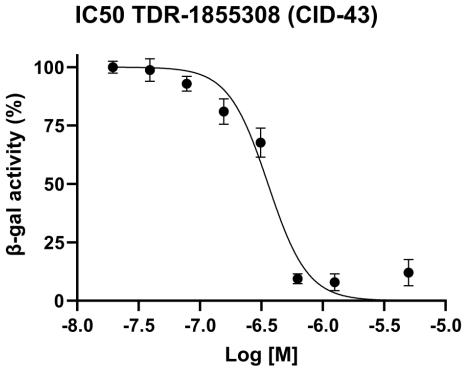


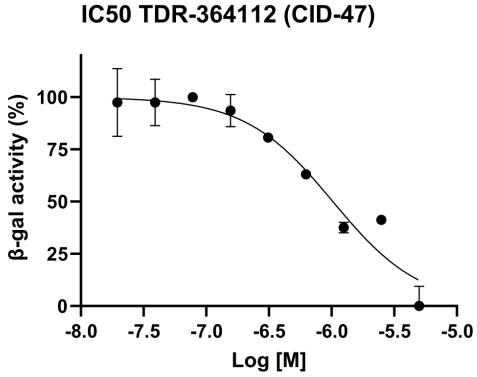
Figura 4.6 – Top-5 *Hits* del screening primario, subgrafos extraídos TDR Targets y blancos putativos. Estructura química de los *hits* y sub-grafo recogido de TDR Targets para cada compuesto. En el grafo, los nodos verdes y el nodo rojo son compuestos, los nodos naranjas, grises o negros son blancos protéicos, y los nodos celestes son anotaciones funcionales. Los nodos naranjas a la derecha de cada grafo son los blancos de putativos de *T. cruzi*, que fueron manualmente aislados para facilitar la construcción y comprensión de la figura. *Se omitieron esta figura los compuestos CID-12 y CID-22.*











	LogIC50	Hill Slope	IC50
CID14	-5.269	-1.380	$5.39e^{-6}$
CID17	-4.835	-9.308	$1.46e^{-5}$
CID18	-6.427	-6.609	$3.74e^{-7}$
CID43	-6.446	-2.690	$3.58e^{-7}$
CID47	-5.990	-1.240	$1.02e^{-6}$

Figura 4.7 – Curvas IC50. Los primeros *hits* obtenidos en el *screening* primario se re-ensayaron en distintas concentraciones para la obtención de sus curvas de IC50. En el panel inferior derecho se resumen los datos para cada compuesto.

(`cherry-P`). La figura 4.8 esquematiza el orden en el que se combinaron los *datasets* y la cantidad de entidades obtenidas en cada caso.

Todas las operaciones tabulares (concatenaciones, combinaciones, intersecciones entre datasets, etc) se realizaron usando Pandas (`pandas==2.0.2`). El código necesario para replicar la construcción de esta librería puede hallarse en Github en un notebook Python (libraryPreparation.ipynb).

Caracterización y armado de biblioteca de compuestos

Para caracterizar la biblioteca se utilizaron distintos algoritmos de clustering sobre descriptores moleculares y métricas de distancia (estructural) entre compuestos. La extracción de descriptores, medidas de similitud y renderizaciones de moléculas se realizaron con Rdkit (`rdkit-pypi>=2022.3.4`). Para el clustering y análisis estadístico se uso Scikit-Learn y SciPy (`scikit-learn==1.2.2`, `scipy==1.10.1`).

Validación experimental

Para determinar la actividad tripanocida de los compuestos, se realizó un ensayo colorimétrico utilizando parásitos transgénicos *T. cruzi* Tulahuen que expresan β-galactosidasa

Origen	Dataset	Label	Tamaño aprox	Tipo
TDR Targets	`NDPI`	Compuestos obtenidos por transformación de lista de blancos en lista de compuestos	180.023	(C)
TDR Targets	`tested`	Compuestos ensayados (activos, inactivos, inconcluyentes) contra tripanosomátidos	6.619	(C)
DrugBank	`drugbank`	Fármacos aprobados por la FDA, en circulación y retirados	~5.000	(C)
Molport	`available`	Disponibilidad comercial	+5.000.000	(C)
Varios*	`NDPI`	Curación manual de *screenings* de alto rendimiento contra *T. cruzi* todavía **no** incorporadas a TDR Targets	36	(C)
TDR Targets	`cherry-D`	Proteínas de *T. cruzi* en grupo de drogabilidad ≥ 4	327	(P)
TDR Targets	`cherry-P`	Proteínas de *T. cruzi* involucradas en metabolismo energético y de aminoácidos	44	(P)

Tabla 4.2 – Lista de *datasets* obtenidos para su combinación posterior en la generación de las bibliotecas. Se muestra cada *dataset* obtenido en forma independiente con su respectivo **origen** y **tamaño**. El **tipo** de cada uno representa la entidad que lo compone, sea éste un *dataset* de compuestos (C) o de proteínas (P). En el caso del dataset compuesto de Varios orígenes (*), se trata de un dataset creado manualmente con los resultados de dos trabajos que describen screenings fenótipicos *High-througput* para *T. cruzi* [225, 226]. NDPI = network derived putative interactions.

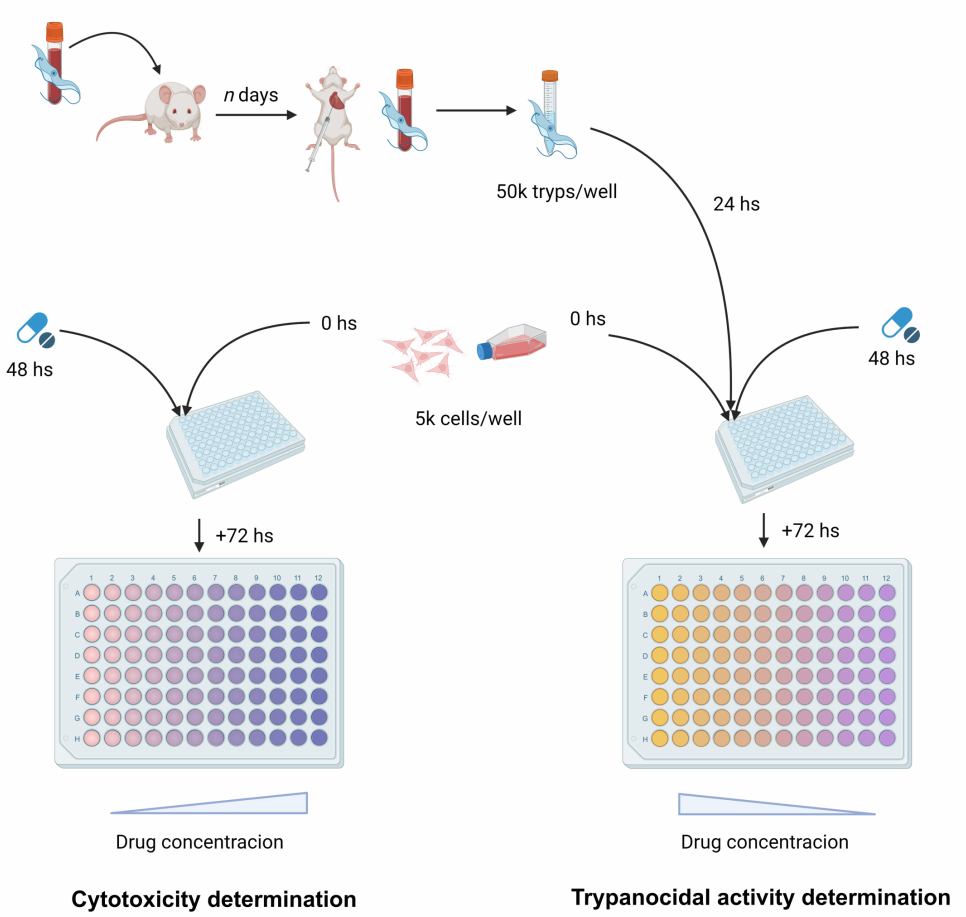


Figura 4.9 – Esquema experimental de los ensayos de actividad tripanocida y citotoxicidad *in vitro*. El ensayo de actividad tripanocida comienza con la obtención de tripomastigotes de sangre de ratón, conteo e infección de placas previamente sembradas con células Vero. Pasadas 24 hs de la infección, se lavan los tripomastigotes no infectivos y se agrega una solución de medio fresco con el compuesto de interés. Finalmente, se incuba la placa durante 72 hs para luego realizar la determinación bioquímica de actividad β-gal. En el caso de la determinación de citotoxicidad, el ensayo se hace sobre placas previamente sembradas con células Vero, a las que pasadas 48 hs desde la siembra se les adiciona la solución de medio fresco con el compuesto de interés. Luego, la placa se incuba durante 72 hs y finalmente se le agrega la resazurina. Al cabo de 6-8 hs se puede determinar la actividad metabólica (*proxy* de viabilidad celular) indirectamente midiendo la cantidad de resazurina reducida en comparación con el control de células sin tratar.

bacteriana (LacZ), utilizando como proxy la actividad enzimática (degradación del rojo de clorofenol-β-D-galactopiranósido, CPRG). para el crecimiento de parásitos. Para ello, se utilizaron $5{,}00 \times 10^4$ tripomastigotes purificados/pocillo para infectar células Vero previamente sembradas (24 hs antes), en una placa de 96 pocillos a razón de $5{,}00 \times 10^3$ células/pocillo. Después de 24 horas de infección, los cultivos se lavaron gentilmente con PBS para eliminar los tripomastigotes libres y se alimentaron con medio RPMI sin rojo fenol (Gibco #11835030) suplementado con compuesto (20 µM). Se utilizó benznidazol (BNZ), también a 20 µM, como control positivo, junto con DMSO (0,5 %) para el control negativo/carrier, y controles infectados/no tratados y no infectados/no tratados para las mediciones esperadas de actividad máxima y mínima de β-gal del ensayo. Tanto los compuestos como los controles se probaron por duplicado. Para algunos compuestos, este ensayo se repitió en idénticas condiciones con excepción de la concentración de compuesto, que fue de 2 µM.

Después de la incubación del cultivo con compuesto 20 µM durante 96 h, se agregó a cada pocillo

100 µL de una solución preparada en el momento de NP40 al 1 % y CPRG 100 µM (Roche n.º 10884308001), hasta una concentración final (en pocillo) de 0,5 % y 50 µM, respectivamente. Como parte del ensayo se utilzó BNZ (20 µM) como control positivo, DMSO (0,5 %) como control de vehículo y PBS como control negativo. A continuación, las placas se incubaron durante 4 horas en la estufa y se mantuvieron al reparo de la luz durante toda la incubación. Finalmente, la actividad de β-gal se determinó mediante la lectura de absorbancia a 595 nm en un lector de placas (FilterMax F5 Multimode Microplate Reader, Molecular Devices). Todos los compuestos se ensayaron por duplicado. La figura 4.9 esquematiza los pasos seguidos para la determinación de actividad tripanocida (derecha).

Para determinar la citotoxicidad, se realizó un ensayo de resazurina (RZ) utilizando las mismas concentraciones de compuesto en cultivos de células Vero no infectadas. Las células se sembraron en placa de 96 pocillos a razón de $5{,}00 \times 10^3$ células/pocillo y se cultivaron durante 48 hs. Luego se lavaron cuidadosamente con PBS, se alimentaron con medio fresco RPMI suplementado con cada uno de los compuestos (20 µM) y se incubaron durante 96 hs. Posteriormente, se agregaron a cada pocillo 10 µL de solución RZ 10X recién preparada. También se agregaron los controles de BNZ (20 µM) y DMSO (0,5 %) a este ensayo, junto con un control de células no tratadas y un control sin células para lecturas de reducción de reactivo máxima y mínima. Además, se añadió a la placa un control de RZ reducido al 100 % (solución de RZ esterilizada en autoclave 1X en medio RPMI) como un medio para determinar cuándo tomar la lectura de reducción final: se tomaron lecturas de absorbancia a 600 nm cada hora hasta que el control de células no tratadas alcanzó el ~100 % de lecturas de control RZ reducidas. La lectura final se tomó después de ~7 horas de incubación, donde se registró la absorbancia a 600 nm y 570 nm. Todas las mediciones se realizaron utilizando el mismo lector de placas (FilterMax F5 Multimode Microplate Reader, Molecular Devices). Los compuestos y controles se probaron por duplicado. La figura 4.9 esquematiza los pasos seguidos para la determinación de citotoxicidad en células Vero (izquierda).

Los parásitos transgénicos Tul-β-gal se obtuvieron a partir sangre anticoagulada de ratones infectados. Para su purificación, el mismo día del sangrado de los ratones, la muestra se diluyó en 1 o 2 volúmenes de PBS y se centrifugó a 300 g (1000 rpm) por 5 minutos; luego se dejó en estufa a 37ºC durante 40 minutos para favorecer el nadado de los tripomastigotes. El centrifugado y nadado se repitió una segunda vez. Luego de 40 min, se tomó nuevamente el sobrenadante y se centrifugó a 7000 rpm (~4500g) 7-10 minutos, esta vez para bajar los tripomastigotes obtenidos. El *pellet* de parásitos obtenido se lavó repetidas veces con medio RPMI, volviendo a centrifugar en cada lavado. Antes del último lavado, se sacó alícuota para contar y chequear viabilidad. Los parásitos obtenidos pueden usarse directamente para ensayos de actividad tripanocida, pero para los experimentos desarrollados en este capítulo se realizó al menos 1 pasaje por botella de cultivo con células Vero al 30-50 % de confluencia. Esto reduce considerablemente la cantidad de plaquetas y favorece a un ensayo más limpio, en general. Para purificar parásitos de cultivo, el protocolo es el mismo, aunque pueden acortarse considerablemente los tiempos de centrifugado y nadado de tripomastigotes.

Selección de hits

Los compuestos con lecturas bajas de CPRG a 595 nm y alto % de reducción de Resazurina se consideraron *hits* directos. Aquellos con lecturas bajas de CPRG y un % de reducción de RZ bajo fueron re-analizados en un ensayo idéntico al presentado más arriba. Para el ensayo CPRG, el umbral se determinó como la lectura CPRG para el control no infectado más 3 veces la desviación estándar para dicha medición (0,2084, ~ 60 % de inhibición del crecimiento). Para citotoxicidad,

el umbral se fijó arbitrariamente en 35 %. En el caso del ensayo de reducción de Resazurina, la actividad se expresa directamente en %; para las lecturas de CPRG, las lecturas de abosbancia a 595 nm se normalizaron utilizando una interpolación lineal entre el valor mínimo (según lectura media para control sin infectar) y el valor máximo (según lectura media para control sin tratamiento). Luego, los valores obtenidos para cada compuesto, ahora normalizados entre 0-1, se expresaron como porcentuales.

Curvas dosis-respuesta (IC50)

Los experimentos para determinar las curvas de dosis-respuesta y medir el IC50 se realizaron en placa de 96 pocillos (tal cual fue descrita en el apartado de validación experimental) para distintas concentraciones de cada uno de los compuestos, partiendo de 50 μM y realizando diluciones al medio hasta alcanzar una concentración aproximada de 200 nM. Todas las condiciones y mediciones se hicieron por duplicado. A las lecturas de absorbancia 595 nm se les restó el blanco (control sin infección). Los datos fueron importados en Graph Prism v9.5.1 (Graphpad Software LLC). Allí, los datos fueron normalizados utilizando la función integrada de *Normalización por subcolumna*, tomando el valor más bajo de cada dataset como 0 % y el más alto como el 100 %. Luego, los valores normalizados se utilizaron para generar una curva de *log(dosis)-respuesta utilizando un ajuste no-lineal con la pendiente como grado de libertad*. Los datos de cada compuesto se analizaron por separado.

4.1.4. Discusión

Sobre el uso de TDR Targets

La base de datos TDR Targets es un recurso valioso que puede utilizarse de maneras diversas. Cuando fue conceptualizada, y en las sucesivas actualizaciones, la intención fue crear un recurso flexible y autónomo, proveyendo herramientas para manipular datos y generar listas priorizadas de entidades. Si bien esto está al menos parcialmente cumplido, haber usado la herramienta con un objetivo concreto de reposicionamiento puso en evidencia la necesidad de un conjunto de *features* indispensables para que la herramienta sea verdaderamente útil como instrumento de democratización. Por ejemplo, la disponibilidad comercial de los compuestos hizo que solo el 15 % de las moléculas inicialmente propuestas fueran realmente adquiribles (y por lo tanto ensayables). Tener la capacidad de filtrar por disponibilidad comercial dentro de la base de datos como parte del pipeline de preparación de las bibliotecas sería ideal. Otra funcionalidad que sería deseable tener es la de herramientas integradas de *clustering* de compuestos. Aunque el agrupamiento de compuestos en base a similitud química está presente en forma subyacente en la capa de drogas de la red de targets-compuestos, incorporar estrategias simples de agrupamiento de compuestos, sería una gran ayuda para asistir a los usuarios en el re-análisis y construcción de bibliotecas para luego adquirir o producir los compuestos.

Finalizada la instancia de generación de bibliotecas, TDR Targets vuelve a ser útil a la hora de explorar *hits*. Dado que el origen de los compuestos es la misma base, se puede volver al recurso para ver en detalle la página de cada compuesto, los compuestos similares a éste y sus blancos putativos; también se puede acceder fácilmente a cada blanco y decidir cuál vale la pena explorar. Sin embargo, esto solo es posible cuando los compuestos están registrados en la base de datos. Para moléculas nuevas, no presentes en TDR Targets, pueden obtenerse y explorar entidades similares (mediante búsquedas de similitud por índice de Tanimoto, subestructura) lo que permite analizar

un compuesto de interés de forma indirecta. Un *feature* interesante, en este sentido, sería permitir al usuario obtener una recopilación de información relevante sobre compuestos similares a uno de interés dentro de TDR Targets, integrándolo en tiempo real y determinando sus propiedades fisicoquímicas y las posibles conexiones que tendría de haber sido incluido en la red.

En nuestro caso, la falta de estas funciones dentro de la base de datos no impusieron un grave impedimento, por que contamos con *know-how* para obtener esas respuestas por fuera de la aplicación web. No obstante, son *features* que podrían resultar de interés en la comunidad.

Finalmente, debe destacarse que durante este trabajo solo se exploró la actividad tripanocida de poco más del 5 % de las moléculas priorizadas, lo que sugiere que hay un gran potencial aún oculto en los resultados de esta priorización.

Sobre los blancos putativos

De los 7 compuestos positivos obtenidos durante screening primario, dos de ellos (TDR-1855308 y TDR-1325841) se presentaron como posibles ihibidores de la glucosa-6-fosfato 1-deshidrogenasa, G6PDH (TcCLB.509287.50), TDR-1017337 como un inhibidor potencial para la C-8 esterol isomerasa (TcCLB.510329.90) y TDR-947132 se asoció como inhibidor de una monoacilglicerol lipasa, MAGL (TcCLB.506489.20) de *T. cruzi*. Tanto TDR-2165262 como TDR-364112 se obtuvieron por asociación con la fosfatidilinositol 3-kinasa 2, PI3K2 (TcCLB.510167.10). Finalmente TDR-2141063 sería un inhibidor putativo de una o ambas dihidrolipoamida deshidrogenasas, DLD (TcCLB.511025.110 / TcCLB.507089.270).

Es importante notar que ninguna de las asociaciones droga-*target* obtenidas por el modelo de redes en este trabajo es infalible, por lo que no se puede afirmar fehacientemente que el blanco de estas moléculas sea inequívocamente el señalado. En la discusión que sigue a continuación se trabajó desde la hipótesis de que los blancos propuestos son una aproximación suficiente para tomar decisiones sobre cuál de las moléculas (y sus respectivos blancos putativos) merece la pena seguir estudiando.

La G6PDH es un blanco ampliamente estudiado [227–229]. Además de participar en la vía glucolítica para la generación de energía, se cree que posee un rol defensivo/regulatorio ante el stress oxidativo causado por el ambiente, lo que ha posicionado a esta enzima – a principios de los 2000 – como un posible blanco secundario para terapias en combinación con BNZ o NFX. Al igual que su homólogo en humanos, la enzima puede modularse químicamente mediante el uso de moléculas esteroideas [230, 231] y se han logrado moléculas altamente selectivas, capaces de inhibir solo la G6PDH del parásito sin afectar la G6PDH humana [232], con potencias sub-micromolares. A pesar de estos resultados promisorios, el desarrollo de inhibidores para este blanco parece haberse estancado; quizás debido a que los esteroides son generalmente poco atractivos como fármacos para enfermedades desatendidas dado su alto costo de producción. Curiosamente, ninguno de los potenciales inhibidores de G6PDH hallados durante este screening es un esteroide, lo cual podría suponer una nueva familia de compuestos para un blanco ampliamente validado.

Aunque menos estudiada en comparación con la G6PDH, la dihdrolipoamida deshidrogenasa (DLD, EC 1.8.1.4) también ha sido aislada y caracterizada en tripanosomátidos [233]. En otros organismos en los que la enzima puede hallarse, la DLD forma complejos con otras enzimas para regenerar la dihidrolopoamida; proceso que tiene lugar en la mitocondria y cuyo producto (el ácido lipoico) es un cofactor esencial para la catálisis de multiples deshidrogrenasas. Dentro de estos complejos, la DLD es la subunidad E3 de la 2-oxoglutarato deshidrogenasa, la

deshidrogenasa de aminoacidos ramificados, la piruvato deshidrogenasa y la subunidad L de la glicina decarboxilasa (también conocida como *Glycine Cleavage System* (GCS) en el metabolismo de aminoacidos) [234, 235]. Además, el ácido dihidrolipoico tiene un rol como *scavenger* de óxido nítrico y radicales libres ante stress oxidativo [236]. De hecho, se ha observado una expresión diferencial de esta enzima en aislamientos naturalmente resistentes a BNZ [237]. Por otra parte su silenciamiento reduce significativamente el *fitness* en *T. brucei* [238, 239], y su inhibición produce la muerte del parásito en *T. cruzi* [240].

En este contexto, DLD podría condiderarse un blanco ideal para regímenes de terapias combinadas con BNZ o NFX, en donde la inhibición de DLD funcione como agente sensibilizante y permita usar dosis menores de BNZ o NFX con igual o mejor efectividad.

La C-8 esterol isomerasa (mejor conocida como ERG2, en *S. cerevisiae*) participa en la biosíntesis de ergosterol, y cataliza la reacción que resulta en la insaturación de C7 en el anillo B de estas moléculas, convirtiendo fecosterol en episterol [241]. Dada la relevancia del ergosterol en las membranas de tripanosomátidos y la evidencia de que su interferencia conduce a la muerte celular *in vitro* e *in vivo*, esta ruta metabólica ha sido abordada por muchas campañas de descubrimiento y reposicionamiento de fármacos [242–244]. Sin embargo, el objetivo es la ruta mebólica en general. Cuando se trata de inhibidores específicos, CYP51 (Lanosterol 14-α-demethylasa) es el blanco más comúnmente abordado [245–247]. A pesar de la multiplicidad de inhibidores de CYP51 surgidos en la última década, este target perdió relevancia luego de los ensayos clínicos fallidos para Posaconazol [200] y E1224 [248]. Dado que es esperable que la inhibición de una enzima río abajo en la vía de biosíntesis de ergosterol tenga un efecto similar al de la inhibición de CYP51, la molécula no parece revestir un interés adicional para seguir siendo estudiada.

La PI3K2 es una integrante de una superfamilia de kinasas involucradas en la transducción de señales que dirigen procesos celulares complejos diversos, como crecimiento celular, proliferación, diferenciación, motilidad y tráfico, entre otras. En tripanosomátidos, esta y otras kinasas de la misma superfamilia han sido evaluadas como posibles blancos terapéuticos [249]. En screenings con inhibidores conocidos de PI3Ks humanas, algunos compuestos como el NVP-BEZ235 (Dactolisib – una droga oncológica que inhibe PI3K y mTOR en mamíferos y que inhibe el crecimiento celular de células cancerosas) mostraron potencia sub-nanomolar [250], y su efectividad fue demostrada tanto *in vitro* como *in vito* para *L. donovani* y *T. brucei* [251]. Este antecedente supone un terreno fértil para el reposicionamiento de inhibidores de PI3K y mTOR humanas. No obstante, todos los ensayos clínicos asociados a esta droga han sido cancelados o detenidos, con el más exitoso de ellos habiendo alcanzado Fase 2 [252]. En contraste, las moléculas halladas en este trabajo y potencialmente asociadas a esta enzima no están relacionadas estructuralmente entre sí o con el Dactolisib u otras moléculas ya estudiadas anteriormente en tripanosomátidos, ofreciendo así dos posibles nuevas familias de inhibidores para esta enzima.

Finalmente, la monoacilglicerato lipasa (MAGL, EC 3.1.1.23) es la enzima que cataliza la degradación de monoglicéridos (lípidos intermediarios derivados de la degradación de fosfolípidos) en glicerol y ácidos grasos; con el ácido araquidónico como principal especie química, por su abundancia pero también por su relevancia en distintas cascadas de señalización que regulan la función sináptica y la inflamación en mamíferos [253]. En tripanosomátidos, su función no es tan clara, pero existe evidencia de que podría funcionar como factor de virulencia, posiblemente modulando la inmunidad innata del hospedador [254, 255]. A la fecha, no existen reportes de inhibición de MAGL en tripanosomátidos; el TDR-947132 sería el primer compuesto reportado capaz de inhibir esta enzima en *T. cruzi*. Todas las enzimas involucradas en el metabolismo de eicosanoides están vastamente estudiadas en mamíferos, y a la fecha existen

múltiples inhibidores para la vía completa y para MAGL en particular, lo cual supone un escenario rico tanto para el reposicionamiento directo de drogas, como para la caracterización biológica de la enzima y su estudio como potencial blanco terapéutico.

Perspectivas futuras

Como se mencionó al inicio de estas conclusiones, la red integrada en TDR Targets [2, 152] es la que sugiere cuál es el blanco de cada compuesto. Esto deberá ser validado experimentalmente como parte de los ensayos que se diseñen. Hay una serie de estrategias posibles, demostradas y recomendadas [256], como las que se exponen en los lineamientos GOT-IT (*Guidelines on Target Assessment for Innovative Therapeutics)*) [257], entre otras.

Demostrar la correlación de actividad de los compuestos *in vitro* contra el parásito y contra un blanco definido (por ejemplo la proteína recombinante correspondiente en un ensayo enzimático) es una de las estrategias a seguir. Si la proteína recombinante es activa en el ensayo pero no hay correlación de actividad del compuesto (inhibición a concentraciones similares) esto sería una señal de que la proteína recombinante no sería el blanco molecular de esta droga [258].

Un análisis similar con mutantes específicas de residuos clave del sitio activo (o de unión de un sustrato o cofactor) de cada uno de estos blancos permitiría también sugerir fuertemente que la proteína expresada es el blanco molecular del compuesto, así como demostrar que el mecanismo de inhibición está relacionado con el mecanismo catalítico de la enzima [258].

Otras alternativas para validar los blancos son la sobre-expresión del gen que codifica un blanco en parásitos transgénicos; y demostrar la aparición de resistencia (IC50 incrementado) debido a la necesidad de utilizar mayores concentraciones de compuesto para lograr el mismo efecto. Al igual que en el caso anterior, una falta de corrimiento del IC50 en parásitos sobre-expresantes sería un indicador de que la proteína sobre-expresada no sería el blanco molecular del compuesto.

Estos pasos futuros ya están siendo explorados en el laboratorio como parte de nuevos proyectos de Tesis de Licenciatura y Doctorado.

5. Conclusiones Finales

El desarrollo de nuevas terapias para enfermedades relegadas o desatendidas es imperiosamente necesario. Muchas de las drogas actualmente indicadas para estas aflicciones presentan variedad de efectos adversos que dificultan la adhesión de los pacientes a los tratamientos. El desarrollo de mecanismos de resistencia para estas drogas es también un problema que pone en relevancia no solo la obtención de nuevas terapias, sino la agilización del proceso mediante el cual éstas son llevadas a la clínica. El reposicionamiento de drogas es una estrategia ampliamente utilizada en la industria farmacéutica, con gran potencial en el campo de las enfermedades desatendidas, por que permite acelerar el desarrollo de nuevas terapias.

La práctica de reposicionamiento alcanza no solo a los fármacos sino también a los blancos terapéuticos. Los primeros ofrecen un punto de entrada más avanzado en el *pipeline* de desarrollo de drogas, pero los segundos brindan un mayor número de candidatos y son apropiados cuando se dispone de poca o ninguna información adicional al genoma del organismo de interés. Cualquiera sea el caso, los flujos de reposicionamiento precisan de un uso integrativo de datos de distintas disciplinas, en especial de la bio- y la quimioinformática, que permiten estandarizar/homogeneizar datos provenientes de distintas fuentes y trazar relaciones entre ellos para la creación de flujos de priorización o el entrenamiento modelos de propensión.

En este trabajo se utilizaron distintas técnicas de quimioinformática para asistir en distintos puntos del proceso de *drug discovery*. La potencia de armonizar datos e integrarlos se demuestra no solo en las prácticas exploradas en el **capítulo 3**, cuando mostramos el racional detrás de la construcción de la última versión de TDR Targets (v6.1). Aquí se combinan la bio y la quimioinformática, junto a la quimiogenómica, para dar luz a formas inteligibles de explorar datos complejos, realizar priorizaciones de entidades de interés de forma flexible y verificar predicciones de actividad o drogabilidad para un compuesto o blanco proteico dado. Esto es posible para un total de 37 especies de patógenos que causan enfermedades relegadas, 21 de ellas relevantes para la salud humana. También en este campo encontramos limitaciones y desafíos complejos, como la dificultad de ofrecer visibilidad sobre la disponibilidad comercial o la escalabilidad de los servicios.

Parte de la potencia de TDR Targets (y algunas de sus limitaciones) se vieron en el **capitulo 4**. Aquí, el repositorio de datos fue muy valioso para la generación primigenia de los datasets filtrados por características de interés (como drogabilidad o novedad), pero debieron usarse recursos externos para obtener en forma programática la disponibilidad comercial o para obtener alguna intuición acerca de la diversidad estructural explorada en el conjunto de moléculas finalmente obtenido. Este tipo de *features* serían de suma utilidad en TDR Targets. Amén de estas dificultades, se obtuvo un número reducido de moléculas cuya actividad tripanocida pudo evaluarse *in vitro*. De las cerca de 20 moléculas ensayadas, 5 de ellas resultaron activas con un buen perfil de citotoxicidad en los *screenings* primarios. Dado que las 20 moléculas fueron seleccionadas al azar luego de corroborar que no habían sido ya ensayadas en tripanosomátidos, podría pensarse que el *ratio* "5/20" (moleculas activas / moléculas totales) es ampliamente superador en contraste con una búsqueda de *leads* a ciegas. No obstante, debe tenerse en cuenta que el cálculo no compone una tasa de éxito convencional que pueda hablar del éxito o fracaso del flujo de priorización, dado que no se han probado experimentalmente todas las moléculas priorizadas, sino que se han elegido al azar solamente algunas moléculas de algunas de las librerías generadas a partir del proceso planteado.

Una posible ventaja adicional de este tipo de enfoques es que la priorización de moléculas viene acompañada de de su blanco putativo. Aunque no hay garantías de que el blanco recuperado sea exactamente el que señala TDR Targets, su asociación es altamente probable. Esto provee un punto de partida para moléculas con perfil tripanocida interesante, y también un dato adicional para re-priorizar en caso de que sea necesario seguir achicando el número de moléculas: si el *target*, por el motivo que sea, no es interesante, no tiene sentido proceder con la validación del mismo ni mucho menos con la optimización de la molécula en sí. En este trabajo se obtuvieron inhibidores putativos para G6PDH, DLD, ERG2, MAGL y PI3K2. De éstos, la enzima MAGL es posiblemente el blanco más interesante surgido en este estudio. A la vez, ortólogos de ésta están vastamente estudiados en organismos modelo, lo que supone una amplia batería de moléculas disponibles para reposicionar y ensayos bioquímicos para adaptar.

A lo largo de este trabajo se ha demostrado el potencial de usar quimiogenómica integrativa para la búsqueda y reposicionamiento de fármacos en enfermedades desatendidas, enfocándonos especialmente en la Enfermedad de Chagas. Los resultados obtenidos en este trabajo serán la base para nuevas lineas de investigación centradas en la validación de los blancos putativos, la expansión de series químicas sobre la base de los *leads* hallados, y la elucidación de los posibles mecanismos de acción involucrados en la actividad tripanocida de éstos.

Bibliografía

[1] Urán Landaburu, L., Didier-Garnham, M. & Aguero, F. Targeting trypanosomes: how chemogenomics and artificial intelligence can guide drug discovery. *Biochem Soc Trans* **51**, 195–206 (2022). URL https://portlandpress.com/biochemsoctrans/article-abstract/doi/10.1042/BST20220618/232416/Targeting-trypanosomes-how-chemogenomics-and.

[2] Urán Landaburu, L. *et al.* TDR Targets 6: driving drug discovery for human pathogens through intensive chemogenomic data integration. *Nucleic Acids Research* gkz999 (2019). URL https://academic.oup.com/nar/advance-article/doi/10.1093/nar/gkz999/5611677.

[3] Salas-Sarduy, E. *et al.* Novel scaffolds for inhibition of cruzipain identified from high-throughput screening of anti-kinetoplastid chemical boxes. *Sci. Rep.* **7**, 12073 (2017). URL http://www.nature.com/articles/s41598-017-12170-4.

[4] Salas-Sarduy, E. *et al.* Potent and selective inhibitors for M32 metallocarboxypeptidases identified from high-throughput screening of anti-kinetoplastid chemical boxes. *PLOS Neglected Tropical Diseases* **13**, e0007560 (2019). URL https://dx.plos.org/10.1371/journal.pntd.0007560.

[5] World Health Organization. *World health statistics 2018: monitoring health for the SDGs* (WHO, 2018). URL http://apps.who.int/iris/bitstream/handle/10665/272596/9789241565585-eng.pdf?ua=1. OCLC: 1040265127.

[6] David Molyneux. Neglected tropical diseases. *Community Eye Health* **26**, 21–24 (2013).

[7] Paul G Wyatt, Ian H Gilbert, Kevin D Read & Alan H Fairlamb. Target validation: linking target and chemical properties to desired product profile. *Curr Top Med Chem* **11**, 1275–1283 (2011).

[8] Lidani, K. C. F. *et al.* Chagas disease: From discovery to a worldwide health problem. *Frontiers in Public Health* **7** (2019). URL https://www.frontiersin.org/article/10.3389/fpubh.2019.00166.

[9] Stephanie A Robertson & Adam R Renslo. Drug discovery for neglected tropical diseases at the Sandler Center. *Future Med Chem* **3**, 1279–1288 (2011). URL http://dx.doi.org/10.4155/fmc.11.85.

[10] Patrice Trouiller *et al.* Drug development for neglected diseases: a deficient market and a public-health policy failure. *Lancet* **359**, 2188–2194 (2002). URL http://dx.doi.org/10.1016/S0140-6736(02)09096-7.

[11] Peter Schmidtke & Xavier Barril. Understanding and predicting druggability. A high-throughput method for detection of drug binding sites. *J Med Chem* **53**, 5858–5867 (2010). URL http://dx.doi.org/10.1021/jm100574m.

[12] Spradlin, J. N., Zhang, E. & Nomura, D. K. Reimagining druggability using chemoproteomic platforms. *American Chemical Society* **54** (2021). URL https://pubs.acs.org/doi/abs/10.1021/acs.accounts.1c00065.

[13] Albert Pujol, Roberto Mosca, Judith Farrés & Patrick Aloy. Unveiling the role of network and systems biology in drug discovery. *Trends Pharmacol Sci* **31**, 115–123 (2010). URL http://dx.doi.org/10.1016/j.tips.2009.11.006.

[14] David C Swinney & Jason Anthony. How were new medicines discovered? *Nat Rev Drug Discov* **10**, 507–519 (2011). URL http://dx.doi.org/10.1038/nrd3480.

[15] Xiaofeng Liu *et al.* PharmMapper server: a web server for potential drug target identification using pharmacophore mapping approach. *Nucleic Acids Res* **38**, W609–W614 (2010). URL http://dx.doi.org/10.1093/nar/gkq300.

[16] Garnham Didier, M., Uran Landaburu, L. & Agüero, F. *Reposicionamiento in silico de compuestos bioactivos: identificación de módulos drogables conservados entre levaduras y tripanosomátidos* (UNSAM, 2021). URL https://github.com/trypanosomatics/didier-tesina.

[17] Solomon Nwaka & Alan Hudson. Innovative lead discovery strategies for tropical diseases. *Nat Rev Drug Discov* **5**, 941–955 (2006). URL http://dx.doi.org/10.1038/nrd2144.

[18] Crowther, G. J. *et al.* Identification of attractive drug targets in neglected-disease pathogens using an in silico approach. *PLoS Negl. Trop. Dis.* **4**, e804 (2010). URL https://doi.org/10.1371/journal.pntd.0000804.

[19] Bern, C. *et al.* Evaluation and treatment of chagas disease in the united states: a systematic review. *JAMA* **298**, 2171–2181 (2007). URL https://jamanetwork.com/journals/jama/fullarticle/209410.

[20] Hertweck, C. Natural products as source of therapeutics against parasitic diseases. *Angew. Chem. Int. Ed Engl.* **54**, 14622–14624 (2015). URL https://onlinelibrary.wiley.com/doi/abs/10.1002/anie.201509828.

[21] Steverding, D. The development of drugs for treatment of sleeping sickness: a historical review. *Parasit. Vectors* **3**, 15 (2010). URL https://www.sciencedirect.com/science/article/pii/S0960894X16303365.

[22] Klug, D. M., Gelb, M. H. & Pollastri, M. P. Repurposing strategies for tropical disease drug discovery. *Bioorg. Med. Chem. Lett.* **26**, 2569–2576 (2016). URL https://www.sciencedirect.com/science/article/pii/S0960894X16303365.

[23] William C Campbell. History of avermectin and ivermectin, with notes on the history of other macrocyclic lactone antiparasitic agents. *Curr Pharm Biotechnol* **13**, 853–865 (2012).

[24] Horton, J. The development of albendazole for lymphatic filariasis. *Ann. Trop. Med. Parasitol.* **103 Suppl 1**, S33–40 (2009). URL https://link.springer.com/article/10.1186/1756-3305-3-15.

[25] Donato Cioli & Livia Pica Mattoccia. Praziquantel. *Parasitol Res* **90 Supp 1**, S3–S9 (2003). URL http://dx.doi.org/10.1007/s00436-002-0751-z.

[26] Poulin, R., Lu, L., Ackermann, B., Bey, P. & Pegg, A. E. Mechanism of the irreversible inactivation of mouse ornithine decarboxylase by alpha-difluoromethylornithine. characterization of sequences at the inhibitor and coenzyme binding sites. *J. Biol. Chem.* **267**, 150–158 (1992). URL https://www.jbc.org/article/S0021-9258(18)48472-4/pdf.

[27] Bacchi, C. J., Nathan, H. C., Hutner, S. H., McCann, P. P. & Sjoerdsma, A. Polyamine metabolism: a potential therapeutic target in trypanosomes. *Science* **210**, 332–334 (1980). URL https://www.science.org/doi/10.1126/science.6775372.

[28] Osborne, C. K. Tamoxifen in the treatment of breast cancer. *N. Engl. J. Med.* **339**, 1609–1618 (1998). URL https://www.nejm.org/doi/full/10.1056/nejm199811263392207.

[29] Agüero, F. *et al.* Genomic-scale prioritization of drug targets: the TDR targets database. *Nat. Rev. Drug Discov.* **7**, 900–907 (2008). URL http://dx.doi.org/10.1038/nrd2684.

[30] Sugiyama, N., Imamura, H. & Ishihama, Y. Large-scale discovery of substrates of the human kinome. *Sci. Rep.* **9**, 10503 (2019). URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6642169/.

[31] Wu, P., Nielsen, T. E. & Clausen, M. H. FDA-approved small-molecule kinase inhibitors. *Trends Pharmacol. Sci.* **36**, 422–439 (2015). URL https://pubmed.ncbi.nlm.nih.gov/25975227/.

[32] Gaulton, A. *et al.* The ChEMBL database in 2017. *Nucleic Acids Research* **45**, D945–D954 (2016). URL https://doi.org/10.1093/nar/gkw1074.

[33] Parsons, M., Worthey, E. A., Ward, P. N. & Mottram, J. C. Comparative analysis of the kinomes of three pathogenic trypanosomatids: Leishmania major, trypanosoma brucei and trypanosoma cruzi. *BMC Genomics* **6**, 127 (2005). URL https://bmcgenomics.biomedcentral.com/articles/10.1186/1471-2164-6-127.

[34] Dichiara, M. *et al.* Repurposing of human kinase inhibitors in neglected protozoan diseases. *ChemMedChem* **12**, 1235–1253 (2017). URL https://pubmed.ncbi.nlm.nih.gov/29148875/.

[35] Patel, G. *et al.* Kinase scaffold repurposing for neglected disease drug discovery: discovery of an efficacious, lapatinib-derived lead compound for trypanosomiasis. *J. Med. Chem.* **56**, 3820–3832 (2013). URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3663906/.

[36] Peter J Hotez *et al.* Control of neglected tropical diseases. *N Engl J Med* **357**, 1018–1027 (2007). URL http://dx.doi.org/10.1056/NEJMra064142.

[37] Rassi, A. J., Rassi, A. & Marin Neto, J. A. Chagas disease. *Lancet* **375**, 1388–1402 (2010). URL http://dx.doi.org/10.1016/S0140-6736(10)60061-X.

[38] Peter J Hotez. Neglected infections of poverty in the United States of America. *PLoS Negl Trop Dis* **2**, e256 (2008). URL http://dx.doi.org/10.1371/journal.pntd.0000256.

[39] Macedo, A. M. & Pena, S. D. Genetic Variability of Trypanosoma cruzi:Implications for the Pathogenesis of Chagas Disease. *Parasitol Today* **14**, 119–124 (1998).

[40] Filardi, L. S. & Brener, Z. Susceptibility and natural resistance of Trypanosoma cruzi strains to drugs used clinically in Chagas disease. *Trans R Soc Trop Med Hyg* **81**, 755–759 (1987).

[41] Mejia, A. M. *et al.* Benznidazole-resistance in trypanosoma cruzi is a readily acquired trait that can arise independently in a single population. *The Journal of Infectious Diseases* **206**, 220–228 (2012). URL https://doi.org/10.1093/infdis/jis331.

[42] Najib M El Sayed *et al.* The genome sequence of Trypanosoma cruzi, etiologic agent of Chagas disease. *Science* **309**, 409–415 (2005). URL http://dx.doi.org/10.1126/science.1112631.

[43] Matthew Berriman *et al.* The genome of the African trypanosome Trypanosoma brucei. *Science* **309**, 416–422 (2005). URL http://dx.doi.org/10.1126/science.1112642.

[44] Alasdair C Ivens *et al.* The genome of the kinetoplastid parasite, Leishmania major. *Science* **309**, 436–442 (2005). URL http://dx.doi.org/10.1126/science.1112680.

[45] Higo, H. *et al.* Genotypic variation among lineages of trypanosoma cruzi and its geographic aspects. *Parasitol. Int.* **53**, 337–344 (2004). URL https://www.sciencedirect.com/science/article/abs/pii/S1383576904000637.

[46] Zingales, B. *et al.* The revised trypanosoma cruzi subspecific nomenclature: rationale, epidemiological relevance and research applications. *Infect. Genet. Evol.* **12**, 240–253 (2012). URL https://www.sciencedirect.com/science/article/abs/pii/S1567134811004564.

[47] Lambrecht, F. L. Biological variations in trypanosomes and their relation to the epidemiology of Chagas' disease. *Rev Inst Med Trop Sao Paulo* **7**, 346–352 (1965).

[48] Zeledón, R. & Rabinovich, J. E. Chagas' disease: an ecological appraisal with special emphasis on its insect vectors. *Annu Rev Entomol* **26**, 101–133 (1981). URL http://dx.doi.org/10.1146/annurev.en.26.010181.000533.

[49] Tanowitz, H. B. *et al.* Chagas' disease. *Clin Microbiol Rev* **5**, 400–419 (1992).

[50] María-Jesús Pinazo *et al.* Tolerance of benznidazole in treatment of Chagas' disease in adults. *Antimicrob Agents Chemother* **54**, 4896–4899 (2010). URL http://dx.doi.org/10.1128/AAC.00537-10.

[51] Diana L Fabbro *et al.* Trypanocide treatment among adults with chronic Chagas disease living in Santa Fe city (Argentina), over a mean follow-up of 21 years: parasitological, serological and clinical evolution. *Rev Soc Bras Med Trop* **40**, 1–10 (2007).

[52] Thomas L. Lemke, D. A. W. *Foye's Principles of Medicinal Chemistry* (Lippincott Williams & Wilkins, 2008).

[53] Despina Smirlis & Milena Botelho Pereira Soares. Selection of molecular targets for drug development against trypanosomatids. *Subcell Biochem* **74**, 43–76 (2014). URL http://dx.doi.org/10.1007/978-94-007-7305-9_2.

[54] Moreno, S. N., Docampo, R., Mason, R. P., Leon, W. & Stoppani, A. O. Different behaviors of benznidazole as free radical generator with mammalian and Trypanosoma cruzi microsomal preparations. *Arch Biochem Biophys* **218**, 585–591 (1982).

[55] Belinda S Hall & Shane R Wilkinson. Activation of benznidazole by trypanosomal type I nitroreductases results in glyoxal formation. *Antimicrob Agents Chemother* **56**, 115–123 (2012). URL http://dx.doi.org/10.1128/AAC.05135-11.

[56] Gauthier, J., Vincent, A. T., Charette, S. J. & Derome, N. A brief history of bioinformatics. *Brief. Bioinform.* **20**, 1981–1996 (2019).

[57] Sanger, F. & Thompson, E. O. P. The amino-acid sequence in the glycyl chain of insulin. 1. the identification of lower peptides from partial hydrolysates. *Biochem. J.* **53**, 353–366 (1953).

[58] Sanger, F. & Thompson, E. O. P. The amino-acid sequence in the glycyl chain of insulin. 2. the investigation of peptides from enzymic hydrolysates. *Biochem. J.* **53**, 366–374 (1953).

[59] Dayhoff, M. O. & Ledley, R. S. Comprotein. In *Proceedings of the December 4-6, 1962, fall joint computer conference on - AFIPS '62 (Fall)* (ACM Press, New York, New York, USA, 1962).

[60] IUPAC-IUB Comm. on Biochem. Nomenclature. A one-letter notation for amino acid sequences. tentative rules. *Biochemistry* **7**, 2703–2705 (1968).

[61] Dayhoff, M. O. *Atlas of protein sequence and structure*, vol. 4 (National Biomedical Research Foundation., 1969).

[62] Mistry, J. *et al.* Pfam: The protein families database in 2021. *Nucleic Acids Res.* **49**, D412–D419 (2021).

[63] Sigrist, C. J. A. *et al.* New and continuing developments at PROSITE. *Nucleic Acids Res.* **41**, D344–7 (2013).

[64] Jones, P. *et al.* InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240 (2014).

[65] Paysan-Lafosse, T. *et al.* InterPro in 2022. *Nucleic Acids Res.* **51**, D418–D427 (2023).

[66] Letunic, I., Khedkar, S. & Bork, P. SMART: recent updates, new developments and status in 2020. *Nucleic Acids Res.* **49**, D458–D460 (2021).

[67] Lu, S. *et al.* CDD/SPARCLE: the conserved domain database in 2020. *Nucleic Acids Res.* **48**, D265–D268 (2020).

[68] Gene Ontology Consortium. The gene ontology resource: enriching a GOld mine. *Nucleic Acids Res.* **49**, D325–D334 (2021).

[69] Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25**, 25–29 (2000). URL http://dx.doi.org/10.1038/75556.

[70] Koonin, E. V. Orthologs, paralogs, and evolutionary genomics. *Annu. Rev. Genet.* **39**, 309–338 (2005).

[71] Nichio, B. T. L., Marchaukoski, J. N. & Raittz, R. T. New tools in orthology analysis: A brief review of promising perspectives. *Front. Genet.* **8** (2017).

[72] Kim, K., Kim, W. & Kim, S. ReMark: an automatic program for clustering orthologs flexibly combining a recursive and a markov clustering algorithms. *Bioinformatics* **27**, 1731–1733 (2011). URL https://doi.org/10.1093/bioinformatics/btr259.

[73] Wang, Y., Coleman-Derr, D., Chen, G. & Gu, Y. Q. OrthoVenn: a web server for genome wide comparison and annotation of orthologous clusters across multiple species. *Nucleic Acids Research* **43**, W78–W84 (2015). URL https://doi.org/10.1093/nar/gkv487.

[74] Petersen, M. *et al.* Orthograph: a versatile tool for mapping coding nucleotide sequences to clusters of orthologous genes. *BMC Bioinformatics* **18** (2017). URL https://doi.org/10.1186/s12859-017-1529-8.

[75] Bitard-Feildel, T., Kemena, C., Greenwood, J. M. & Bornberg-Bauer, E. Domain similarity based orthology detection. *BMC Bioinformatics* **16**, 154 (2015).

[76] Schreiber, F. & Sonnhammer, E. L. L. Hieranoid: hierarchical orthology inference. *J. Mol. Biol.* **425**, 2072–2081 (2013).

[77] Gupta, S., Guru Nanak Dev Engineering College, Ludhiana, India & Singh, M. Phylogenetic method for high-throughput ortholog detection. *Int. J. Inf. Eng. Electron. Bus.* **7**, 51–59 (2015).

[78] Chen, F., Mackey, A. J., Vermunt, J. K. & Roos, D. S. Assessing performance of orthology detection strategies applied to eukaryotic genomes. *PLoS ONE* **2**, e383 (2007). URL https://doi.org/10.1371/journal.pone.0000383.

[79] Fischer, S. *et al.* Using OrthoMCL to assign proteins to OrthoMCL-DB groups or to cluster proteomes into new ortholog groups. *Curr. Protoc. Bioinformatics* **Chapter 6**, 6.12.1–6.12.19 (2011).

[80] William Lingran Chen. Chemoinformatics: past, present, and future. *J Chem Inf Model* **46**, 2230–2255 (2006). URL http://dx.doi.org/10.1021/ci060016u.

[81] David S Wishart. Introduction to cheminformatics. *Curr Protoc Bioinformatics* **Chapter 14**, Unit 14.1 (2007). URL http://dx.doi.org/10.1002/0471250953.bi1401s18.

[82] Saber A Akhondi, Jan A Kors & Sorel Muresan. Consistency of systematic chemical identifiers within and between small-molecule databases. *J Cheminform* **4**, 35 (2012). URL http://dx.doi.org/10.1186/1758-2946-4-35.

[83] Accelerys. *CT File Formats, Symyx Solutions* (2010). URL http://accelrys.com/.

[84] Daylight Chemical Information Systems, Inc. *Daylight Theory Manual* (2011). URL http://www.daylight.com/dayhtml/doc/theory/.

[85] Stephen Heller, Alan McNaught, Stephen Stein, Dmitrii Tchekhovskoi & Igor Pletnev. InChI - the worldwide chemical structure identifier standard. *J Cheminform* **5**, 7 (2013). URL http://dx.doi.org/10.1186/1758-2946-5-7.

[86] InChI FAQ (2012). URL http://www.inchi-trust.org/fileadmin/user_upload/software/inchi-faq/inchi-faq.pdf.

[87] Christopher Southan. InChI in the wild: an assessment of InChIKey searching in Google. *J Cheminform* **5**, 10 (2013). URL http://dx.doi.org/10.1186/1758-2946-5-10.

[88] InChI version 1, software version 1.04 (September 2011) User's Guide (2011). URL http://www.inchi-trust.org/fileadmin/user_upload/software/inchi-v1.04/InChI_UserGuide.pdf.

[89] Stephen E. Stein, Stephen R. Heller, Dmitrii V. Tchekhovskoi & Igor V. Pletnev. InChI version 1, software version 1.04 (2011) Technical Manual (2011). URL http://www.inchi-trust.org/fileadmin/user_upload/software/inchi-v1.04/InChI_TechMan.pdf.

[90] Gisbert Schneider. Madame Curie Bioscience Database. Drug Design. URL http://www.ncbi.nlm.nih.gov/books/NBK6122/#top.

[91] Ertl, P., Rohde, B. & Selzer, P. Fast calculation of molecular polar surface area as a sum of fragment-based contributions and its application to the prediction of drug transport properties. *J Med Chem* **43**, 3714–3717 (2000).

[92] Cramer, C. *Essentials of Computational Chemistry: Theories and Models* (Wiley, 2005). URL http://books.google.com.ar/books?id=tNiyZjAZqKkC.

[93] Ulf Norinder & Christel A S Bergström. Prediction of ADMET Properties. *ChemMedChem* **1**, 920–937 (2006). URL http://dx.doi.org/10.1002/cmdc.200600155.

[94] Lipinski, C. A., Lombardo, F., Dominy, B. W. & Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv Drug Deliv Rev* **46**, 3–26 (2001).

[95] Daniel F Veber *et al.* Molecular properties that influence the oral bioavailability of drug candidates. *J Med Chem* **45**, 2615–2623 (2002).

[96] Barry Hardy *et al.* Collaborative development of predictive toxicology applications. *J Cheminform* **2**, 7 (2010). URL http://dx.doi.org/10.1186/1758-2946-2-7.

[97] Andreas Bender & Robert C Glen. Molecular similarity: a key technique in molecular informatics. *Org Biomol Chem* **2**, 3204–3218 (2004). URL http://dx.doi.org/10.1039/B409813G.

[98] Ray M Marín, Nestor F Aguirre & Edgar E Daza. Graph theoretical similarity approach to compare molecular electrostatic potentials. *J Chem Inf Model* **48**, 109–118 (2008). URL http://dx.doi.org/10.1021/ci7001878.

[99] Brown, N. Chemoinformatics&Mdash;an Introduction for Computer Scientists. *ACM Comput. Surv.* **41**, 8:1–8:38 (2009). URL http://doi.acm.org/10.1145/1459352.1459353.

[100] Kirkpatrick, P. & Ellis, C. Chemical Space. *Nature* **432**, 823 (2004).

[101] Christopher M Dobson. Chemical space and biology. *Nature* **432**, 824–828 (2004). URL http://dx.doi.org/10.1038/nature03192.

[102] Rogers, D. & Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inf. Model.* **50**, 742–754 (2010).

[103] Flower, D. On the Properties of Bit String-Based Measures of Chemical Similarity. *J. Chem. Inf. Comput. Sci.* **38**, 379–386 (1998).

[104] Pierre Baldi, Ryan W Benz, Daniel S Hirschberg & Joshua Swamidass, S. Lossless compression of chemical fingerprints using integer entropy codes improves storage and retrieval. *J Chem Inf Model* **47**, 2098–2109 (2007). URL http://dx.doi.org/10.1021/ci700200n.

[105] Landrum, G. *et al.* rdkit/rdkit: 2020_03_1 (q1 2020) release (2020).

[106] Dalke, A. The chemfp project. *Journal of Cheminformatics* **11**, 76 (2019). URL https://doi.org/10.1186/s13321-019-0398-8.

[107] Bero, S. A., Muda, A. K., Choo, Y.-H., Muda, N. A. & Pratama, S. F. Weighted tanimoto coefficient for 3D molecule structure similarity measurement. *ArXiv* (2018).

[108] Daylight Chemical Information Systems, Inc. *Daylight Clustering Manual* (2011). URL http://www.daylight.com/dayhtml/doc/cluster/index.html.

[109] Willett, J. . *Similarity and Clustering in Chemical Information Systems* (John Wiley & Sons, Inc., New York, NY, USA, 1987).

[110] Jarvis, R. A. & Patrick, E. A. Clustering Using a Similarity Measure Based on Shared Near Neighbors. *IEEE Trans. Comput.* **22**, 1025–1034 (1973). URL http://dx.doi.org/10.1109/T-C.1973.223640.

[111] Downs, G. M. & Barnard, J. M. Clustering methods and their uses in computational chemistry. In *Reviews in Computational Chemistry, Volume 18*, 1–40 (John Wiley & Sons, Inc., Hoboken, New Jersey, USA, 2003).

[112] Wagen, C. Dimensionality reduction in cheminformatics. https://corinwagen.github.io/public/blog/20230417_dimensionality_reduction.html. Accessed: 2023-09-13.

[113] Cacace, E., Kritikos, G. & Typas, A. Chemical genetics in drug discovery. *Current Opinion in Systems Biology* **4**, 35–42 (2017). URL https://doi.org/10.1016/j.coisb.2017.05.020.

[114] Yeung, C. H. L., Sahin, N. & Andrews, B. Phenomics approaches to understand genetic networks and gene function in yeast. *Biochemical Society Transactions* **50**, 713–721 (2022). URL https://doi.org/10.1042/bst20210285.

[115] Lee, A. Y., Bader, G. D., Nislow, C. & Giaever, G. Chemogenomic profiling. In *Handbook of Systems Biology*, 153–176 (Elsevier, 2013). URL https://doi.org/10.1016/b978-0-12-385944-0.00008-3.

[116] Xue, A., Robbins, N. & Cowen, L. E. Advances in fungal chemical genomics for the discovery of new antifungal agents. *Annals of the New York Academy of Sciences* **1496**, 5–22 (2020). URL https://doi.org/10.1111/nyas.14484.

[117] Wong, J. H. *et al.* Chemogenomic profiling in yeast reveals antifungal mode-of-action of polyene macrolactam auroramycin. *PLOS ONE* **14**, e0218189 (2019). URL https://doi.org/10.1371/journal.pone.0218189.

[118] Fletcher, E., Mercurio, K., Walden, E. A. & Baetz, K. A yeast chemogenomic screen identifies pathways that modulate adipic acid toxicity. *iScience* **24**, 102327 (2021). URL https://doi.org/10.1016/j.isci.2021.102327.

[119] Alsford, S. *et al.* High-throughput decoding of antitrypanosomal drug efficacy and resistance. *Nature* **482**, 232–236 (2012). URL https://doi.org/10.1038/nature10771.

[120] Baker, N., Alsford, S. & Horn, D. Genome-wide RNAi screens in african trypanosomes identify the nifurtimox activator NTR and the eflornithine transporter AAT6. *Molecular and Biochemical Parasitology* **176**, 55–57 (2011). URL https://doi.org/10.1016/j.molbiopara.2010.11.010.

[121] Thomas, J. A. *et al.* Insights into antitrypanosomal drug mode-of-action from cytology-based profiling. *PLOS Neglected Tropical Diseases* **12**, e0006980 (2018). URL https://doi.org/10.1371/journal.pntd.0006980.

[122] Collett, C. F. *et al.* Chemogenomic profiling of antileishmanial efficacy and resistance in the related kinetoplastid parasite trypanosoma brucei. *Antimicrobial Agents and Chemotherapy* **63** (2019). URL https://doi.org/10.1128/aac.00795-19.

[123] Huesken, D. *et al.* Design of a genome-wide siRNA library using an artificial neural network. *Nature Biotechnology* **23**, 995–1001 (2005). URL https://doi.org/10.1038/nbt1118.

[124] Qiu, S. A computational study of off-target effects of RNA interference. *Nucleic Acids Research* **33**, 1834–1847 (2005). URL https://doi.org/10.1093/nar/gki324.

[125] Stortz, J. A. *et al.* Genome-wide and protein kinase-focused RNAi screens reveal conserved and novel damage response pathways in trypanosoma brucei. *PLOS Pathogens* **13**, e1006477 (2017). URL https://doi.org/10.1371/journal.ppat.1006477.

[126] elodie Gazanion, Fernandez-Prada, C., Papadopoulou, B., Leprohon, P. & Ouellette, M. Cos-seq for high-throughput identification of drug target and resistance mechanisms in the protozoan parasite leishmania. *Proceedings of the National Academy of Sciences* **113** (2016). URL https://doi.org/10.1073/pnas.1520693113.

[127] Fernandez-Prada, C. *et al.* High-throughput cos-seq screen with intracellular leishmania infantum for the discovery of novel drug-resistance mechanisms. *International Journal for Parasitology: Drugs and Drug Resistance* **8**, 165–173 (2018). URL https://doi.org/10.1016/j.ijpddr.2018.03.004.

[128] Leprohon, P., Fernandez-Prada, C., elodie Gazanion, Monte-Neto, R. & Ouellette, M. Drug resistance analysis by next generation sequencing in leishmania. *International Journal for Parasitology: Drugs and Drug Resistance* **5**, 26–35 (2015). URL https://doi.org/10.1016/j.ijpddr.2014.09.005.

[129] Wall, R. J. *et al.* Clinical and veterinary trypanocidal benzoxaboroles target CPSF3. *Proceedings of the National Academy of Sciences* **115**, 9616–9621 (2018). URL https://doi.org/10.1073/pnas.1807915115.

[130] Dickie, E. A. *et al.* New drugs for human african trypanosomiasis: A twenty first century success story. *Tropical Medicine and Infectious Disease* **5**, 29 (2020). URL https://doi.org/10.3390/tropicalmed5010029.

[131] Mateus, A., Määttä, T. A. & Savitski, M. M. Thermal proteome profiling: unbiased assessment of protein state through heat-induced stability changes. *Proteome Science* **15** (2016). URL https://doi.org/10.1186/s12953-017-0122-4.

[132] Corpas-Lopez, V. & Wyllie, S. Utilizing thermal proteome profiling to identify the molecular targets of anti-leishmanial compounds. *STAR Protocols* **2**, 100704 (2021). URL https://doi.org/10.1016/j.xpro.2021.100704.

[133] Jafari, R. *et al.* The cellular thermal shift assay for evaluating drug target interactions in cells. *Nature Protocols* **9**, 2100–2122 (2014). URL https://doi.org/10.1038/nprot.2014.138.

[134] Mateus, A. *et al.* Thermal proteome profiling for interrogating protein interactions. *Molecular Systems Biology* **16** (2020). URL https://doi.org/10.15252/msb.20199232.

[135] Corpas-Lopez, V. *et al.* Pharmacological validation of n-myristoyltransferase as a drug target in leishmania donovani. *ACS Infectious Diseases* **5**, 111–122 (2018). URL https://doi.org/10.1021/acsinfecdis.8b00226.

[136] den Kerkhof, M. V. *et al.* Antileishmanial aminopyrazoles: Studies into mechanisms and stability of experimental drug resistance. *Antimicrobial Agents and Chemotherapy* **64** (2020). URL https://doi.org/10.1128/aac.00152-20.

[137] Douanne, N. *et al.* MRPA-independent mechanisms of antimony resistance in leishmania infantum. *International Journal for Parasitology: Drugs and Drug Resistance* **13**, 28–37 (2020). URL https://doi.org/10.1016/j.ijpddr.2020.03.003.

[138] Yasur-Landau, D., Jaffe, C. L., David, L., Doron-Faigenboim, A. & Baneth, G. Resistance of leishmania infantum to allopurinol is associated with chromosome and gene copy number variations including decrease in the s-adenosylmethionine synthetase (METK) gene copy number. *International Journal for Parasitology: Drugs and Drug Resistance* **8**, 403–410 (2018). URL https://doi.org/10.1016/j.ijpddr.2018.08.002.

[139] Roy, G., Bhattacharya, A., Leprohon, P. & Ouellette, M. Decreased glutamate transport in acivicin resistant leishmania tarentolae. *PLOS Neglected Tropical Diseases* **15**, e0010046 (2021). URL https://doi.org/10.1371/journal.pntd.0010046.

[140] Wyllie, S. *et al.* Nitroheterocyclic drug resistance mechanisms in trypanosoma brucei. *Journal of Antimicrobial Chemotherapy* **71**, 625–634 (2015). URL https://doi.org/10.1093/jac/dkv376.

[141] Graf, F. E. *et al.* Comparative genomics of drug resistance in trypanosoma brucei rhodesiense. *Cellular and Molecular Life Sciences* **73**, 3387–3400 (2016). URL https://doi.org/10.1007/s00018-016-2173-6.

[142] Giordani, F. *et al.* Veterinary trypanocidal benzoxaboroles are peptidase-activated prodrugs. *PLOS Pathogens* **16**, e1008932 (2020). URL https://doi.org/10.1371/journal.ppat.1008932.

[143] Mondelaers, A. *et al.* Genomic and molecular characterization of miltefosine resistance in leishmania infantum strains with either natural or acquired resistance through experimental selection of intracellular amastigotes. *PLOS ONE* **11**, e0154101 (2016). URL https://doi.org/10.1371/journal.pone.0154101.

[144] Rosa-Teijeiro, C. *et al.* Three different mutations in the DNA topoisomerase 1b in leishmania infantum contribute to resistance to antitumor drug topotecan. *Parasites and Vectors* **14** (2021). URL https://doi.org/10.1186/s13071-021-04947-4.

[145] Bhattacharya, A. *et al.* Coupling chemical mutagenesis to next generation sequencing for the identification of drug resistance mutations in leishmania. *Nature Communications* **10** (2019). URL https://doi.org/10.1038/s41467-019-13344-6.

[146] Hendrickx, S., Reis-Cunha, J. L., Forrester, S., Jeffares, D. C. & Caljon, G. Experimental selection of paromomycin resistance in leishmania donovani amastigotes induces variable genomic polymorphisms. *Microorganisms* **9**, 1546 (2021). URL https://doi.org/10.3390/microorganisms9081546.

[147] Barja, P. P. *et al.* Haplotype selection as an adaptive mechanism in the protozoan pathogen leishmania donovani. *Nature Ecology & Evolution* **1**, 1961–1969 (2017). URL https://doi.org/10.1038/s41559-017-0361-x.

[148] Wooller, S. K., Benstead-Hume, G., Chen, X., Ali, Y. & Pearl, F. M. G. Bioinformatics in translational drug discovery. *Biosci. Rep.* **37** (2017).

[149] Magari nos, M. P. *et al.* TDR Targets: a chemogenomics resource for neglected diseases. *Nucleic Acids Res* **40**, D1118–D1127 (2012). URL http://dx.doi.org/10.1093/nar/gkr1053.

[150] Lykins, J. D. *et al.* CSGID solves structures and identifies phenotypes for five enzymes in toxoplasma gondii. *Front. Cell. Infect. Microbiol.* **8**, 352 (2018).

[151] Shanmugam, D. *et al.* *Integrating and Mining Helminth Genomes to Discover and Prioritize Novel Therapeutic Targets* (Wiley, 2012). URL https://doi.org/10.1002/9783527652969.ch3.

[152] Berenstein, A. J., Magariños, M. P., Chernomoretz, A. & Agüero, F. A multilayer network approach for guiding drug repositioning in neglected diseases. *PLoS Negl. Trop. Dis.* **10**, e0004300 (2016). URL http://dx.plos.org/10.1371/journal.pntd.0004300.

[153] Kim, K. & Weiss, L. M. Toxoplasma gondii: the model apicomplexan. *Int. J. Parasitol.* **34**, 423–432 (2004).

[154] Sidik, S. M. *et al.* A genome-wide CRISPR screen in toxoplasma identifies essential apicomplexan genes. *Cell* **166**, 1423–1435.e12 (2016). URL https://linkinghub.elsevier.com/retrieve/pii/S0092867416310704.

[155] Gajria, B. *et al.* ToxoDB: an integrated toxoplasma gondii database resource. *Nucleic Acids Res.* **36**, D553–6 (2008).

[156] Warrenfeltz, S. *et al.* EuPathDB: The eukaryotic pathogen genomics database resource. In Kollmar, M. (ed.) *Methods in Molecular Biology*, vol. 1757 of *Methods in molecular biology (Clifton, N.J.)*, 69–113 (Springer New York, New York, NY, 2018). URL http://link.springer.com/10.1007/978-1-4939-7737-6_5.

[157] Sayers, E. W. *et al.* Database resources of the national center for biotechnology information in 2023. *Nucleic Acids Res.* (2022).

[158] Hertz-Fowler, C. & Peacock, C. S. Introducing GeneDB: a generic database. *Trends Parasitol.* **18**, 465–467 (2002). URL https://linkinghub.elsevier.com/retrieve/pii/S1471492202023619.

[159] Bolt, B. J. *et al.* Using WormBase ParaSite: An integrated platform for exploring helminth genomic data. In *Methods in Molecular Biology*, Methods in molecular biology (Clifton, N.J.), 471–491 (Springer New York, New York, NY, 2018).

[160] Lechat, P., Hummel, L., Rousseau, S. & Moszer, I. GenoList: an integrated environment for comparative analysis of microbial genomes. *Nucleic Acids Res.* **36**, D469–74 (2008).

[161] Kapopoulou, A., Lew, J. M. & Cole, S. T. The MycoBrowser portal: a comprehensive and manually annotated resource for mycobacterial genomes. *Tuberculosis (Edinb.)* **91**, 8–13 (2011). URL https://linkinghub.elsevier.com/retrieve/pii/S1472979210001095.

[162] Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009). URL http://www.biomedcentral.com/1471-2105/10/421.

[163] Hancock, J. M. & Bishop, M. J. EMBOSS (the european molecular biology open software suite). In *Dictionary of Bioinformatics and Computational Biology*, dob0206.pub2 (John Wiley & Sons, Ltd, Chichester, UK, 2004). URL http://doi.wiley.com/10.1002/9780471650126.dob0206.pub2.

[164] Krogh, A., Larsson, B., von Heijne, G. & Sonnhammer, E. L. Predicting transmembrane protein topology with a hidden markov model: application to complete genomes. *J. Mol. Biol.* **305**, 567–580 (2001).

[165] Almagro Armenteros, J. J. *et al.* SignalP 5.0 improves signal peptide predictions using deep neural networks. *Nat. Biotechnol.* **37**, 420–423 (2019). URL http://www.nature.com/articles/s41587-019-0036-z.

[166] S. Punla, C., , C. Farro, R. & Bataan Peninsula State University Dinalupihan, Bataan, Philippines. Are we there yet?: An analysis of the competencies of BEED graduates of BPSU-DC. *International Multidisciplinary Research Journal* **4**, 50–59 (2022).

[167] Moriya, Y., Itoh, M., Okuda, S., Yoshizawa, A. C. & Kanehisa, M. KAAS: an automatic genome annotation and pathway reconstruction server. *Nucleic Acids Res.* **35**, W182–5 (2007).

[168] Chen, F., Mackey, A. J., Stoeckert, C. J., Jr & Roos, D. S. OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. *Nucleic Acids Res.* **34**, D363–8 (2006).

[169] Mitchell, A. L. *et al.* InterPro in 2019: improving coverage, classification and access to protein sequence annotations. *Nucleic Acids Res.* **47**, D351–D360 (2019).

[170] Burley, S. K. *et al.* Protein data bank: the single global archive for 3D macromolecular structure data. *Nucleic Acids Res* **47**, D520–D528 (2019).

[171] Pieper, U. *et al.* ModBase, a database of annotated comparative protein structure models and associated resources. *Nucleic Acids Res.* **42**, D336–46 (2014).

[172] Zhu, L. *et al.* New insights into the plasmodium vivax transcriptome using RNA-Seq. *Sci. Rep.* **6** (2016).

[173] Smircich, P. *et al.* Ribosome profiling reveals translation control as a key mechanism generating differential gene expression in trypanosoma cruzi. *BMC Genomics* **16**, 443 (2015).

[174] Lasonder, E. *et al.* Integrated transcriptomic and proteomic analyses of p. falciparum gametocytes: molecular insight into sex-specific processes and translational repression. *Nucleic Acids Res.* **44**, 6087–6101 (2016).

[175] Otto, T. D. *et al.* New insights into the blood-stage transcriptome of plasmodium falciparum using RNA-Seq. *Mol. Microbiol.* **76**, 12–24 (2010).

[176] Otto, T. D. *et al.* A comprehensive evaluation of rodent malaria parasite genomes and gene expression. *BMC Biol.* **12** (2014).

[177] Fernandes, M. C. *et al. Dual Transcriptome Profiling ofLeishmania-Infected Human Macrophages Reveals Distinct Reprogramming Signatures*, vol. 7 (American Society for Microbiology, 2016). URL https://doi.org/10.1128/mbio.00027-16.

[178] Fritz, H. M. *et al.* Transcriptomic analysis of toxoplasma development reveals many novel functions and structures specific to sporozoites and oocysts. *PLoS One* **7**, e29998 (2012).

[179] Hon, C.-C. *et al.* Quantification of stochastic noise of splicing and polyadenylation in entamoeba histolytica. *Nucleic Acids Res.* **41**, 1936–1952 (2013).

[180] Siegel, T. N., Hekstra, D. R., Wang, X., Dewell, S. & Cross, G. A. M. Genome-wide analysis of mRNA abundance in two life-cycle stages of trypanosoma brucei and identification of splicing and polyadenylation sites. *Nucleic Acids Res.* **38**, 4946–4957 (2010).

[181] Yeoh, L. M., Goodman, C. D., Mollard, V., McFadden, G. I. & Ralph, S. A. Comparative transcriptomics of female and male gametocytes in plasmodium berghei and the evolution of sex in alveolates. *BMC Genomics* **18**, 734 (2017).

[182] Hehl, A. B. *et al.* Asexual expansion of toxoplasma gondii merozoites is distinct from tachyzoites and entails expression of non-overlapping gene families to attach, invade, and replicate within feline enterocytes. *BMC Genomics* **16**, 66 (2015).

[183] Bushell, E. *et al.* Functional profiling of a plasmodium genome reveals an abundance of essential genes. *Cell* **170**, 260–272.e8 (2017).

[184] Mendez, D. *et al.* ChEMBL: towards direct deposition of bioassay data. *Nucleic Acids Res.* **47**, D930–D940 (2019).

[185] Peña, I. *et al.* New compound sets identified from high throughput phenotypic screening against three kinetoplastid parasites: An open resource. *Sci. Rep.* **5**, 8771 (2015). URL http://www.nature.com/articles/srep08771.

[186] Spangenberg, T. *et al.* The open access malaria box: a drug discovery catalyst for neglected diseases. *PLoS One* **8**, e62906 (2013). URL http://dx.plos.org/10.1371/journal.pone.0062906.

[187] Haider, N. Functionality pattern matching as an efficient complementary structure/reaction search tool: an open-source approach. *Molecules* **15**, 5079–5092 (2010). URL http://dx.doi.org/10.3390/molecules15085079.

[188] Cheng, T. *et al.* Computation of octanol-water partition coefficients by guiding an additive model with knowledge. *J. Chem. Inf. Model.* **47**, 2140–2148 (2007). URL http://dx.doi.org/10.1021/ci700257y.

[189] Heller, S. R., McNaught, A., Pletnev, I., Stein, S. & Tchekhovskoi, D. InChI, the IUPAC international chemical identifier. *J. Cheminform.* **7**, 23 (2015).

[190] Al-Lazikani, B. Rule of five (lipinski rule of five). In *Dictionary of Bioinformatics and Computational Biology*, dob1075 (John Wiley & Sons, Ltd, Chichester, UK, 2004). URL http://doi.wiley.com/10.1002/9780471650126.dob1075.

[191] Congreve, M., Carr, R., Murray, C. & Jhoti, H. A 'rule of three' for fragment-based lead discovery? drug discov. *Drug Discov. Today* **8**, 876–877 (2003).

[192] Rogers, D. J. & Fleming, H. A computer program for classifying plants II. a numerical handling of non-numerical data. *Bioscience* **14**, 15–28 (1964).

[193] WJ, Y. Index for rating diagnostic tests. *Cancer* **3**, 32–35 (1950).

[194] Bostock, M., Ogievetsky, V. & Heer, J. D3: Data-Driven documents. *IEEE Trans. Vis. Comput. Graph.* **17**, 2301–2309 (2011).

[195] Probst, D. & Reymond, J.-L. SmilesDrawer: Parsing and drawing SMILES-encoded molecular structures using client-side JavaScript. *J. Chem. Inf. Model.* **58**, 1–7 (2018). URL http://pubs.acs.org/doi/10.1021/acs.jcim.7b00425.

[196] Mosleh, M., Dalili, K. & Heydari, B. Distributed or monolithic? a computational architecture decision framework. *IEEE Syst. J.* **12**, 125–136 (2018).

[197] Al-Debagy, O. & Martinek, P. A comparative review of microservices and monolithic architectures (2019).

[198] Andrews, K. T., Fisher, G. & Skinner-Adams, T. S. Drug repurposing and human parasitic protozoan diseases. *International Journal for Parasitology: Drugs and Drug Resistance* **4**, 95–111 (2014). URL https://www.sciencedirect.com/science/article/pii/S2211320714000050.

[199] Miguel, D. C. *et al.* Tamoxifen as a potential antileishmanial agent: efficacy in the treatment of Leishmania braziliensis and Leishmania chagasi infections. *The Journal of Antimicrobial Chemotherapy* **63**, 365–368 (2009).

[200] Molina, I. *et al.* Randomized trial of posaconazole and benznidazole for chronic chagas' disease. *N. Engl. J. Med.* **370**, 1899–1908 (2014).

[201] Volochnyuk, D. M. *et al.* Evolution of commercially available compounds for HTS. *Drug Discovery Today* **24**, 390–402 (2019). URL https://www.sciencedirect.com/science/article/pii/S1359644618302423.

[202] Grygorenko, O. O. *et al.* Generating Multibillion Chemical Space of Readily Accessible Screening Compounds. *iScience* **23**, 101681 (2020). URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7593547/.

[203] Rex, D. A. B. *et al.* Dissecting plasmodium yoelii pathobiology: Proteomic approaches for decoding novel translational and post-translational modifications. *ACS Omega* **7**, 8246–8257 (2022).

[204] Knoll, K. E., van der Walt, M. M. & Loots, D. T. In silico drug discovery strategies identified ADMET properties of decoquinate RMB041 and its potential drug targets against mycobacterium tuberculosis. *Microbiol. Spectr.* **10**, e0231521 (2022).

[205] Beteck, R. M. *et al.* Accessible and distinct decoquinate derivatives active against mycobacterium tuberculosis and apicomplexan parasites. *Commun. Chem.* **1** (2018).

[206] Padalino, G., Chalmers, I. W., Brancale, A. & Hoffmann, K. F. Identification of 6-(piperazin-1-yl)-1,3,5-triazine as a chemical scaffold with broad anti-schistosomal activities. *Wellcome Open Res.* **5**, 169 (2020).

[207] Reis, I. M. A. *et al.* γ-Lactones from persea americana and persea fulva - in vitro and in silico evaluation of trypanosoma cruzi activity. *Chem. Biodivers.* **18**, e2100362 (2021).

[208] Shah-Simpson, S., Lentini, G., Dumoulin, P. C. & Burleigh, B. A. Modulation of host central carbon metabolism and in situ glucose uptake by intracellular trypanosoma cruzi amastigotes. *PLoS Pathog.* **13**, e1006747 (2017).

[209] Marchese, L. *et al.* The uptake and metabolism of amino acids, and their unique role in the biology of pathogenic trypanosomatids. *Pathogens* **7**, 36 (2018).

[210] James, G., Witten, D., Hastie, T. & Tibshirani, R. (eds.) *An introduction to statistical learning: with applications in R.* No. 103 in Springer texts in statistics (Springer, New York, 2013). OCLC: ocn828488009.

[211] Maaten, L. v. d. & Hinton, G. Visualizing Data using t-SNE. *Journal of Machine Learning Research* **9**, 2579–2605 (2008). URL http://jmlr.org/papers/v9/vandermaaten08a.html.

[212] Vanden Eynde, J. J. *et al.* Alkanediamide-linked bisbenzamidines are promising antiparasitic agents. *Pharmaceuticals (Basel)* **9**, 20 (2016).

[213] Chohan, Z. H. *et al.* Sulfonamide-metal complexes endowed with potent anti-trypanosoma cruzi activity. *J. Enzyme Inhib. Med. Chem.* **29**, 230–236 (2014).

[214] Mezencev, R., Galizzi, M., Kutschy, P. & Docampo, R. Trypanosoma cruzi: antiproliferative effect of indole phytoalexins on intracellular amastigotes in vitro. *Exp. Parasitol.* **122**, 66–69 (2009).

[215] Ciammaichella, A. *et al.* Optimization of 2-(1h-imidazo-2-yl)piperazines series of trypanosoma brucei growth inhibitors as potential treatment for the second stage of HAT. *Bioorg. Med. Chem. Lett.* **30**, 127207 (2020).

[216] Goad, L. J., Berens, R. L., Marr, J. J., Beach, D. H. & Holz, G. G., Jr. The activity of ketoconazole and other azoles against trypanosoma cruzi: biochemistry and chemotherapeutic action in vitro. *Mol. Biochem. Parasitol.* **32**, 179–189 (1989).

[217] de Oliveira Filho, G. B. *et al.* Structural design, synthesis and pharmacological evaluation of thiazoles against trypanosoma cruzi. *Eur. J. Med. Chem.* **141**, 346–361 (2017).

[218] Rocha, Y. M. *et al.* Antiparasitary and antiproliferative activities in vitro of a 1,2,4-oxadiazole derivative on trypanosoma cruzi. *Parasitol. Res.* **121**, 2141–2156 (2022).

[219] Zuma, A. A. *et al.* Furan derivatives impair proliferation and affect ultrastructural organization of trypanosoma cruzi and leishmania amazonensis. *Exp. Parasitol.* **224**, 108100 (2021).

[220] Batista, J. M., Jr *et al.* Natural chromenes and chromene derivatives as potential anti-trypanosomal agents. *Biol. Pharm. Bull.* **31**, 538–540 (2008).

[221] Racané, L. *et al.* Synthesis, antiproliferative and antitrypanosomal activities, and DNA binding of novel 6-amidino-2-arylbenzothiazoles. *J. Enzyme Inhib. Med. Chem.* **36**, 1952–1967 (2021).

[222] Moreira, D. R. M. *et al.* Structural design, synthesis and structure-activity relationships of thiazolidinones with enhanced anti-trypanosoma cruzi activity. *ChemMedChem* **9**, 177–188 (2014).

[223] Buckner, F. S., Verlinde, C. L., La Flamme, A. C. & Van Voorhis, W. C. Efficient technique for screening drugs for activity against trypanosoma cruzi using parasites expressing beta-galactosidase. *Antimicrob. Agents Chemother.* **40**, 2592–2597 (1996).

[224] Rolón, M., Vega, C., Escario, J. A. & Gómez-Barrio, A. Development of resazurin microtiter assay for drug sensibility testing of Trypanosoma cruzi epimastigotes. *Parasitology Research* **99**, 103–107 (2006).

[225] Fleau, C. *et al.* Chagas disease drug discovery: Multiparametric lead optimization against trypanosoma cruzi in acylaminobenzothiazole series. *J. Med. Chem.* **62**, 10362–10375 (2019).

[226] Bettiol, E. *et al.* Identification of three classes of heteroaromatic compounds with activity against intracellular trypanosoma cruzi by chemical library screening. *PLoS Negl. Trop. Dis.* **3**, e384 (2009).

[227] Igoillo-Esteve, M. & Cazzulo, J. J. The glucose-6-phosphate dehydrogenase from trypanosoma cruzi: its role in the defense of the parasite against oxidative stress. *Mol. Biochem. Parasitol.* **149**, 170–181 (2006).

[228] Gupta, S., Igoillo-Esteve, M., Michels, P. A. M. & Cordeiro, A. T. Glucose-6-phosphate dehydrogenase of trypanosomatids: characterization, target validation, and drug discovery. *Mol. Biol. Int.* **2011**, 135701 (2011).

[229] Mercaldi, G. F., Dawson, A., Hunter, W. N. & Cordeiro, A. T. The structure of a trypanosoma cruzi glucose-6-phosphate dehydrogenase reveals differences from the mammalian enzyme. *FEBS Lett.* **590**, 2776–2786 (2016).

[230] Gupta, S., Cordeiro, A. T. & Michels, P. A. M. Glucose-6-phosphate dehydrogenase is the target for the trypanocidal action of human steroids. *Mol. Biochem. Parasitol.* **176**, 112–115 (2011).

[231] Ortiz, C. *et al.* Binding mode and selectivity of steroids towards glucose-6-phosphate dehydrogenase from the pathogen trypanosoma cruzi. *Molecules* **21**, 368 (2016).

[232] Fredo Naciuk, F., do Nascimento Faria, J., Gonçalves Eufrásio, A., Torres Cordeiro, A. & Bruder, M. Development of selective steroid inhibitors for the glucose-6-phosphate dehydrogenase from trypanosoma cruzi. *ACS Med. Chem. Lett.* **11**, 1250–1256 (2020).

[233] Else, A. J. *et al.* Dihydrolipoamide dehydrogenase in the trypanosoma subgenus, trypanozoon. *Mol. Biochem. Parasitol.* **64**, 233–239 (1994).

[234] Solmonson, A. & DeBerardinis, R. J. Lipoic acid metabolism and mitochondrial redox regulation. *The Journal of Biological Chemistry* **293**, 7522–7530 (2018).

[235] Berg, J., Tymoczko, J. & Stryer, L. *Biochemistry*. Biochemistry (Berg) (W. H. Freeman, 2007). URL https://books.google.com.ar/books?id=Uhm_ngEACAAJ.

[236] Gutierrez-Correa, J. Trypanosoma cruzi dihydrolipoamide dehydrogenase as target for phenothiazine cationic radicals. effect of antioxidants. *Curr. Drug Targets* **7**, 1155–1179 (2006).

[237] Dos Santos, P. F. *et al.* Molecular characterization of lipoamide dehydrogenase gene in trypanosoma cruzi populations susceptible and resistant to benznidazole. *Exp. Parasitol.* **170**, 1–9 (2016).

[238] Roldán, A., Comini, M. A., Crispo, M. & Krauth-Siegel, R. L. Lipoamide dehydrogenase is essential for both bloodstream and procyclic trypanosoma brucei. *Mol. Microbiol.* **81**, 623–639 (2011).

[239] Alsford, S. *et al.* High-throughput phenotyping using parallel sequencing of RNA interference targets in the african trypanosome. *Genome Res.* **21**, 915–924 (2011). URL http://genome.cshlp.org/cgi/doi/10.1101/gr.115089.110.

[240] Vacchina, P., Lambruschi, D. A. & Uttaro, A. D. Lipoic acid metabolism in trypanosoma cruzi as putative target for chemotherapy. *Exp. Parasitol.* **186**, 17–23 (2018).

[241] Cosentino, R. & Agüero, F. Genetic profiling of genes from the isoprenoid and sterol biosynthesis pathways of Trypanosoma cruzi. (2014). URL http://dx.doi.org/10.7287/peerj.preprints.44v1. Submitted.

[242] Dixon, H., Ginger, C. D. & Williamson, J. Trypanosome sterols and their metabolic origins. *Comp. Biochem. Physiol. B* **41**, 1–18 (1972).

[243] Kessler, R. L., Soares, M. J., Probst, C. M. & Krieger, M. A. Trypanosoma cruzi response to sterol biosynthesis inhibitors: morphophysiological alterations leading to cell death. *PLoS One* **8**, e55497 (2013).

[244] de Macedo-Silva, S. T. *et al.* Benzylamines as highly potent inhibitors of the sterol biosynthesis pathway in leishmania amazonensis leading to oxidative stress and ultrastructural alterations. *Sci. Rep.* **12**, 11313 (2022).

[245] Choi, J. Y., Podust, L. M. & Roush, W. R. Drug strategies targeting CYP51 in neglected tropical diseases. *Chem. Rev.* **114**, 11242–11271 (2014).

[246] Chen, C.-K. *et al.* Structural characterization of CYP51 from trypanosoma cruzi and trypanosoma brucei bound to the antifungal drugs posaconazole and fluconazole. *PLoS Negl. Trop. Dis.* **4**, e651 (2010).

[247] Chen, C.-K. *et al.* Trypanosoma cruzi CYP51 inhibitor derived from a mycobacterium tuberculosis screen hit. *PLoS Negl. Trop. Dis.* **3**, e372 (2009).

[248] Torrico, F. *et al.* New regimens of benznidazole monotherapy and in combination with fosravuconazole for treatment of chagas disease (BENDITA): a phase 2, double-blind, randomised trial. *Lancet Infect. Dis.* **21**, 1129–1140 (2021).

[249] Naula, C., Parsons, M. & Mottram, J. C. Protein kinases as drug targets in trypanosomes and leishmania. *Biochim. Biophys. Acta* **1754**, 151–159 (2005).

[250] Diaz-Gonzalez, R. *et al.* The susceptibility of trypanosomatid pathogens to PI3/mTOR kinase inhibitors affords a new opportunity for drug repurposing. *PLoS Negl. Trop. Dis.* **5**, e1297 (2011).

[251] Phan, T.-N. *et al.* In vitro and in vivo activity of mTOR kinase and PI3K inhibitors against leishmania donovani and trypanosoma brucei. *Molecules* **25**, 1980 (2020).

[252] Salazar, R. *et al.* Phase II study of BEZ235 versus everolimus in patients with mammalian target of rapamycin inhibitor-naïve advanced pancreatic neuroendocrine tumors. *Oncologist* **23**, 766–e90 (2018).

[253] Grabner, G. F., Zimmermann, R., Schicho, R. & Taschler, U. Monoglyceride lipase as a drug target: At the crossroads of arachidonic acid metabolism and endocannabinoid signaling. *Pharmacol. Ther.* **175**, 35–46 (2017).

[254] Machado, F. S., Mukherjee, S., Weiss, L. M., Tanowitz, H. B. & Ashton, A. W. Bioactive lipids in trypanosoma cruzi infection. In *Advances in Parasitology*, Advances in parasitology, 1–31 (Elsevier, 2011).

[255] López-Muñoz, R. A. *et al.* Inflammatory and pro-resolving lipids in trypanosomatid infections: A key to understanding parasite control. *Front. Microbiol.* **9**, 1961 (2018).

[256] Chan, J. N. Y., Nislow, C. & Emili, A. Recent advances and method development for drug target identification. *Trends in Pharmacological Sciences* **31**, 82–88 (2010).

[257] Emmerich, C. H. *et al.* Improving target assessment in biomedical research: the GOT-IT recommendations. *Nature Reviews Drug Discovery* **20**, 64–81 (2021). URL https://www.nature.com/articles/s41573-020-0087-3.

[258] Schenone, M., Dančík, V., Wagner, B. K. & Clemons, P. A. Target identification and mechanism of action in chemical biology and drug discovery. *Nature Chemical Biology* **9**, 232–240 (2013).

Siglas

ADME Absorción, Distribución, Metabolsimo y Excreción. 9, 26

API Application Programming Interface. 61, 63

ARNi RNA de interferencia. 34

CSV *Comma-separated values file.* 63

DDD *Domain Driven Design.* 57

dsRNA *RNA doble cadena.* 34

DTU *Discrete Typing Unit.* 11

InChI *IUPAC International Chemical Identifier.* 23

IUPAC *International Union of Pure and Applied Chemistry.* 23

MOI *Multiplicity of infection.* 71

MVC *Model-View-Controller.* 57

NDP *Network Driven Prioritizations.* 41

NDS *Network Druggability Score.* 41, 53

PAINs *Pan-assay interference compounds.* 62, 65

PCA *Principal components analysis.* 63, 65

PDB *Protein Data Bank.* 49

SDF *Structure-Data File.* 18

SMILES *Simplified Molecular Input Line Entry System.* 19, 20, 63

TDR *Special Programme for Research and Traning in Tropical Diseases.* 1

TPP *Thermal proteome profiling.* 36

tSNE *t-distributed stochastic neighbor embedding.* 65

WHO *World Health Organization.* 1

Glosario

hit Una molécula pequeña que ha mostrado actividad de interés en un ensayo de *screening* y cuya actividad ha sido confirmada en ensayos similares reiterados. La potencia es tipicamente de 100 nM – 5 M. 105

IC50 Concentración Inhibitoria 50 es la concentración de una droga o inhibidor necesaria para inhibir un proceso o respuesta biológica en un 50 %. . 71

Indice de Siluetas El índice de siluetas o *Silhouette Score* es una métrica para evaluar la performance de algoritmos de agrupamiento (*clustering*). Combina una medida de compactación (distancias *intra*-cluster) y una de separación (distancias *inter*-cluster para medir un puntaje global representativo de cuán bien quedan agrupados los datos en sus respectivos clusters.. 65

lead Una molécula pequeña que ha mostrado la actividad de interés (un hit) y que además ha pasado varios filtros adicionales (selectividad, potencia, solubilidad, estudios de relación estructura-función, etc.) se convierte en un *lead* y suele avanzar hacia otros tipos de optimización. 8

target Un blanco terapéutico drogable cuya modulación química constituye una potencial vía de tratamiento. 8

Agradecimientos

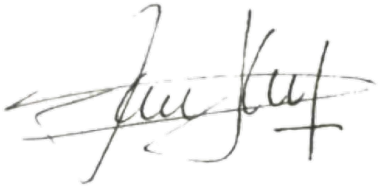
A mis amigos y compas de trabajo por enseñarme y bancarme todos los días (muy especialmente a Mechi, que puso el hombro para finalizar algunos de los ensayos de medición de actividad tripanocida). A mi director y mentor, Fernán, por dejarme crecer a mi modo sin perder la capacidad de guiarme cada vez que fue preciso. A Emir, mi compa-mentor, que supo ser amigo sin dejar de ser maestro.

A este país hermoso, gracias al cual pude estudiar todo lo que me propuse y más. Al CONICET y la Agencia Nacional de Promoción de la Investigación, el Desarrollo Tecnológico y la Innovación, que financiaron mi trabajo por 6 años; a la Universidad Pública en general, y a la UNSAM y al IIBio en particular, por ser mi segunda casa. Guardaré siempre en alta estima los años que me tocó pasar por estos lares.

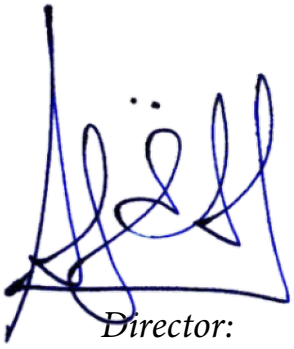
Por último, y por sobre todas las cosas, a mis padres, por el gen absolutamente dominante de la perseverancia. A mis hermanos por el amor y la protección, casi siempre invisible. A mi persona favorita en todo el mundo, Mich, por sanarme a risotadas cualquier mal y mostrarme día a día la fuerza del cariño incondicional. Son ellos el núcleo indisputable de mi existencia. Los amo con alma y vida.

Gracias, de todo corazón.

Firmas



Autor:
Lionel Uran Landaburu



Director:
Fernán Agüero

Corresponde a última versión del manuscrito. Incluye correcciones acordadas con los jurados.