Universidad Nacional Mayor de San Marcos
Universidad del Perú. Decana de América
Facultad de Farmacia y Bioquímica
Escuela Profesional de Farmacia y Bioquímica

Desarrollo de modelos QSAR para la predicción de la permeabilidad aparente en células Caco-2 de productos naturales provenientes de la biodiversidad del Perú

TESIS

Para optar el Título Profesional de Químico Farmacéutico

AUTOR

Victor Anderson ACUÑA GUZMAN

ASESOR

Christian SOLIS CALERO

Lima, Perú

2023

Referencia bibliográfica

Acuña V. Desarrollo de modelos QSAR para la predicción de la permeabilidad aparente en células Caco-2 de productos naturales provenientes de la biodiversidad del Perú [Tesis de pregrado]. Lima: Universidad Nacional Mayor de San Marcos, Facultad de Farmacia y Bioquímica, Escuela Profesional de Farmacia y Bioquímica; 2023.

Metadatos complementarios

Datos de autor	
Nombres y apellidos	Victor Anderson Acuña Guzman
Tipo de documento de identidad	DNI
Número de documento de identidad	76397073
URL de ORCID	https://orcid.org/0000-0002-1404-3680
Datos de asesor	
Nombres y apellidos	Christian Solis Calero
Tipo de documento de identidad	DNI
Número de documento de identidad	10373255
URL de ORCID	https://orcid.org/0000-0003-0016-3750
Datos del jurado	
Presidente del jurado	
Nombres y apellidos	Gabriela Norma Solano Canchaya
Tipo de documento	DNI
Número de documento de identidad	41679636
Miembro del jurado 1	
Nombres y apellidos	Oscar Herrera Calderón
Tipo de documento	DNI
Número de documento de identidad	44789288
Miembro del jurado 2	
Nombres y apellidos	Juan Roberto Pérez León Camborda
Tipo de documento	DNI
Número de documento de identidad	06050022
Datos de investigación	
Línea de investigación	B.2.5.4. Tecnologías ómicas y bioinformática aplicadas en salud

Grupo de investigación	Genética, ómicas, bioinformática y desarrollo computacional en biomedicina, farmacia, toxicología y alimentos - GENOBIDC
Agencia de financiamiento	Perú. Universidad Nacional Mayor de San Marcos. Vicerrectorado de Investigación y Posgrado. RR N° 006081-R-23 con código de proyecto A23042271.
Ubicación geográfica de la investigación	Edificio: Universidad Nacional Mayor de San Marcos País: Perú Departamento: Lima Provincia: Lima Distrito: Lima Latitud: -12.05819215 Longitud: -77.0189181894387
Año o rango de años en que se realizó la investigación	2020 - 2023
URL de disciplinas OCDE	Química medicinal https://purl.org/pe-repo/ocde/ford#3.01.06 Bioinformática https://purl.org/pe-repo/ocde/ford#1.02.03

Universidad Nacional Mayor de San Marcos
Universidad del Perú. Decana de América
Facultad de Farmacia y Bioquímica
Decanato

ACTA DE SUSTENTACIÓN DE TESIS

Los miembros del Jurado Examinador y Calificador de la Tesis titulada:

"Desarrollo de modelos QSAR para la predicción de la permeabilidad aparente en células Caco-2 de productos naturales provenientes de la biodiversidad del Perú"

Que presenta el Bachiller en Farmacia y Bioquímica:

VICTOR ANDERSON ACUÑA GUZMAN

Que reunidos en la fecha se llevó a cabo la **SUSTENTACIÓN** de la **TESIS**, y después de las respuestas satisfactorias a las preguntas y objeciones formuladas por el Jurado, ha obtenido la siguiente calificación final:

18 (DIECIOCHO) APROBADO CON MENCIÓN HONROSA

de conformidad con el Art. 14.º del Reglamento General de Grados y Títulos de la Universidad Nacional Mayor de San Marcos para la obtención del Título Profesional de Químico Farmacéutico (a) de la Facultad de Farmacia y Bioquímica.

Lima, 10 de julio de 2023.

Dra. Gabriela Norma Solano Canchaya
Presidenta

Dr. Oscar Herrera Calderón
Miembro

Mg. Juan Roberto Pérez León Camborda
Miembro

UNIVERSIDAD NACIONAL MAYOR DE SAN MARCOS
Universidad del Perú. DECANA DE AMÉRICA.
FACULTAD DE FARMACIA Y BIOQUÍMICA
ESCUELA PROFESIONAL DE FARMACIA Y BIOQUÍMICA

INFORME DE EVALUACIÓN DE CRITERIOS DE ORIGINALIDAD

1	Facultad	**FARMACIA Y BIOQUÍMICA**
2	Escuela	**FARMACIA Y BIOQUÍMICA**
3	Autoridad que emite el informe de originalidad	**Director de la Escuela Profesional**
4	Apellidos y nombres de la autoridad académica	**Luis Miguel V. Felix Veliz**
5	Operador del programa informático de similitudes	**Luis Miguel V. Felix Veliz**
6	Documento evaluado	Tesis para optar al título profesional de Químico Farmacéutico: "Desarrollo de modelos QSAR para la predicción de la permeabilidad aparente en células Caco-2 de productos naturales provenientes de la biodiversidad del Perú"
7	Autor(es) del documento	Br. Acuña Guzman, Victor Anderson
8	Fecha de recepción del documento	**20/06/2023**
9	Fecha de aplicación del programa informático de similitudes	**20/06/2023**
10	Software utilizado	**Turnitin**
11	Configuración del programa detector de similitudes	**Excluye:** - **Textos entrecomillados** - **Bibliografía** - **Cadenas menores de 40 palabras**
12	Porcentaje de similitud según programa detector de similitudes	**8 % (El % de similitud debe ser ≤ 10%)**
13	Fuentes originales de las similitudes encontradas	• **Fuentes de internet varias 8 %** • **Publicaciones 6 %** • **Trabajo de estudiantes entregados a otras universidades 3 %**
14	Observaciones	Realizar la edición final de la tesis. Procede la sustentación.
15	Calificación de originalidad	Documento cumple con los criterios de originalidad.
16	Fecha del informe	**20/06/2023**

Nota: se adjunta archivo de reporte del sistema Turnitin en el que se resaltan las similitudes detectadas.

UNMSM
Firmado digitalmente por FELIX VELIZ Luis Miguel Visitacion FAU 20148092282 soft
Motivo: Soy el autor del documento
Fecha: 20.06.2023 21:03:02 -05:00

Dr. Luis Miguel V. Felix Veliz

DEDICATORIA

Dedicado a mis padres Victor y Zoraida, mis tíos Jacob, Artemisa y Linda, mi abuelito Pedro, mi abuelita Modesta y Aida, quienes me apoyaron y motivaron a que siempre alcance mis metas y me esfuerce por conseguir una carrera universitaria.

AGRADECIMIENTOS

Agradezco a Dios por permitirme desarrollar este trabajo y brindarme la fortaleza y perseverancia para desarrollarlo y concluirlo.

A mis padres, tíos, abuelos y amigos, por acompañarme en estos años que desarrollé este trabajo de investigación.

A la Escuela de Farmacia y Bioquímica y al Jurado Evaluador por la revisión del presente trabajo de investigación.

A mi asesor, Dr. Christian Solís Calero, por su asesoría y guía durante el desarrollo de este trabajo de investigación.

Al Instituto de Investigaciones de la Amazonía Peruana (IIAP) del Centro de Alto Rendimiento Computacional, por el acceso al supercomputador MANATI para la ejecución de los experimentos computacionales.

Esta investigación fue financiada por la Universidad Nacional Mayor de San Marcos – RR N° 006081-R-23 con código de proyecto A23042271.

ÍNDICE

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

ÍNDICE DE ANEXOS

ABREVIATURAS

QSAR **:** Relación cuantitativa estructura-actividad

Caco-2 **:** Línea celular de adenocarcinoma de colon humano

SMILES **:** Simplified Molecular Input Line Entry Specification

logPeff **:** Permeabilidad efectiva

Papp **:** Permeabilidad aparente

logPapp : Logaritmo de la permeabilidad aparente

MLR **:** Regresión lineal múltiple (*multiple linear regression*)

PLS **:** Mínimos cuadrados parciales (*partial least squares*)

Lasso **:** Least absolute shrinkage and selection operator

SVM **:** Support Vector Machine

RF **:** Random Forest

GBM **:** Gradient Boosting Machine

SVM-RF-GBM: Modelo de conjunto por apilamiento SVM, RF y GBM

CV5 **:** Validación cruzada de 5 iteraciones (*five-fold cross validation*)

PCA **:** Análisis de componentes principales (*principal component analysis*)

RMSE **:** Raíz del error cuadrático medio (*root-mean-square error*)

R^2 **:** Coeficiente de determinación

OCDE **:** Organización para la Cooperación y el Desarrollo Económicos

RFE-RF **:** Eliminación recursiva de características con random forest

RESUMEN

El Perú es un país biodiverso que posee recursos naturales que han sido empleados para tratar diferentes enfermedades. Sin embargo, se desconocen las propiedades de los compuestos presentes en estos recursos, como la seguridad y eficacia. La línea celular Caco-2 es un modelo *in vitro* utilizado para estimar la absorción intestinal humana mediante la medición de la permeabilidad aparente. El modelamiento QSAR es un enfoque alternativo para predecir la permeabilidad aparente. En este estudio, se desarrollaron modelos QSAR para predecir la permeabilidad aparente en células Caco-2 de productos naturales provenientes de la biodiversidad del Perú. Se desarrollaron 6 modelos QSAR para predecir la permeabilidad aparente en células Caco-2, donde el modelo SVM-RF-GBM mostró la mejor capacidad predictiva (RMSE = 0.38, R^2 = 0.76). Se generó una base de datos de 516 productos naturales de la biodiversidad del Perú clasificados en 6 vías metabólicas que fueron obtenidos de 59 especies. Finalmente, se predijo la permeabilidad aparente de 516 productos naturales, y se realizaron análisis posteriores a 502 compuestos que se encontraban dentro del dominio de aplicabilidad. Se encontró que 338 compuestos presentan una alta permeabilidad aparente y, en consecuencia, una alta absorción intestinal.

Palabras clave: absorción, descubrimiento de fármacos, recursos naturales

ABSTRACT

Peru is a biodiverse country with natural resources for treating different diseases. However, the properties of the compounds present in these resources, such as safety and efficacy, are unknown. The Caco-2 cell line is an *in vitro* model used to estimate human intestinal absorption by measuring apparent permeability. QSAR modelling is an alternative approach to predicting apparent permeability. In this study, we developed QSAR models to predict the apparent permeability in Caco-2 cells of natural products from the biodiversity of Peru. We developed 6 QSAR models to predict the apparent permeability in Caco-2 cells, generated a natural products database of the biodiversity of Peru and predicted the apparent permeability in Caco-2 using the natural products database. We developed 6 QSAR models to predict apparent permeability in Caco-2 cells, where the SVM-RF-GBM model showed the best predictive ability (RMSE = 0.38, R^2 = 0.76). We generated a database of 516 natural products of the biodiversity of Peru classified into six metabolic pathways obtained from 59 species. Finally, we predicted the apparent permeability of 516 natural products, of which 502 compounds were within the applicability domain. We found that 338 compounds present a high apparent permeability and, consequently, a high intestinal absorption.

Keywords: absorption, drug discovery, natural resources

I. INTRODUCCIÓN

Los fármacos son moléculas que interactúan con un sistema biológico para producir una respuesta, y se emplean en el diagnóstico, tratamiento o prevención de una enfermedad.[1,2] La vía de administración preferida es la vía oral debido a un mayor cumplimiento por el paciente, menor costo de manufactura y fácil administración. Asimismo, representan alrededor del 60 % de productos farmacéuticos de moléculas pequeñas que se comercializan.[3] Sin embargo, cuando se ingiere un compuesto por vía oral, debe pasar por la boca, estómago y llegar al intestino delgado, donde debe atravesar las células que recubren la pared intestinal, luego llegar a los vasos sanguíneos y distribuirse mediante el suministro de sangre alrededor del cuerpo a su objetivo final.[2] Por ello, se han estudiados diferentes factores que influyen en la absorción oral de fármacos, como la solubilidad, la permeabilidad y la estabilidad en el entorno del tracto gastrointestinal.[3]

La historia del desarrollo de fármacos orales proviene desde tiempos inmemoriales en la que se empleaban hierbas para tratar enfermedades.[4] En las últimas décadas, el uso de productos naturales como medicamentos y suplementos ha aumentado.[5] Asimismo, el Perú es uno de los países más biodiversos del mundo[6], donde la riqueza de esta biodiversidad se ve reflejada en el uso de recursos naturales empleados por la medicina tradicional desde las culturas precolombinas hasta la actualidad[6,7], para el tratamiento de diferentes condiciones como problemas respiratorios, infecciones urinarias o enfermedades gastrointestinales.[8–10] No obstante, se desconocen las propiedades de los compuestos que componen estos recursos naturales, así como su seguridad y eficacia.[11,12]

La eficacia de un compuesto está influenciada por la absorción intestinal, que se evalúa mediante la perfusión en voluntarios humanos sanos; sin embargo, este parámetro no se puede determinar hasta las últimas fases de los estudios preclínicos; por ello, se han estudiado líneas celulares epiteliales para sustituir estos datos.[13,14] La línea celular de adenocarcinoma colónico humano (Caco-2) es uno de los modelos más usados para reemplazar la determinación *in vivo* del

intestino debido a la similitud morfológica y funcional con los enterocitos humanos, a las uniones estrechas entre células adyacentes y los niveles de expresión de sus enzimas[13]; sin embargo, es costoso y requiere tiempo.

Los modelos de relación cuantitativa estructura-actividad (QSAR) permiten predecir la permeabilidad intestinal *in vitro*, correlacionando los valores experimentales con descriptores moleculares calculados[15–18] y presentan ventajas respecto a otros métodos como su alta eficiencia y bajo costo.[19]

Por lo tanto, este estudio se plantea desarrollar modelos QSAR para la predicción de la permeabilidad aparente en células Caco-2 de productos naturales provenientes de la biodiversidad del Perú que ayudará a obtener un mejor conocimiento de la permeabilidad aparente e inferir la posible absorción intestinal de los productos naturales.

I.1 Planteamiento del problema

I.1.1 Determinación del problema

El epitelio intestinal es una barrera conformada por una capa de células que recubren la luz, impide el paso de microorganismos, toxinas y antígenos extraños, y actúa como un filtro selectivo para el paso de nutrientes, electrolitos y agua desde el lumen intestinal a la circulación.[20] La capacidad de los fármacos de atravesar membranas es esencial para muchos procesos farmacocinéticos y farmacodinámicos.[21]

En el mundo, ha aumentado el uso de productos naturales como medicamentos y suplementos en las últimas tres décadas.[5] El Perú es uno de los 12 países más biodiversos del mundo.[6] La riqueza de esta diversidad se refleja en el uso de productos naturales que han sido empleados por la medicina tradicional desde las culturas precolombinas hasta la actualidad,[6,7] para el tratamiento de diferentes condiciones como problemas respiratorios, infecciones urinarias o enfermedades gastrointestinales.[8–10] Sin embargo, se desconocen las propiedades de los productos naturales, como su seguridad y eficacia[11,12]; por ello, en el presente

estudio, planteamos desarrollar modelos QSAR para predecir la permeabilidad aparente en células Caco-2 para poder obtener conocimiento de las propiedades de permeabilidad intestinal de estos productos, lo cual permitirá inferir sus propiedades farmacocinéticas y farmacodinámicas, como la absorción y biodisponibilidad potencial, que proporcionarán indicios sobre los compuestos responsables de la actividad farmacológica en estos productos naturales.

I.1.2 Formulación del problema

¿Cuál es la permeabilidad aparente en células Caco-2 de productos naturales provenientes de la biodiversidad del Perú, según los modelos QSAR desarrollados?

I.2 Objetivos

I.2.1 Objetivo general

Desarrollar modelos QSAR para la predicción de la permeabilidad aparente en células Caco-2 de productos naturales provenientes de la biodiversidad del Perú

I.2.2 Objetivos específicos

- Desarrollar modelos QSAR para predecir la permeabilidad aparente en células Caco-2
- Generar la base de datos de productos naturales de la biodiversidad del Perú
- Predecir la permeabilidad aparente en células Caco-2 usando la base de datos de productos naturales mediante los modelos QSAR

I.3 Importancia y alcance de la investigación

La absorción es un criterio importante en la selección de un fármaco durante el descubrimiento y desarrollo[22]; asimismo, la permeabilidad intestinal presenta una correlación con la absorción intestinal humana[19]; en consecuencia, existe la necesidad de métodos de detección fiables para evaluar la permeabilidad intestinal.[22] La línea celular Caco-2 se ha empleada en estudios sobre la función de la barrera del epitelio intestinal, vías para el transporte de fármacos o componentes

alimentarios, efectos tóxicos en la mucosa intestinal y la difusión pasiva a través del epitelio intestinal.[23]

Los modelos QSAR han sido empleados en la predicción de la permeabilidad aparente de Caco-2 debido a su alta eficiencia y bajo costo[19]. El Perú es uno de los países con una gran biodiversidad que representa una ventaja competitiva, y un importante activo de ingresos económicos y de exportación. Asimismo, en los últimos años, las tendencias del consumo se han orientado hacia productos naturales, orgánicos, y con cualidades nutritivas y nutracéuticas especiales[24]; por ello, el conocimiento químico sobre los productos naturales es necesario para otorgarles un mayor valor agregado. En el Perú, se han realizado pocos estudios QSAR,[25,26] los cuales han empleado un bajo número de moléculas para la construcción del modelo y, en consecuencia, un limitado número de moléculas que se puedan predecir, una baja rigurosidad en la validación y no se han aplicado en la predicción de actividades biológicas como la permeabilidad aparente en Caco-2.

Los recursos naturales han sido empleados por la medicina tradicional desde hace miles de años en el tratamiento y prevención de enfermedades. En el Perú, la riqueza de su biodiversidad se refleja en los diferentes usos tradicionales que se atribuyen a los productos naturales, como en el tratamiento de problemas respiratorios, infecciones parasitarias, infecciones urinarias, problemas gastrointestinales y problemas de la piel[6,27,28]; sin embargo, existe una falta de información sobre la composición de sus extractos, sus propiedades farmacológicas y su seguridad.[11,12]

I.4 Limitaciones de la investigación

Las limitaciones de este estudio están relacionadas con la cantidad de compuestos disponibles en la literatura con permeabilidad aparente reportada experimentalmente. Estos compuestos son empleados para construir los modelos QSAR e influyen en su capacidad predictiva.

II. REVISIÓN DE LA LITERATURA

II.1 Marco teórico

II.1.1 Permeabilidad intestinal

La permeabilidad intestinal es una propiedad que consiste en la facilidad con que el epitelio intestinal permite que las moléculas pasen por difusión pasiva no mediada.[29]

II.1.1.1 Tracto gastro intestinal

El tracto gastrointestinal es un canal cubierto de epitelio que se extiende desde la boca hasta el ano, y se encarga de la digestión de alimentos, absorción de nutrientes, expresión de productos de desecho y reabsorción de agua. La región absorbente, que incluye el estómago, el intestino delgado y parte del intestino grueso, está compuesta de una capa de células epiteliales denominada enterocitos.[29]

Los enterocitos son la población más grande de células intestinales, las cuales presentan una membrana apical hacia la luz gastrointestinal y una membrana basolateral hacia el lado seroso. Estas membranas presentan variaciones en su composición de fosfolípidos y expresión de proteínas.[29,30] Asimismo, presentan pliegues, vellosidades (proyecciones en forma de dedos) y microvellosidades (en la membrana celular apical de los enterocitos) que aumenta el área de superficie de la mucosa y mejora su capacidad de absorción.[29]

Las células secretoras del intestino delgado están conformadas por las células caliciformes, células enteroendocrinas y células de Paneth. Las células caliciformes son el segundo grupo más abundante en el intestino y se encargan de la producción y secreción de mucina, un componente de la capa de moco. Las células de Paneth secretan proteínas antimicrobianas y defensinas, involucradas en el sistema inmune innato, y las células enteroendocrinas producen y secretan hormonas intestinales en respuesta a estímulos como la absorción de nutrientes y la composición del medio luminal (Figura 1).[30]

Las células M se encuentran en las placas de Peyer, regiones especializadas en el intestino sin protección de la capa de moco, y exhiben una baja actividad de aminopeptidasa. El alto potencial endocítico de las células M permite el transporte de macromoléculas, antígenos y microorganismos. Además, estas células también son cruciales para iniciar la respuesta inmunitaria de la mucosa.[30]

La barrera intestinal separa la luz intestinal del huésped interno, permite el intercambio de moléculas y la absorción de nutrientes, previene la pérdida de agua, electrolitos, y la entrada de antígenos y microorganismos en el cuerpo.[21,31]. Asimismo, está compuesta de elementos mecánicos (moco, capa epitelial), elementos humorales (defensinas, IgA), elementos inmunológicos (linfocitos, células inmunitarias innatas), elementos musculares y neurológicos.[21]

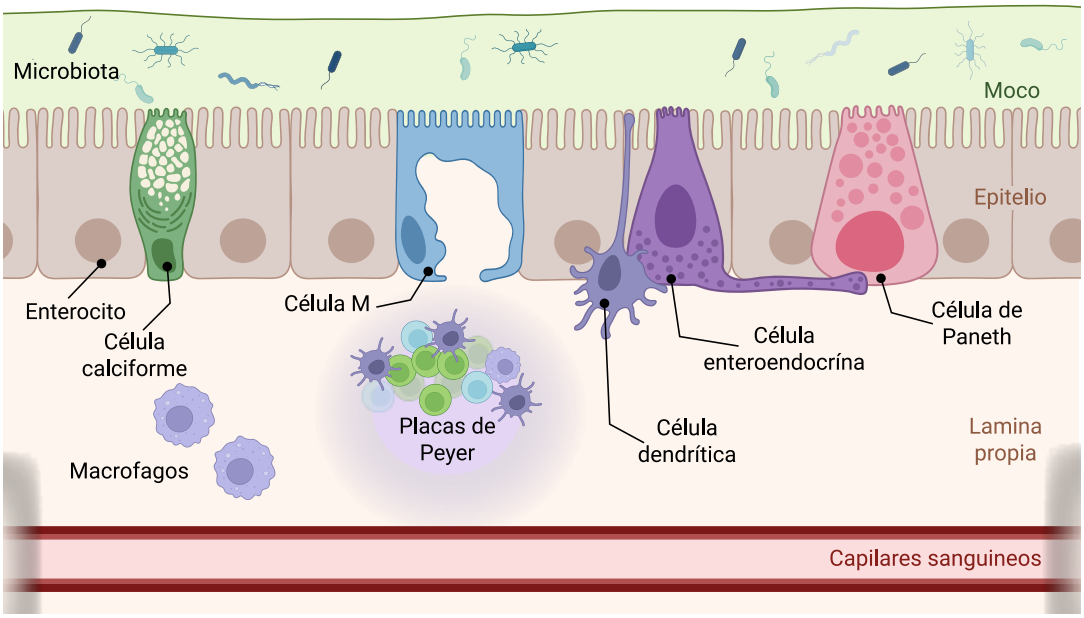


Figura 1. Monocapa epitelial del intestino. Adaptado de Xu et al.[30] Creado con BioRender.com

II.1.1.2 Absorción intestinal

La absorción, permeabilidad y biodisponibilidad son propiedades que están estrechamente relacionadas.[32] La absorción oral es el porcentaje de fármaco absorbido desde la luz gastrointestinal hacia la sangre de la vena porta, donde

están involucrados procesos fisicoquímicos y biológicos (transportadores, enzimas metabolizadoras).[32]

II.1.1.3 Mecanismos de transporte

Los compuestos, tanto nutrientes como xenobióticos, tienen que atravesar la barrera intestinal para llegar a la circulación sanguínea.[33] El transporte de estos compuestos a través del epitelio intestinal puede darse mediante cuatro vías: (1) pasiva transcelular, (2) pasiva paracelular, (3) mediada por portadores y (4) mediada por transcitosis, como se muestra en la Figura 2.[34]

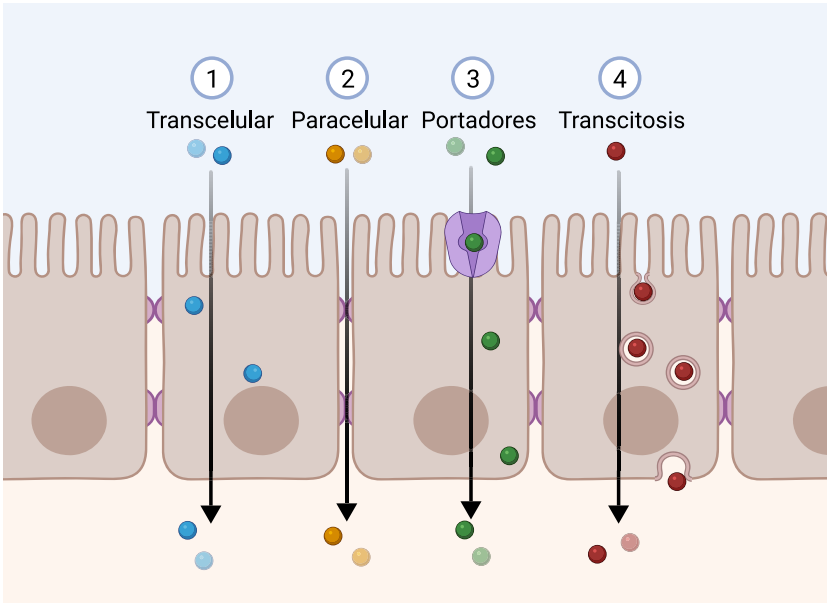


Figura 2. Transporte de fármacos por 4 vías. (1) Vía transcelular. (2) Vía paracelular. (3) Vía mediada por portadores. (4) Transcitosis. Adaptado de Artursson et al.[34] Creado con BioRender.com

El transporte pasivo se rige por la ley de difusión de Fick y se caracteriza por el movimiento de moléculas a favor de un gradiente de concentración.[32] En contraste, el transporte activo implica el desplazamiento de moléculas en contra de un gradiente de concentración, lo cual demanda un consumo energético.[32]

(a) Pasiva transcelular

El transporte transcelular es el paso de un compuesto a través del enterocito por difusión simple.[33] La mayoría de los medicamentos aprobados que se absorben rápida y completamente después de la administración oral se transportan por la ruta transcelular pasiva.[34]

(b) Pasiva paracelular

El transporte paracelular o intercelular es el paso de un compuesto a través de las uniones estrechas de los enterocitos.[32,33]

(c) Mediada por portadores

El transporte mediado por portadores puede ser facilitado o activo. El transporte facilitado emplea portadores que permiten el paso del soluto como la glucosa, urea y aminoácidos, a favor del gradiente electroquímico y sin gasto de energía. El transporte activa crea gradientes de iones o solutos a través de las membranas, que implica un gasto de energía.[35]

(d) Transcitosis

El transporte mediado por transcitosis emplea vesículas de membrana para transportar compuestos desde la mucosa hasta la serosa del epitelio intestinal. Esta vía es la menos frecuente para el transporte de fármacos. Asimismo, las vesículas de membrana contienen enzimas proteolíticas encargadas de degradar las proteínas exógenas. La vitamina B12 emplea esta ruta para atravesar los enterocitos desde la mucosa a la serosa.[34]

II.1.1.4 Factores fisicoquímicos en la permeabilidad y absorción

Los mecanismos de transporte de los compuestos dependen de sus propiedades fisicoquímicas, como el tamaño, estructura química y equilibrio hidrofílico-lipofílico.[30]

(a) Peso Molecular (MW)

Es un factor limitante en la absorción oral y debe ser menor a 500 según la regla 5 de Lipinski.[32]

(b) Enlaces de Hidrogeno

La capacidad de formar enlaces de hidrógeno de un compuesto se correlaciona con su difusión pasiva. El número de enlaces de hidrógeno y el área de superficie polar (PSA) son descriptores de enlaces de hidrogeno. El PSA es la suma de las contribuciones fraccionales al área de superficie de todos los átomos de nitrógeno y oxígeno, y átomos de hidrógeno; por ello, los compuestos con baja absorción presentan un PSA mayor a 140 Å.[32]

(c) Lipofilicidad

Es la afinidad de una molécula o un residuo por un entorno lipofílico.[32] Las moléculas hidrófobas pueden atravesar la bicapa lipídica de los enterocitos y las moléculas hidrófilas de bajo peso molecular, por transporte paracelular.[30] Se emplean los coeficientes de partición octanol/agua (logP) y distribución para estimar la penetración y permeabilidad de la membrana incluyendo la absorción intestinal.[32] Los fármacos que se absorben rápida y completamente son generalmente lipofílicos y se distribuyen fácilmente en las membranas celulares del epitelio intestinal, debido a que el área superficial de las membranas en cepillo es 1000 veces mayor que el área de superficie paracelular.[34] Los fármacos que se absorben de forma pasiva lenta e incompleta, como los fármacos hidrófilos y los péptidos, se distribuyen mal en las membranas celulares y pueden ser transportados por la vía paracelular.[34]

II.1.1.5 Tipos de Permeabilidad

(a) Permeabilidad efectiva

La permeabilidad efectiva (Peff) o tasa de permeabilidad yeyunal humana es la permeabilidad determinada a partir de estudios de perfusión intestinal en humanos.

[36] Se ha encontrado una correlación entre la Peff y la fracción absorbida del fármaco, la cual se obtuvo de estudios farmacocinéticos o de balance de masas. Se ha empleado ampliamente para determinar la velocidad y el grado de absorción intestinal de un fármaco en humanos.[21] Sin embargo, debido a la complejidad y altos costos de cada procedimiento, existen pocos estudios de permeabilidad yeyunal en humanos disponibles, con alrededor de 30 fármacos.[36]

(b) Permeabilidad aparente

La permeabilidad intestinal aparente (Papp) es la cantidad de compuesto transportado por tiempo. Es medida *in vitro* en líneas celulares[36] y se calcula mediante la siguiente expresión:[14]

$$P_{app} = \frac{dQ\ /\ dt}{C_0 \times A} \qquad (1)$$

Donde Papp es medida en cm s^{-1}; dQ/dt es la velocidad de permeación del fármaco a través de las células; es decir, es la cantidad de sustancia que se transfiere a través de la membrana por unidad de tiempo (µmol s^{-1}); C_0 es la concentración inicial del compartimento del donador (µM); y A, el área de la monocapa celular (cm^2).[14]

II.1.2 Línea celular de adenocarcinoma de colon humano (Caco-2)

La línea celular Caco-2 es un modelo *in vitro* empleado para estudios de absorción, transporte y biodisponibilidad.[37]. Caco-2 se aisló de un adenocarcinoma colorrectal humano y fue establecida por Fogh y Trempe en 1974.[37]

Caco-2 diferenciada expresa propiedades morfológicas y funcionales características del intestino delgado, como un borde en cepillo apical con microvellosidades, formación de uniones estrechas entre células adyacentes y expresión de enzimas como lactasa, aminopeptidasa N, sacarasa-isomaltasa y dipeptidil peptidasa IV, las cuales son características de los enterocitos (Figura 3)[23,37]. Estas características lo convierten en uno de los métodos más usados para reemplazar la determinación *in vivo* del intestino.[13]

Asimismo, Caco-2 presenta algunas limitaciones, como una monocapa heterogénea vinculada al tiempo de cultivo, presencia de zonas multicapa e incapacidad de producir moco debido a la ausencia de células calciformes.[37] A pesar de ello, la línea Caco-2 es empleada en la predicción *in vitro* de permeabilidad y absorción intestinal debido a la correlación entre la permeabilidad aparente *in vitro* y la absorción en humanos. En la Figura 4, se muestra la correlación entre la fracción absorbida (FA) en humanos y su permeabilidad en la línea celular Caco-2, en un grupo fármacos orales transportados pasivamente.[14]

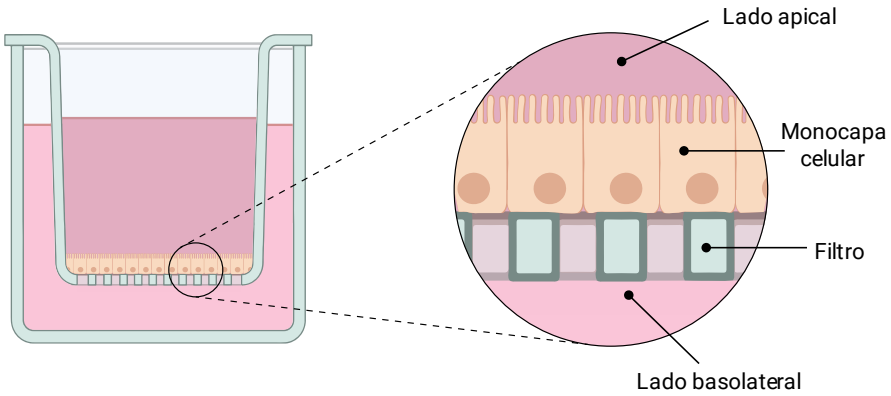


Figura 3. Diagrama de una monocapa Caco-2. Adaptado de Hubatsch et al.[14] Creado con BioRender.com

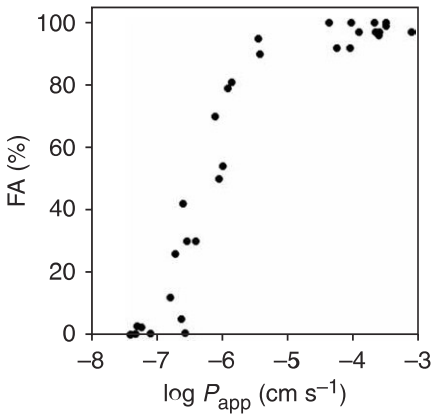


Figura 4. Correlación entre la fracción absorbida y la permeabilidad en Caco-2[14]

II.1.3 Estudio de relación cuantitativa estructura-actividad (QSAR)

La relación cuantitativa estructura-actividad (QSAR) es un modelo matemático que correlaciona la actividad biológica de compuestos y sus características estructurales.[38] El objetivo de un modelo QSAR es predecir la actividad biológica o respuesta (*Y*) en función de los predictores (*X*), que puede ser ser expresado como:

$$\text{Actividad biológica} = f(características\ estructurales) \qquad (2)$$

La respuesta o actividad biológica puede ser pIC50, pEC50, Ki, logBB o logPapp.[39]

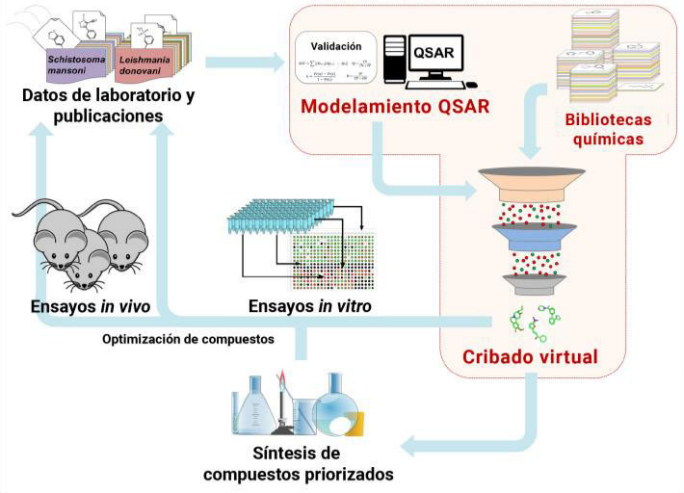


Figura 5. Flujo de trabajo en el desarrollo de modelos QSAR.[39]

El empleo de modelos QSAR presenta diferentes ventajas como la reducción del tiempo y el costo, y la predicción racional de las actividades/propiedades biológicas, físicas y químicas.[40]

II.1.3.1 Descriptores moleculares

Son parámetros fisicoquímicos en términos numéricos que describen a las moléculas. Se determinan de forma experimental o computacional y se pueden emplear en el desarrollo de modelos QSAR.[38]

Se pueden clasificar en descriptores constitucionales, topológicos, geométricos, termodinámicos y electrónicos, y en base a su dimensionalidad, como 0D, 1D, 2D y 3D.[41]

- **Constitucionales:** Son descriptores que emplean la información química de la molécula sin información de la conectividad de átomos. Son los descriptores más comunes y simples. Ej. Número de átomos, número de enlaces, peso molecular.[41]
- **Topológicos:** Están basados en gráficos moleculares, representan la conectividad de átomos en moléculas y son usados para modelar propiedades fisicoquímicas, biológicas y farmacocinéticas. Ej. Wiener, Zagreb, índices de conectividad.[41]
- **Geométricos:** Son calculados a partir de las coordenadas 3D de los átomos, emplean información el tamaño molecular, la forma y la distribución de los átomos. Ej. WHIM, MoRSE, GETAWAY.[41]
- **Termodinámicos:** Cuantifica las propiedades termodinámicas de una molécula. Ej. Refractividad molar, AlogP.[41]
- **Electrónicos:** Describen aspectos electrónicos de los enlaces de moléculas. Ej. Momento dipolo, HOMO, LUMO. [41]

II.1.3.2 División del conjunto de datos

Los modelos QSAR deben ajustarse a los datos empleados en su construcción y predecir con precisión nuevos resultados. Para evaluar la capacidad de generalización del modelo, es necesario dividir el conjunto de datos en conjuntos de entrenamiento y prueba.[42]

(a) Conjunto de entrenamiento (Training set):

Son los datos que se emplean para construir o "entrenar" el modelo. En este conjunto de datos, se entrenan diferentes algoritmos, se ajustan hiperparámetros y comparan diversos modelos para seleccionar el modelo final.[42]

(b) Conjunto de prueba (Testing set):

Son los datos empleados para estimar de manera no sesgada el rendimiento del modelo y comprobar su capacidad de generalización.[42]

II.1.3.3 Selección de características

Los métodos de selección de características tienen como objetivo eliminar los descriptores redundantes, ruidosos o irrelevantes para la construcción de modelos QSAR;[43] asimismo, presentan las siguientes ventajas:

- Mejora de la precisión del modelo al eliminar descriptores irrelevantes o redundantes.
- Simplificación y mayor facilidad de interpretación de un modelo al utilizar un conjunto reducido de descriptores.
- Reducción del tiempo y complejidad en los métodos para construir el modelo.

II.1.3.4 Modelos de regresión

(a) Regresión Lineal Múltiple (MLR)

La forma más simple de regresión lineal se llama regresión lineal simple (SLR) [38]; la cual se basa en la siguiente ecuación:

$$\mathrm{Y} = \mathrm{aX} + \mathrm{b} \qquad (3)$$

Donde Y es la variable dependiente y X, la variable independiente; y los valores constantes a y b son la pendiente e intersección, respectivamente.[38]

En la práctica, se puede emplear más de un predictor; es decir, podemos extender el modelo SLR para múltiples predictores; el cual se conoce como modelo de regresión lineal múltiple (MLR).[42] Los modelos de MLR se expresan de la siguiente forma:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \cdots + \beta_n X_n \qquad (4)$$

Donde β_0 es la constante del modelo; Y, la variable dependiente; X_1, X_2,…, X_n son las variables independientes con sus respectivos coeficientes β_1, β_2,…, β_n. [38]

(b) Regresión de Mínimos Cuadrados Parciales (PLS)

El modelo de mínimos cuadrados parciales (PLS) transforma las variables independientes originales (X_1,…, X_m) en variables latentes (LVs) (t_1,…, t_n), las cuales son combinaciones lineales de variables independientes.[38] La ecuación general para el modelo PLS se expresa de la siguiente forma:

$$Y = a_1 t_1 + a_2 t_2 + \cdots + a_n t_n \quad (5)$$

Las LVs se describen como:

$$t_1 = b_{11} X_1 + b_{12} X_2 + \cdots + b_{1m} X_m \quad (6)$$

$$t_2 = b_{21} X_1 + b_{22} X_2 + \cdots + b_{2m} X_m \quad (7)$$

$$\vdots$$

$$t_n = b_{n1} X_1 + b_{n2} X_2 + \cdots + b_{nm} X_m \quad (8)$$

La aplicación de PLS se vuelve esencial cuando se utiliza un gran número de descriptores multicolineales para análisis QSAR. Las LV no solo expresan las variaciones de los descriptores moleculares, sino que también modelan las actividades biológicas (Y) al mismo tiempo.[38]

(c) Support Vector Machine (SVM)

Es un algoritmo que se emplea para resolver problemas de regresión y clasificación. En la regresión, los datos de entrada son transformados mediante una función llamada kernel, a un espacio de características; luego, encuentra un hiperplano que se ajuste al espacio de características (Figura 6). Algunas funciones kernel son lineal, polinomial y radial o *radial basis function*.[42]

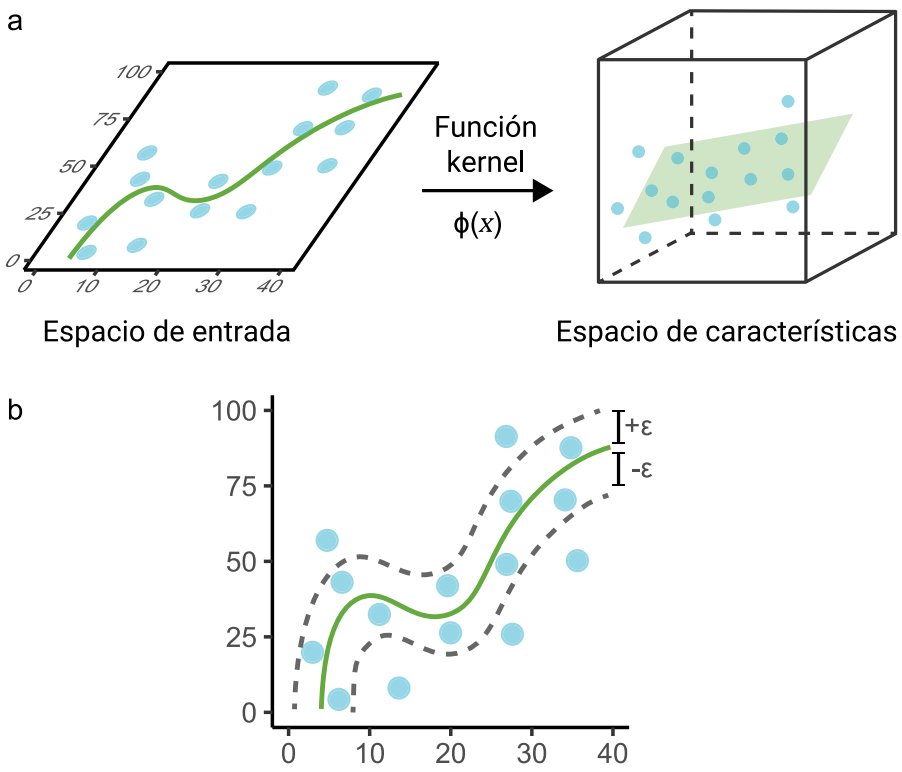


Figura 6. Esquema de un Support Vector Machine. A. Empleo de la función kernel. B. Gráfico de un SVM. ε: parámetro épsilon. línea o superficie verde: hiperplano.[42,44]

(d) Random Forest (RF)

Arboles de decisión

Los árboles de decisión o *decision tree* son un algoritmo no paramétrico que divide los datos en regiones más pequeñas mediante un conjunto de reglas de división. Estos modelos generan reglas simples que son fáciles de interpretar y visualizar con diagramas de árbol (Figura 7); sin embargo, generalmente carecen de rendimiento predictivo en comparación con algoritmos más complejos.[42] Se pueden clasificar según su variable respuesta (cuantitativa o cualitativa) en arboles de regresión o clasificación, respectivamente.

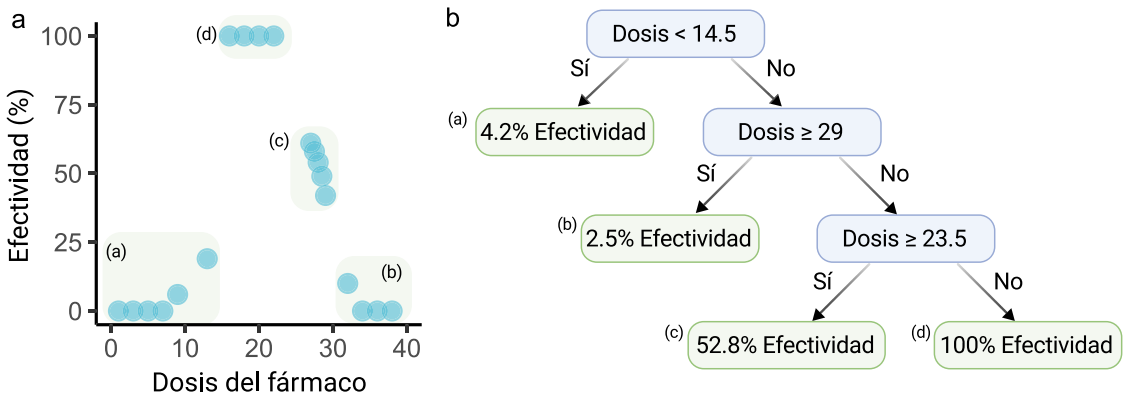


Figura 7. Esquema de un árbol de decisión. A) Gráfico dosis del fármaco vs Efectividad del fármaco (%). B) Árbol de regresión.[45]

Random forest

Los bosques aleatorios o *random forest* son un conjunto de árboles de decisión, donde cada árbol es ligeramente diferente de los demás (Figura 8). El árbol de decisión tiene una buena capacidad predictiva, pero puede sobreajustarse en alguna parte de los datos lo que disminuye su capacidad predictiva. Random forest resuelve el problema del sobreajute construyendo muchos árboles donde cada uno tiene una buena capacidad predictiva y se sobreajusta de diferentes maneras; los resultados de cada árbol de decisión son promediados lo que disminuye el sobreajuste y mejora el rendimiento predictivo.[42,46]

(e) Gradient Boosting Machine

Es un método que combina varios árboles de decisión para crear un modelo con una mejor capacidad predictiva (Figura 9). Se pueden usar para resolver problemas de regresión y clasificación. A diferencia de *random forest*, este modelo emplea el aumento de gradiente que funciona construyendo árboles en serie, donde cada árbol corrige los errores del anterior. Asimismo, se caracterizan por usar árboles de decisión poco profundos lo que ayuda a que las predicciones sean más rápidas.[46]

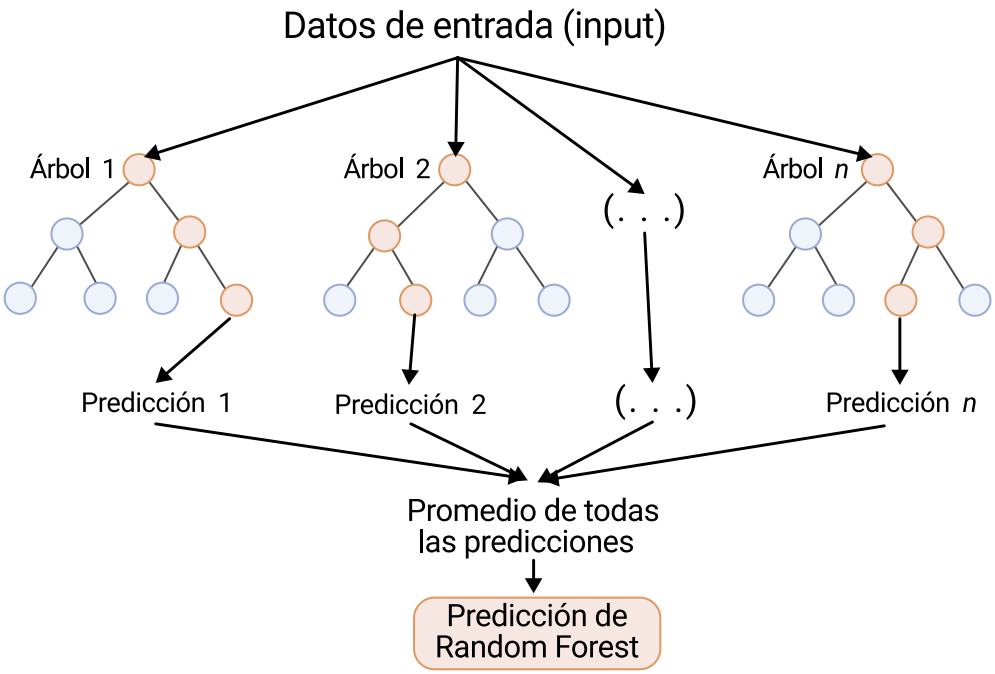


Figura 8. Esquema de un bosque aleatorio (*random forest*)

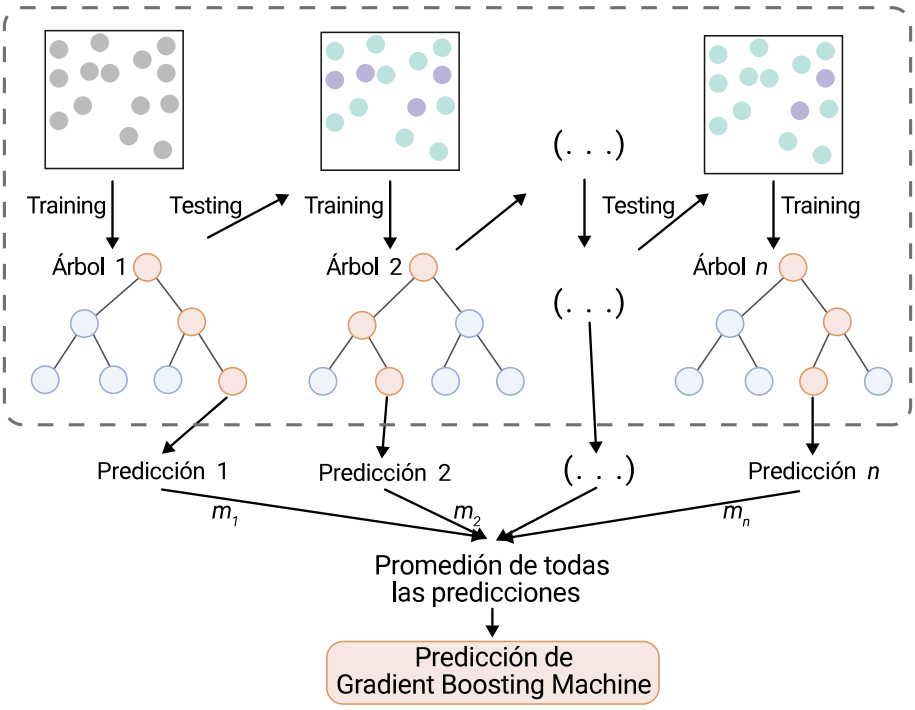


Figura 9. Esquema de un Gradient Boosting Machine. Predicción correcta (círculos celestes), predicción incorrecta (círculos morados)

(f) Métodos de Conjunto (Ensemble Methods)

Los métodos de conjunto combinan los modelos que tienen mejor desempeño para obtener un rendimiento predictivo superior al que se podría lograr con el mejor modelo individual.[42,46] La combinación de múltiples modelos en lugar de seleccionar el mejor ha sido empleado en algoritmos como *random forest* y *gradient boosting machine*, los cuales son métodos de conjunto.[42]

El apilamiento o *stacking* es un método de conjunto que reúne un grupo diverso de modelos con buena capacidad predictiva (*strong learners*) para crear un nuevo modelo o metamodelo (*meta-learner*).[42] Este método mejora el desempeño predictivo al combinar modelos con una alta variabilidad y valores predichos no correlacionados; en consecuencia, si los valores predichos son más similares, será menor la ventaja de combinar los modelos.[42] Para desarrollar un modelo por apilamiento se siguen los siguientes pasos (Figura 10): primero, se entrena individualmente cada modelo; luego, se entrena un modelo (metamodelo o meta-learner) utilizando como predictores los valores predichos por modelo individual; finalmente, se obtiene la predicción final del metamodelo.[42,46]

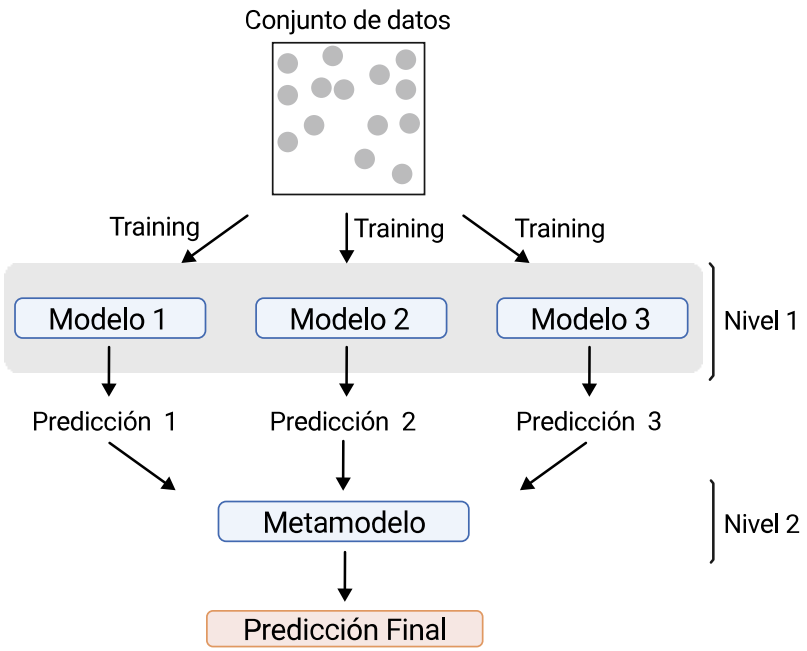


Figura 10. Esquema del método de conjunto por apilamiento o *stacking*

II.1.3.5 Validación de modelos QSAR

La validación es el proceso mediante el cual se establece la confiabilidad y relevancia de un enfoque, método, proceso o evaluación, para un propósito definido.[47]

(a) Principios de la OCDE para la validación de modelos QSAR

La Organización para la Cooperación y el Desarrollo Económico (OCDE) establece 5 principios para facilitar la consideración de un modelo QSAR con fines regulatorios[47], debe presentarse la siguiente información:

1) Una variable respuesta definida
2) Un algoritmo no ambiguo
3) Un dominio de aplicabilidad definido
4) Medidas apropiadas de bondad de ajuste, robustez y capacidad predictiva
5) Una interpretación mecanicista, si es posible

(b) Tipos de Validación

Validación Interna: Se realiza en base a las moléculas empleadas en el desarrollo del modelo. Se realiza la predicción de la actividad biológica y se estiman los parámetros. Se emplea para medir la calidad y bondad del ajuste del modelo; sin embargo, no se puede evaluar la capacidad predictiva del modelo con un nuevo conjunto de datos.[48]

Validación Externa: Se emplea el conjunto de prueba para predecir la actividad biológica y se estima los parámetros, lo cual asegura la capacidad predictiva y aplicabilidad del modelo para la predicción de nuevas moléculas.[48]

Métricas

- **Raíz del error cuadrático medio (RMSE):** Es una métrica del error, el objetivo es minimizar.[42]

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}} \quad (9)$$

Donde n es el número de observaciones, $\hat{y}_i$ son los valores predichos y y_i son los valores observados.

- **Coeficiente de determinación (R^2):** Representa la proporción de la varianza en la variable dependiente que es predecible a partir de las variables independientes.[42]

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2} \quad (10)$$

Donde $\hat{y}_i$ son los valores predichos, y_i son los valores observados y $\bar{y}$ es la media de y_i.

(c) Dominio de aplicabilidad

Un modelo QSAR debe presentar un dominio de aplicabilidad, el cual garantiza que pueda predecir los compuestos inciertos de manera razonable y precisa.[49]

Valores de apalancamiento (hat values)

Los valores de apalancamiento o *hat values* miden la distancia de un punto de datos al centro de la distribución del conjunto de entrenamiento. El apalancamiento permite evaluar si un compuesto se encuentra dentro del dominio o no, que se define de la siguiente manera.[49,50] Para una matriz del conjunto de entrenamiento con los descriptores moleculares de n filas y p columnas, $X_{n\times p}$, se calcula la matriz de sombrero o *hat matrix* (H) con la siguiente ecuación:

$$H = X'(X'X)^{-1}X \quad (11)$$

Donde X' es la matriz transpuesta y $(X'X)^{-1}$ es la matriz inversa de $(X'X)$. Los valores de apalancamiento o *hat values* se obtienen de la diagonal de la matriz H. El apalancamiento de advertencia o *warning leverage* (h^*) se calcula como $3(p+1)/n$, donde n es el número de compuestos y p el número de descriptores.[50] Un valor de apalancamiento mayor que el apalancamiento de advertencia (h^*) indica que la respuesta predicha puede no ser confiable.[50]

Este enfoque no permite calcular los valores de apalancamiento para nuevos conjuntos de datos, por ello, para un conjunto de datos nuevo y desconocido denotado como una matriz (u), se calcula la *hat matrix* mediante la siguiente ecuación:

$$h = u'(X'X)^{-1}u \qquad (12)$$

Donde X es la matriz del conjunto de entrenamiento, u es la matriz del conjunto de datos desconocido y u' es su matriz transpuesta. Finalmente, los valores de apalancamiento de la nueva muestra se obtienen de la diagonal de la matriz h.

El gráfico de Williams o *Williams plot puede* emplearse para identificar compuestos influyentes u *outliers*, que están fuera del dominio de aplicabilidad. En sus ejes, se emplean los valores de apalancamiento y los residuales estandarizados. [50]

II.1.4 Productos Naturales

Los productos naturales son compuestos orgánicos obtenidos de fuentes naturales como plantas, animales y microorganismos, que presentan actividades biológicas;[51,52] los cuales se han empleado históricamente como fuente principal de compuestos para medicamentos, cosméticos y alimentos.[53] Se pueden clasificar en metabolitos primarios y secundarios.[54]

Los metabolitos primarios son compuestos esenciales para la vida y son producidos por el metabolismo primario, mediante las vías de biosíntesis y descomposición de carbohidratos, grasas, proteínas y ácidos nucleicos.[52,54] Por otro lado, los metabolitos secundarios son compuestos que pueden proporcionar una ventaja evolutiva como resultado de la adaptación del organismo a su entorno.[51,54] Los

metabolitos secundarios son producidos a partir de intermediarios biosintéticos como la acetil coenzima A (acetil-CoA), ácido shikímico, ácido mevalónico y 1-desoxixilulosa-5-fosfato, que luego de ser modificados a través de numerosos mecanismos y reacciones, como alquilación, descarboxilación, formación de bases de Claisen y Schiff, permiten producir una diversidad de compuestos.[51,52] Los productos naturales se pueden clasificar en base a 6 vías metabólicas que los producen.[55]

II.1.4.1 Terpenoides

Los terpenoides están formados de unidades o bloques de construcción de 5 carbonos llamados isoprenos. La vía del mevalonato es la vía biosintética más conocida, que incluye al ácido mevalónico para la síntesis de terpenoides. El metabolito primario inicial es el acetil-CoA que pasa por una serie de reacciones para condensar tres moléculas de acetil-CoA en ácido mevalónico.[56] Luego de una serie de reacciones se forman compuestos como los monoterpenos, diterpeno y sesquiterpenos.[56]

II.1.4.2 Policétidos

Los policétidos son productos naturales producidos por bacterias, hongos y plantas. La biosíntesis se inicia mediante la policétido sintasa a partir de una unidad iniciadora que contine grupos alquilo, como acetato o propionato, anillos aromáticos como benzoato, o derivados de aminoácidos. Luego, la cadena se extiende con bloques de malonilo para formar un β-cetotioéster. Posteriormente, se pueden añadir restos de hidroxilo, alquenilo o alquilo.[57]

II.1.4.3 Ácidos grasos

Los ácidos grasos son ácidos carboxílicos con una larga cadena alifática saturada o insaturada, presentan de 4 a 28 átomos de carbono.[58] A partir de los ácidos grasos, las plantas producen metabolitos especializados, cuyas estructuras son generalmente ácidos grasos de cadena larga con algunos enlaces dobles o triples o un grupo funcional con oxígeno. Estas variaciones se encuentran en

determinadas especies como en aceites de semillas, lo que modifica sus propiedades y aplicaciones potenciales.[59]

II.1.4.4 Alcaloides

Son compuestos orgánicos naturales que contienen un átomo de nitrógeno, el cual le brinda propiedades alcalinas, y suele encontrarse en un grupo cíclico.[60] Se sintetizan principalmente a partir de aminoácidos como tirosina, lisina, ornitina, fenilalanina y triptófano.[61] Se han identificado alrededor de 20 000 alcaloides principalmente derivados de plantas. También se encuentran en microorganismos, organismos marinos como algas, peces globo, dinoflagelados y animales, como insectos y sapos.[61] Basándose en su estructura, los alcaloides se clasifican en indoles, quinolinas, isoquinolinas, pirrolidinas, piridinas, pirrolizidinas y tropanos.[60]

II.1.4.5 Shikimatos y fenilpropanoides

La vía del shikimato relaciona el metabolismo de los carbohidratos con la biosíntesis de aminoácidos aromáticos; asimismo, solo se encuentra en plantas y microorganismos. Esta vía empieza con el fosfoenolpiruvato y la eritrosa-4-fosfato, compuestos intermedios de la glucólisis y la ruta de la pentosa fosfato, para formar compuestos como triptófano, tirosina y fenilalanina.[62]

La vía de los fenilpropanoides comienza con la fenilalanina, que se convierte en ácido hidroxicinámico. A través de ramificaciones específicas, se forman ligninas, cumarinas, ácidos benzoicos, estilbenos y flavonoides.[63]

II.1.4.6 Aminoácidos y péptidos

Los péptidos son poliamidas formadas por la unión de α-aminoácidos a través de sus grupos carboxilo y α-amino. Estos compuestos están involucrados en el metabolismo primario y secundario. Algunos péptidos están ampliamente distribuidos en la naturaleza en diversos organismos, con ligeras variaciones, mientras que otros tienen una presencia más limitada.[64]

II.2 Antecedentes del estudio

Los estudios QSAR se basan en la asociación de una estructura molecular y una actividad biológica.[38] En 1863, se documentó el primer informe sobre una relación entre propiedades biológicas y moleculares realizado por Cros, quien observó un aumento en la toxicidad de los alcoholes con solubilidad en agua decreciente.[65] En 1868, Brown y Fraser estudiaron los efectos biológicos de los alcaloides, antes y después de la metilación de un átomo de nitrógeno básico, y observaron diferencias entre los compuestos básicos y los cuaternarios.[66] En consecuencia, platearon una relación matemática entre la estructura y la actividad fisiológica.[66] En 1893, Richet observó que la toxicidad de los éteres, aldehídos, alcoholes, cetonas y otros compuestos tienen una relación inversa con su solubilidad acuosa.[65]

El empleo de descriptores cuantitativos en modelos de correlación fue empleado por Hammett, quien introdujo la constantes electrónica sustituyente de Hammett (σ).[67] En 1962, Hansch publicó un estudio sobre la relación estructura-actividad de los reguladores del crecimiento vegetal asociadas con las constantes de Hammett y la hidrofobicidad, medida como el coeficiente de partición octanol/agua.[68]

En los últimos años, se han desarrollado modelos QSAR para predecir la permeabilidad aparente en Caco-2 con una mayor cantidad de compuestos y una variedad de descriptores moleculares, lo que ha llevado a emplear otros métodos además de la regresión lineal múltiple, como la regresión de mínimos cuadrados parciales, support vector machine, boosting, random forest y redes neuronales.[13,15–18] Asimismo, ha aumentado la rigurosidad en la validación de los modelos y en la determinación de su dominio de aplicabilidad.[19,47]

En el Perú, se han realizado pocos estudios QSAR,[25,26] empleando un número reducido de moléculas para construir el modelo, lo cual limita la capacidad de predicción de nuevos compuestos. Asimismo, se ha presentado una baja rigurosidad en la validación y no se ha abordado la predicción de actividades biológicas como la permeabilidad aparente en Caco-2.

II.3 Glosario de términos

- **Modelo**: Es una ecuación matemática que provee una aproximación de la realidad.[45]
- **Modelo predictivo**: Es un modelo que emplea variables o características para la predicción de una variable respuesta.[42]
- **Aprendizaje supervisado**: Es una técnica que emplea modelos predictivos.[42]
- **Entrenamiento, *training***: Es el proceso de ajustar los parámetros al conjunto de datos de entrenamiento para que realice predicciones precisas de datos nuevos.[42]
- **Sobreajuste u *overfitting***: Es cuando el modelo predice bien los datos del conjunto de entrenamiento, pero hace predicciones deficientes en el conjunto de prueba.[45]
- **Variable independiente**: Es denominada variable predictora, atributo, característica o predictor.[42]
- **Variable dependiente**: Es denominada variable objetivo o respuesta.[42]

III. HIPÓTESIS Y VARIABLES

III.1 Hipótesis

Los modelos QSAR desarrollados pueden predecir la permeabilidad aparente en células Caco-2 de productos naturales provenientes de la biodiversidad del Perú.

III.2 Variables

- **Independiente**: Descriptores moleculares
- **Dependiente**: Logaritmo de la permeabilidad aparente (logPapp).

III.3 Operacionalización de las variables (Anexo 1)

IV. MATERIALES Y MÉTODOS

IV.1 Área de estudio

El presente trabajo está enfocado en el desarrollo de modelos QSAR para la predicción de la permeabilidad aparente en células Caco-2 de productos naturales provenientes de la biodiversidad del Perú.

IV.2 Diseño de investigación

La investigación es de tipo aplicada debido a que se predijo la permeabilidad aparente en células Caco-2 de productos naturales del Perú para inferir su absorción intestinal y biodisponibilidad potencial. Es de tipo correlacional porque busca la relación entre la variable dependiente (logPapp) y la variable independiente (descriptores moleculares). Es no experimental debido a que no se manipularon las variables; se calcularon descriptores moleculares, se recopiló el logPapp experimental y se predijo el logPapp con los modelos QSAR, y es de corte transversal porque los datos fueron recolectados en un determinado momento.

IV.3 Población y muestra

Se analizaron 2 grupos de datos: (1) el conjunto de datos de moléculas con su permeabilidad aparente en Caco-2 medida experimentalmente, junto a sus descriptores moleculares para construir los modelos QSAR, y (2) el conjunto de datos que corresponde a las moléculas de la base de datos productos naturales del Perú junto a sus descriptores moleculares.

IV.4 Procedimientos, técnicas e instrumentos de recolección de información

IV.4.1 Desarrollo de modelos QSAR para predecir la permeabilidad aparente en células caco-2

IV.4.1.1 Búsqueda de compuestos con permeabilidad aparente en células Caco-2

Se buscó en la literatura[19] moléculas reportadas con permeabilidad aparente (Papp) medidas en células Caco-2 y se colectó la siguiente información de la molécula: nombre, formato *Simplified Molecular Input Line Entry Specification* (SMILES), permeabilidad aparente (cm/s), nombre del articulo y DOI. Se calculó el logaritmo de la permeabilidad aparente (logPapp). Se eliminaron los duplicados.

IV.4.1.2 Cálculo de estructuras 3D y optimización

La estructura 2D en formato SMILES se empleó para generar las coordenadas 3D usando Molcovert[69], se añadieron hidrógenos a pH = 7.4 y se optimizaron las estructuras en steepest descent, usando Merck Molecular Force Field 94 (MMFF 94)[70] con 5000 pasos en el software OpenBabel [71]. Se realizó una búsqueda de 200 confórmeros con 500 pasos y se optimizó nuevamente las estructuras en *gradient descent* con MMFF 94 con 5000 pasos. Las moléculas que no se pudieron optimizar con MMFF 94 fueron optimizadas con Generalized Amber Force Field (GAFF)[72].

IV.4.1.3 Cálculo de descriptores moleculares

Se calcularon 7110 descriptores moleculares 2D y 3D, a partir del código SMILES de los compuestos usando Padel-Descriptor[73] y AlvaDesc[74].

IV.4.1.4 División del conjunto de datos

Para evitar el sobreajuste de los modelos, se dividió el conjunto de datos (data set) en (1) conjunto de entrenamiento (training set) y un (2) conjunto de prueba (test set) en relación 80:20, y manteniendo la distribución del logPapp.

IV.4.1.5 Preprocesamiento

Se imputaron los valores perdidos usando la mediana, se eliminaron los descriptores con varianza cercana a cero y los descriptores altamente correlacionados (Valor absoluto del coeficiente de correlación de Pearson mayor a 0.7, P > 0.7). Finalmente, se centraron y escalaron los descriptores.

IV.4.1.6 Selección de características (Feature selection)

Para un correcto desempeño del modelo, es necesario seleccionar el menor número de descriptores informativos. Se empleó la eliminación de características recursivas con random forest (RFE-RF), los parámetros fueron sizes = 1:200 y la validación cruzada de 5 iteraciones (*five-fold cross validation*, CV5). Los descriptores seleccionados por RFE-RF fueron empleados en el algoritmo genético con random forest para seleccionar el número óptimo de descriptores, los parámetros fueron iters = 100 y CV5.

IV.4.1.7 Modelamiento

Para la construcción de los modelos QSAR, se construyeron 6 modelos de regresión usando el logPapp experimental y los descriptores moleculares calculados. Se empleó la regresión lineal múltiple (MLR), regresión de mínimos cuadrados parciales (PLS), *support vector machine regression* (SVM), *random forest* (RF) y *gradient boosting machine* (GBM). Asimismo, se empleó el método de conjunto por apilamiento (*stacking*) para combinar las predicciones de SVM, RF y GBM usando un modelo lineal, denominado modelo SVM-RF-GBM.

IV.4.1.8 Validación

Se validaron los modelos mediante la validación interna, empleando la validación cruzada de 5 iteraciones repetida 5 veces (*five-fold cross validation repeated 5 times*), y la validación externa empleando el conjunto de prueba. Se calculó el coeficiente de determinación (R^2) y el error cuadrático medio (RMSE) del conjunto de entrenamiento, validación cruzada y conjunto de prueba, para comparar la capacidad predictiva de los modelos QSAR.

La Figura 11 resume los pasos para la construcción del modelo QSAR desarrollado en este trabajo.

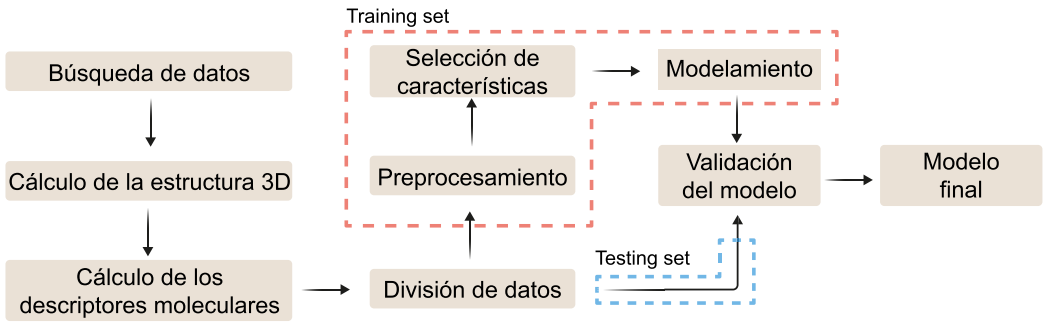


Figura 11. Flujo de trabajo del desarrollo del modelo QSAR

IV.4.1.9 Análisis complementarios de los compuestos y del modelo QSAR

Se determinó el dominio de aplicabilidad del modelo QSAR seleccionado mediante el gráfico de Williams, que emplea el apalancamiento y los residuales estandarizados. Se calculó la importancia relativa de cada variable en los modelos QSAR. Asimismo, se clasificaron los compuestos en alta, media y baja permeabilidad aparente, y se determinó el número de violaciones a las reglas de Lipinski y Veber.

IV.4.2 Generación de la base de datos de productos naturales de la biodiversidad del Perú

Se realizó la búsqueda de artículos hasta el año 2019 empleando las palabras clave "natural products" y "Perú". Se recopilaron 48 artículos[75–122] que contenían estudios sobre plantas medicinales del Perú, cuya actividad farmacológica ha sido reportada y se hayan determinado sus compuestos. La estructura del compuesto fue guardada en formato SMILES. Se construyó una base de datos con los siguientes datos de la molécula: nombre, formato SMILES, actividad farmacológica atribuida, nombre común de la planta, nombre científico de la planta, parte de la planta, lugar de origen o recolección, referencia del articulo y DOI del artículo.

Se clasificaron los productos naturales en base a su vía de biosíntesis, clase, subclase y presencia de un enlace glucosídico, empleando NPClassifier[55] y se revisó manualmente para completar los compuestos que no fueron clasificados. El organismo de procedencia se clasificó a partir del nombre de la especie, en género, familia y clase taxonómica empleando taxonlookup[123].

IV.4.3 Predicción de la permeabilidad aparente en células Caco-2 usando la base de datos de productos naturales mediante el modelo QSAR

Se predijo mediante el modelo QSAR, la permeabilidad aparente en células Caco-2 de los compuestos de la base de datos de productos naturales de la biodiversidad del Perú.

IV.4.4 Análisis estadístico

Se empleó el lenguaje de programación R[124] en el entorno Rstudio[125] y Google Colab. En el análisis de datos, se emplearon los paquetes tidyverse[126], dplyr[127], readr[128], stringr[129], purrr[130], tibble[131], FactoMineR[132], recipes[133] y applicable[134]. Para el modelamiento, se empleó Caret[135], CaretEnsemble[136], pls,[137] e1071,[138] randomForest[139] y gbm[140]. En la visualización de datos, se empleó ggplot2,[141] Factoextra[142], gridExtra[143], Lattice[144] y ggpubr[145]; adicionalmente, se empleó la librería D3Blocks[146] en Python[147]. El Anexo 2 muestra un código de ejemplo escrito en lenguaje R sobre el flujo de trabajo para el desarrollo de un modelo QSAR.

La estructura de los compuestos fue trazada con Accelrys DRAW y guardada en formato SMILES. Las estructuras moleculares fueron graficadas con MarvinSketch.[148]

IV.4.5 Procesamiento computacional

Los experimentos computacionales fueron desarrollados en el Centro de Alto Rendimiento Computacional de la Amazonía Peruana del Instituto de Investigaciones de la Amazonía Peruana (IIAP).[149]

V. RESULTADOS

V.1 Desarrollo de modelos QSAR para predecir la permeabilidad aparente en células caco-2

V.1.1 Desarrollo de modelos QSAR

Se empleó la base de datos reportada por Wang[19] que presenta 1827 moléculas, el cual registró el nombre del compuesto, logPapp y SMILES. Se obtuvo 1817 moléculas únicas luego de eliminar los valores duplicados. Los valores de logPapp mostraron una media = -5.34, valor mínimo (min) = -7.70 y valor máximo (max) = -3.78.

La estructura 3D de las moléculas fue generada a partir del código SMILES. Se calcularon 7112 descriptores moleculares: 5666 descriptores de AlvaDesc y 1444 de PadelDescriptor. Posteriormente, se dividió el conjunto de datos en 2 grupos: un conjunto de en entrenamiento y conjunto prueba en relación 80:20 manteniendo la distribución del logPapp (Figura 12a). Se emplearon 1455 compuestos para el conjunto de entrenamiento (80%) y 362 compuestos para el conjunto de prueba (20%).

Preprocesamiento

Se encontraron descriptores moleculares con valores faltantes; por ello, se imputaron los valores faltantes empleando la mediana en cada columna. Se eliminaron los descriptores con varianza cercana a cero, altamente correlacionados (valor absoluto del coeficiente de correlación de Pearson > 0.70) y, finalmente se escalaron los datos. Se conservaron 523 descriptores moleculares.

Selección de características

Se eliminaron variables innecesarias para mejorar la interpretabilidad del modelo. Se empleó la eliminación recursiva de características con bosque aleatorio (RFE-RF) para seleccionar el menor número de descriptores (Anexo 3). Se seleccionaron 60 descriptores Luego, se empleó un algoritmo genético con bosque aleatorio para

seleccionar el número óptimo de descriptores. Se seleccionaron 41 descriptores moleculares para el modelamiento (Anexo 4).

Se realizó un análisis de componentes principales (PCA) para ver la distribución de moléculas de los 41 descriptores seleccionados (Figura 12b, Figura 12c). La división del conjunto de datos mantuvo la distribución en el conjunto de entrenamiento y prueba según el histograma (Figura 12a); asimismo, están distribuidos en el mismo espacio químico (Figura 12b).

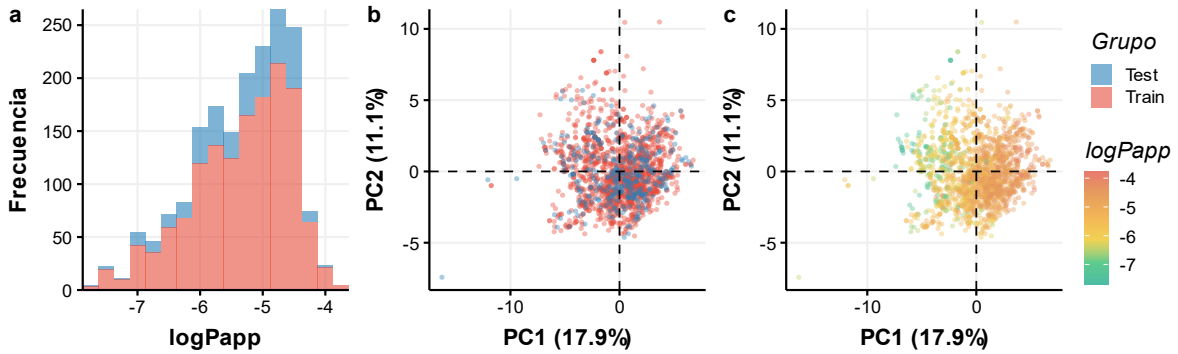


Figura 12. Análisis de datos del conjunto de datos de modelado. a, Distribución de logPapp en conjunto de entrenamiento y conjunto de prueba, b, PCA de 41 descriptores seleccionados que muestran Train y Test. c, PCA de 41 descriptores seleccionados que muestran el valor logPapp.

Modelamiento

Se construyeron 6 modelos QSAR: MLR, PLS, SVM, RF, GBM y SVM-RF-GBM empleando el conjunto de entrenamiento con los 41 descriptores seleccionados. Se calcularon las métricas de RMSE y R^2 del conjunto de entrenamiento, validación cruzada y conjunto de prueba, para comparar el desempeño de cada modelo (Tabla 1).

Los modelos MLR y PLS presentaron un desempeño predictivo similar, como se observa en los valores de RMSE y R^2, para el conjunto de entrenamiento, prueba y validación cruzada; asimismo, presentaron los valores de RMSE más altos y R^2 más bajos, $RMSE_{Test}$ = 0.47 y R^2_{Test} = 0.63. Los modelos SVM, RF y GBM

presentaron un $RMSE_{Test}$ de 0.40, 0.39 y 0.39, respectivamente; y un R^2 de 0.73, 0.74 y 0.74, respectivamente, siendo los de mejor desempeño predictivo.

Se combinaron los 3 modelos para obtener el modelo SVM-RF-GBM, el cual presentó una mejor capacidad predictiva, $RMSE_{test}$ = 0.38 y R^2 = 0.76, por lo que se escogió para los posteriores análisis. Los resultados se compararon con otros estudios QSAR para predecir el logPapp en Caco-2 (Tabla 2), donde se indica el autor, método, descriptores moleculares, número de compuestos empleados (N), $RMSE_{Test}$ y R^2_{Test}.

Para una mejor visualización de la capacidad predictiva de los 6 modelos QSAR, se generó un gráfico de dispersión de los valores de logPapp experimentales y predichos del conjunto de entrenamiento y prueba, para cada modelo (Figura 13). Se observó que los modelos MLR y PLS presentaron una mayor dispersión en el gráfico; es decir, una mayor variación entre el valor experimental y el valor predicho.

Tabla 1. Métricas del conjunto de entrenamiento, conjunto de prueba y validación cruzada e hiperparámetros

Modelo	$RMSE_{Train}$	R^2_{Train}	$RMSE_{CV5}$	R^2_{CV5}	$RMSE_{Test}$	R^2_{Test}	Hiperparámetros
MLR	0.43	0.70	0.44	0.68	0.47	0.63	-
PLS	0.43	0.70	0.44	0.68	0.47	0.63	ncomp = 11
SVM	0.28	0.87	0.40	0.74	0.40	0.73	Sigma = 0.015, C = 2
RF	0.16	0.97	0.40	0.75	0.39	0.74	mtry = 18
GBM	0.19	0.94	0.40	0.74	0.39	0.74	n.trees = 100, interaction.depth = 16
SVM-RF-GBM	0.19	0.94	0.38	0.76	0.38	0.76	-

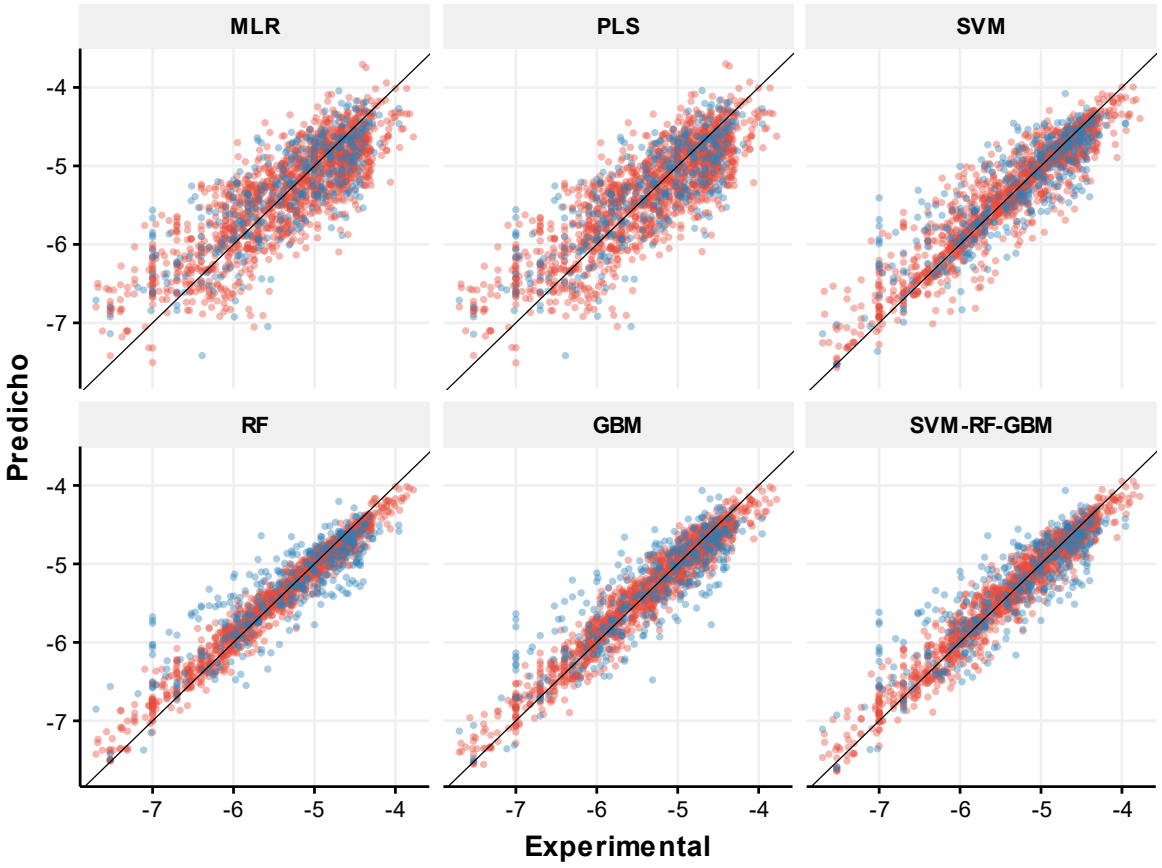


Figura 13. Gráfico de dispersión del logPapp experimental y predicho para el conjunto de entrenamiento (rojo) y el conjunto de prueba (azul)

Tabla 2. Análisis comparativo de optimización de modelos

Estudio	**Método**	**Descriptores**	**N**	**$RMSE_{Test}$**	**R^2_{Test}**	**Referencia**
Este estudio	SVM-RF-GBM	Descriptores 2D y 3D PadelDescriptor y AlvaDesc	1817	0.38	0.76	-
Wang et al., 2020	Dual-RBF neural network	Descriptores PadelDescriptor 2D	1827	0.39	0.76	[19]
Wang et al., 2016	Boosting	Descriptores MOE 2D y 3D	1017	0.31	0.812	[18]
Lanevskij et al., 2019	Nonlinear least squares	log $D_{o/w}$, pKa, N_{HD}, V_x	442	0.49	0.77	[13]

V.1.2 Análisis complementarios de los compuestos y del modelo QSAR

V.1.2.1 Dominio de aplicabilidad

Para identificar los compuestos cuyas predicciones de logPapp son confiables, es necesario determinar el dominio de aplicabilidad. Para ello, se calcularon los apalancamientos del conjunto de entrenamiento y del conjunto de prueba utilizando la *hat matrix* del conjunto de entrenamiento. Se calcularon los residuos estandarizados para el conjunto de entrenamiento y el conjunto de prueba. El gráfico de Williams se creó utilizando los apalancamientos y los residuales estandarizados (Figura 14). Se tomaron como límites del dominio de aplicabilidad las líneas discontinuas horizontales, que representan un residual estándar de ± 3, y la línea discontinua vertical que representa un apalancamiento de advertencia, h* = 0.0866.

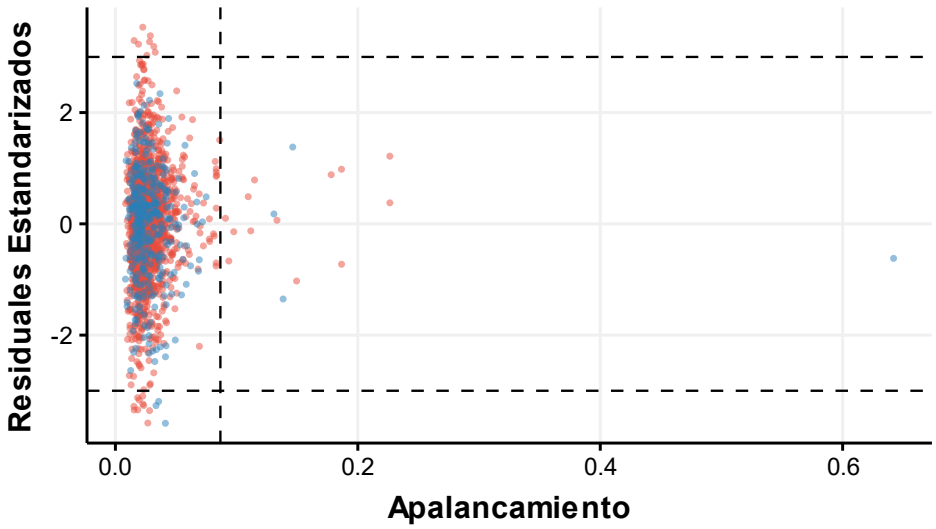


Figura 14. Gráfico de Williams. Conjunto de entrenamiento (rojo), conjunto de prueba (azul).

Se identificaron 38 valores atípicos o *outliers* según el gráfico de Williams que se encuentran fuera del dominio de aplicabilidad: 31 (2.13%) valores atípicos del conjunto de entrenamiento y 7 (1.93%), del conjunto prueba. (Figura 14, Anexo 11)

V.1.2.2 Importancia de las variables en el modelo

Se calculó el valor de importancia relativa para los 6 modelos desarrollados. Cada modelo mostró un valor diferente de importancia relativa a cada variable (Figura 15). En el modelo SVM-RF-GBM, los 10 primeros descriptores son maxHBint7, ALogP, SpMAD_Dzs, maxHBint5, maxHBint9, Eta_D_epsiD, maxHBint3, SHED_DL, maxHBd y Hypertens.80. Asimismo, los descriptores maxHBint7 y Eta_D_epsiD se encontraron entre los 10 primeros descriptores en los 6 modelos QSAR.

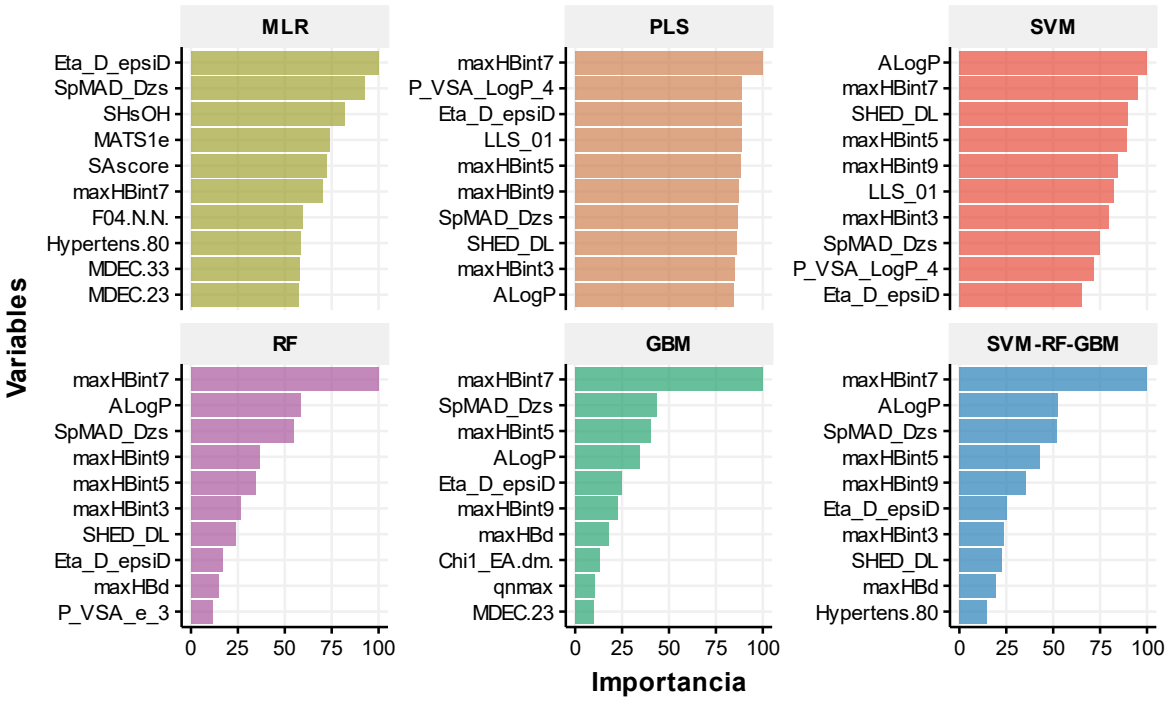


Figura 15. Gráfico de barras de la importancia de los diez principales descriptores moleculares en cada modelo

Los 10 mejores descriptores moleculares según el modelo QSAR SVM-RF-GBM presentaron una moderada o baja correlación lineal (Tabla 3) con el logPapp experimental según el coeficiente de correlación de Pearson (*r*). Además, los 41 descriptores moleculares presentaron una moderada, baja o nula correlación lineal con el logPapp experimental según el gráfico de dispersión (Anexo 7) y presentaron

valores atípicos que se escapan de la tendencia de la mayoría de los puntos (moléculas).

Tabla 3. Los diez mejores descriptores moleculares según el modelo SVM-RF-GBM. r: Coeficiente de correlación de Pearson

Descriptor Molecular	**Grupo**	***r***	**Descripción**
maxHBint7	E-State	-0.50	Descriptor E-State máximo de fuerza para enlaces de hidrógeno potenciales de longitud de camino 7
ALogP	Constitucional	0.46	Ghose-Crippen LogKow
SpMAD_Dzs	Barysz matrix	-0.43	Desviación media absoluta espectral de la matriz de Barysz/ ponderado por el I-State
maxHBint5	E-State	-0.48	Descriptor E-State máximo de fuerza para enlaces de hidrógeno potenciales de longitud de camino 5
maxHBint9	E-State	-0.47	Descriptor E-State máximo de fuerza para enlaces de hidrógeno potenciales de longitud de camino 9
Eta_D_epsiD	Índice ETA	-0.39	Medida Eta de los átomos donantes de enlace de hidrógeno
maxHBint3	E-State	-0.46	Descriptor E-State máximo de fuerza para enlaces de hidrógeno potenciales de longitud de camino 3
SHED_DL	Pharmacophore descriptor	-0.47	Donante- Lipofílico SHED
maxHBd	E-State	-0.35	E-State máximo para donantes de enlaces de hidrógeno (fuertes).
Hypertens.80	Drug-like index	0.41	Índice antihipertensivo similar al Ghose-Viswanadhan-Wendoloski al 80%

V.1.2.3 Aplicación de las reglas de Lipinski y Veber

El conjunto de datos de 1817 de compuestos se clasificó en 3 grupos, según el valor de logPapp, como clase L (logPapp < -6), clase M (logPapp < -5 y logPapp > -6) y clase H (logPapp > -5).[150] Se calculó la media, mediana y rango del logPapp y 6 descriptores moleculares que incluyen el MlogP, número de donadores de hidrógeno (HBD), número de aceptores de hidrógeno (HBA), *topological polar surface área* (TPSA), número de enlaces rotables (RBN). Asimismo, se contaron las violaciones de reglas de Lipinski [151] y Veber [152] y expresaron en frecuencia y porcentaje.(Tabla 4 y Figura 16) Los grupos de baja, media y alta permeabilidad presentaron 364, 730 y 723 compuestos, respectivamente. Los compuestos con alta permeabilidad intestinal (Clase H) presentaron pocos compuestos con violaciones a las reglas de Lipinski y Veber en relación con la cantidad de compuestos de su grupo (Tabla 4), por otro lado, los compuestos con baja permeabilidad aparente presentaron un mayor porcentaje de violaciones a estas reglas.

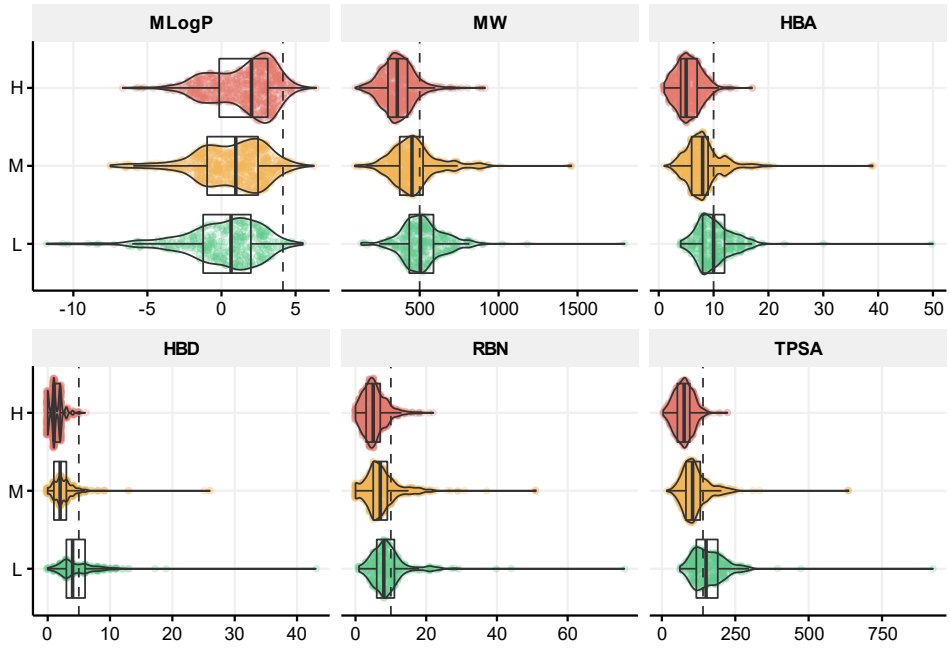


Figura 16. Gráfico de distribución de 6 descriptores moleculares entre 3 clases de permeabilidad. H: permeabilidad alta, M: permeabilidad media, L: permeabilidad baja.

Tabla 4. Resumen del conjunto de datos de modelado: logPapp y 6 descriptores moleculares

	Total	**Clase L**	**Clase M**	**Clase H**
Número de compuestos	1817	**364**	**730**	**723**
Caco-2 Papp media (cm/s)	-7.7	-6.55	-5.46	-4.62
Caco-2 Papp mediana (cm/s)	-3.78	-6.43	-5.42	-4.64
MW rango (g/mol)	89.05–1797.92	129.1–1797.92	89.05–1460.67	92.05–913.56
MW media/mediana (g/mol)	437.13/422.11	521.95/504.77	467.58/451.16	363.67/358.14
MlogP rango	1–50	4–50	1–39	1–17
MlogP media/mediana	7.52/7	10.25/10	8.17/8	5.48/5
HBD rango	0–43	0–43	0–26	0–6
HBD mean/median	2.5/2	4.69/4	2.6/2	1.31/1
HBA rango	-11.82–6.39	-11.82–5.49	-7.5–6.24	-6.67–6.39
HBA media/mediana	0.85/1.29	0.07/0.65	0.63/0.97	1.48/2.03
TPSA (Å)	0–76	1–76	0–51	0–22
TPSA media/mediana (Å)	7.01/6	9.32/8	7.52/7	5.33/5
RBN rango	4.44–923.49	61.36–923.49	16.61–635.68	4.44–222.76
RBN media/mediana	108.34/99.83	161.39/150.51	114.48/104.47	75.44/75.71
Lipinski (Ro5) [151] **Absorption Permeability**				
MW ≤ 500	476 (26.2%)	195 (53.6%)	226 (31.0%)	55 (7.6%)
ClogP5 ≤ 5 (MLogP ≤ 4.15)	72 (4.0%)	10 (2.7%)	26 (3.6%)	36 (5.0%)
HBA ≤ 10	279 (15.4%)	141 (38.7%)	126 (17.3%)	12 (1.7%)
HBD ≤ 5	157 (8.6%)	122 (33.5%)	34 (4.7%)	1 (0.1%)
Veber [152]				
RBN ≤ 10	272 (15.0%)	96 (26.4%)	129 (17.7%)	47 (6.5%)
TPSA ≤ 140	381 (21.0%)	209 (57.4%)	154 (21.1%)	18 (2.5%)

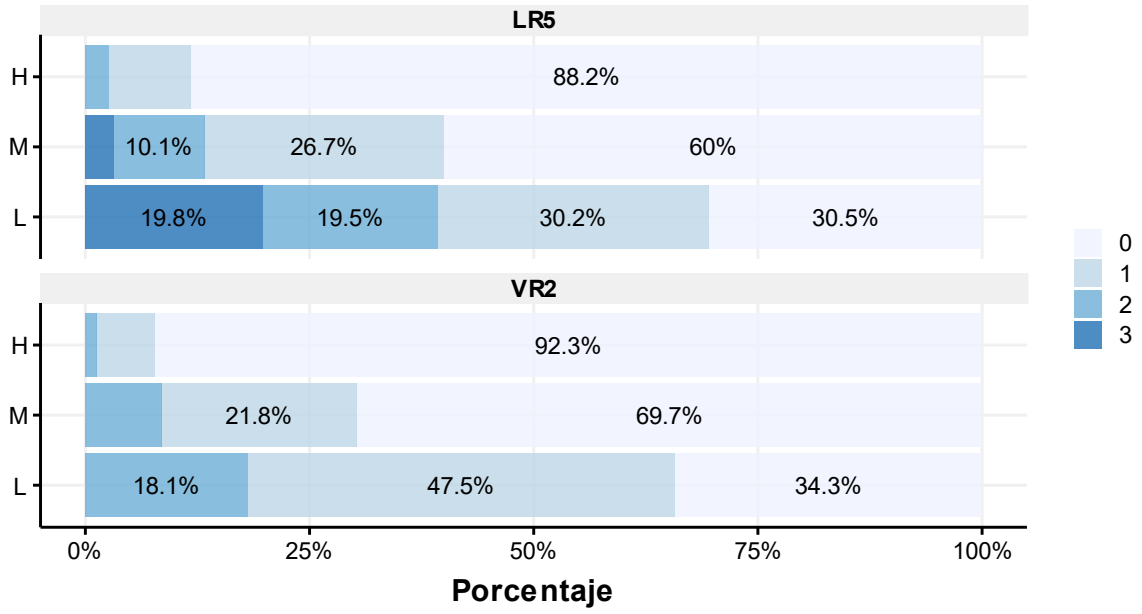


Figura 17. Gráfico de barras del número de violaciones a las Reglas de Lipinski y Veber. Izquierda: Reglas de Lipinski (LR5). Derecha: Reglas de Veber (VR2). H: permeabilidad alta, M: permeabilidad media, L: permeabilidad baja.

V.2 Generación de la base de datos de productos naturales de la biodiversidad del Perú

Se recopilaron 516 compuestos de la biodiversidad del Perú, los cuales se clasificaron en base a su clase metabólica, taxonomía del organismo de origen y actividad farmacológica reportada.

V.2.1 Clasificación metabólica

Se clasificaron en base a las 6 vías metabólicas de producción de metabolitos secundarios: alcaloides (n = 78), aminoácidos y péptidos (n = 11), ácidos grasos (n = 91), policétidos (n = 4), compuestos fenólicos (n = 141) y terpenoides (n = 191). Asimismo, se subdividen en 24 grupos y 97 subgrupos. Los 24 grupos son alcaloides de ornitina, alcaloides de triptófano, alcaloides de tirosina, alcaloides de lisina y pseudoalcaloides, aminoácidos y péptidos, acilos grasos, ácidos grasos y conjugados, amidas grasas, ésteres grasos, policétidos, cumarinas y diarilheptanoides, flavonoides, isoflavonoides, ácidos fenólicos (C6-C1),

fenilpropanoides (C6-C3), estilbenoides, xantonas, apocarotenoides, diterpenoides, meroterpenoides y triterpenoides, monoterpenoides, sesquiterpenoides y esteroides.(Anexo 8) Asimismo, 52 compuestos de las diferentes vías metabólicas presentaron enlace glucosídico.

V.2.2 Clasificación taxonómica del organismo de origen

La base de datos registró 59 especies diferentes pertenecientes a 48 géneros, 29 familias, 21 órdenes, 6 clases o divisiones y 3 reinos (Figura 19). De los 516 compuestos reportados, las angiospermas (n = 474, 91.86%) presentaron la mayor cantidad de reportes y en menor cantidad los demás grupos, como anfibios (n = 31, 6.01%), briófitas (n = 4, 0.77%), bacilos (n = 3, 0.58%), citofagia (n = 2, 0.39%) y actinomicetos (n = 2, 0.39%). (Anexo 9)

V.2.3 Clasificación farmacológica

Se han reportado 53 actividades farmacológicas diferentes, donde un compuesto puede presentar más de una actividad farmacológica. Las actividades farmacológicas reportadas con mayor frecuencia son la actividad citotóxica (n = 146), antioxidante (n = 101), leishmanicida (n = 65), antibacteriana (n = 59) y antiinflamatoria (n = 47).(Figura 24).

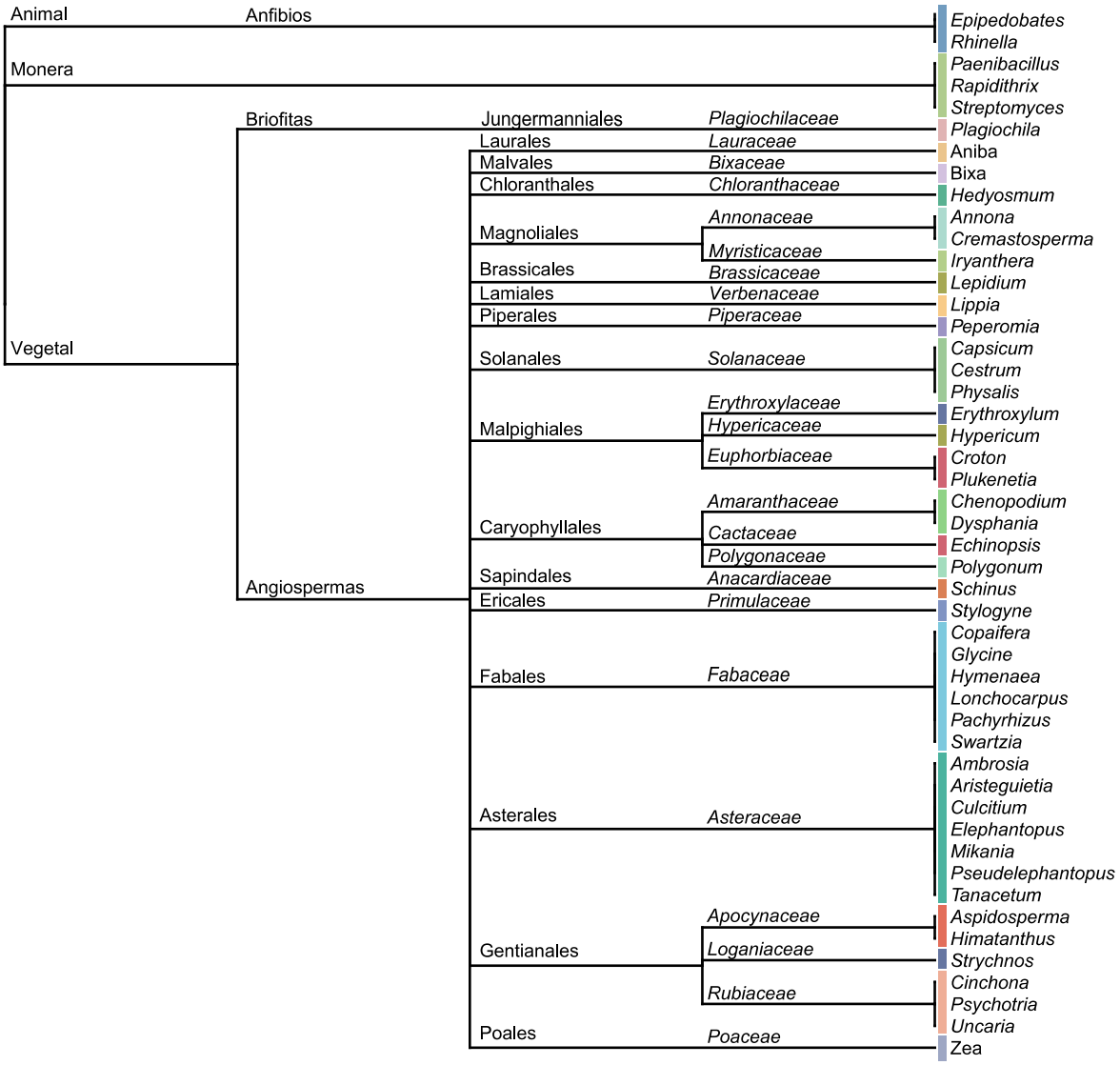


Figura 18. Agrupamiento de jerárquico de las especies de la base de datos de productos naturales del Perú.

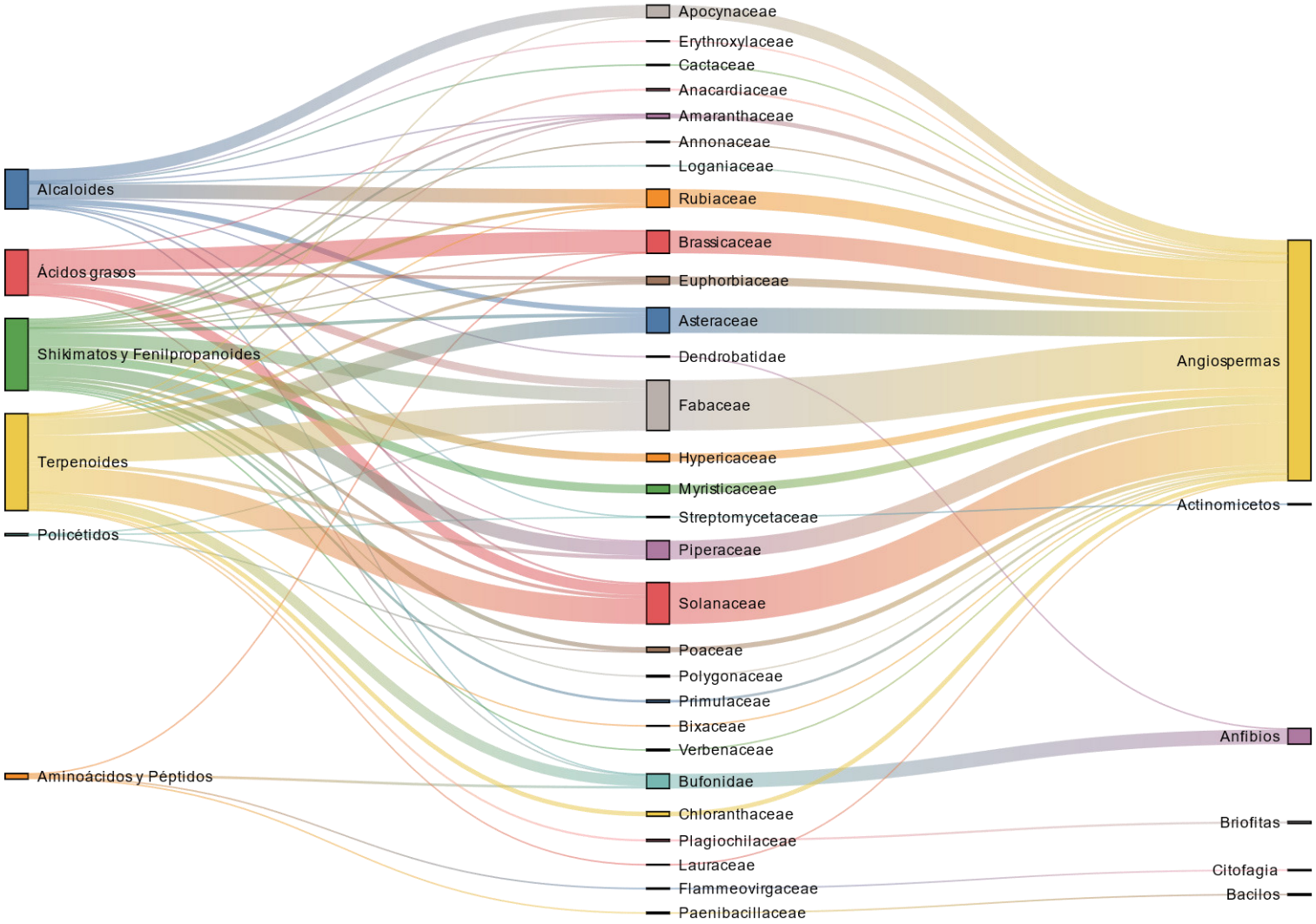


Figura 19. Diagrama de flujo del grupo fitoquímico y familia del producto natural

V.3 Predicción de la permeabilidad aparente en células Caco-2 usando la base de datos de productos naturales mediante el modelo QSAR

V.3.1 Predicción de la permeabilidad aparente en Caco-2

Se predijo el logPapp de 516 productos naturales mediante el modelo SVM-RF-GBM. Se encontraron 347 (67.24%) compuestos con un logPapp > -5, 117 (22.67%) compuestos con logPapp < -5 y logPapp > -6; y 52 compuestos (10.07%) con logPapp < -6. Se observa una distribución del logPapp predicho con una media = -4.93, min = -6.60 y max = -4.00. (Anexo 10)

V.3.2 Determinación del dominio de aplicabilidad

Se determinó el dominio de aplicabilidad para identificar los compuestos que son predichos correctamente por el modelo QSAR. Los compuestos fuera del dominio de aplicabilidad son considerados valores atípicos o *outliers*, por lo que su logPapp predicho no es confiable y, en consecuencia, deben eliminarse. En el gráfico de apalancamientos (Figura 20), la línea discontinua horizontal es el apalancamiento de advertencia, h* = 0.0866, que marca el límite del dominio de aplicabilidad.

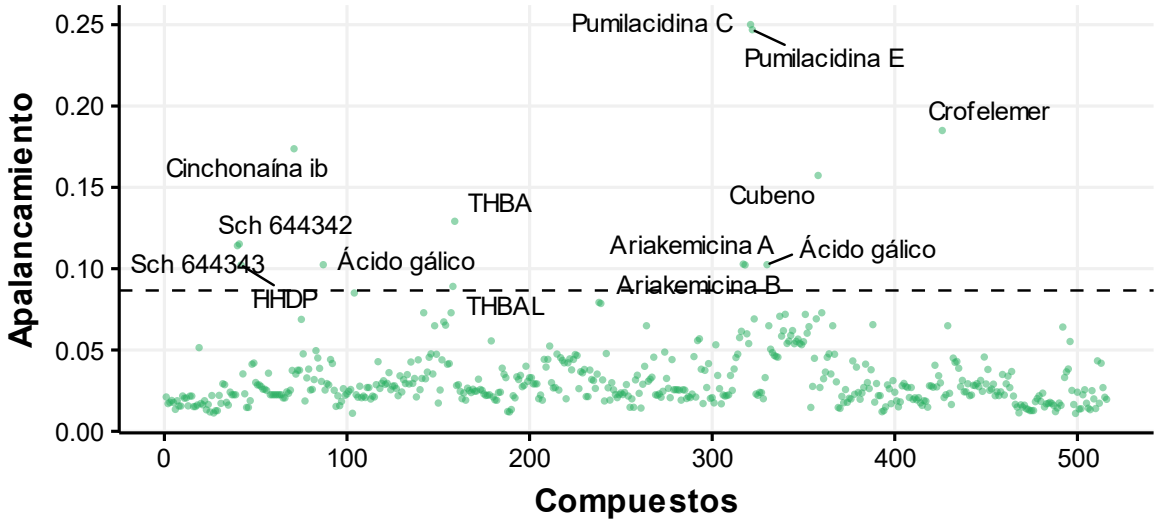
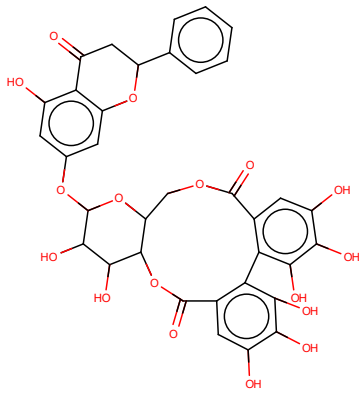
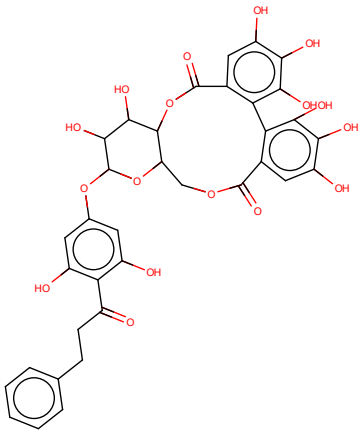
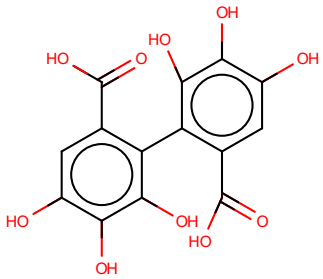
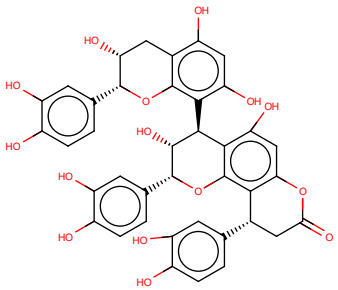
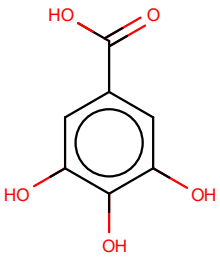
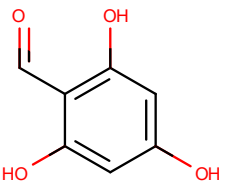


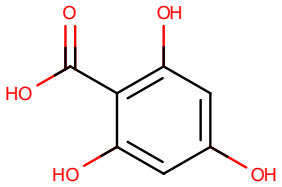
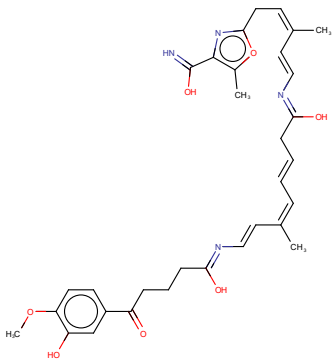
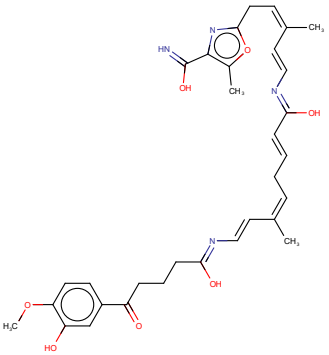
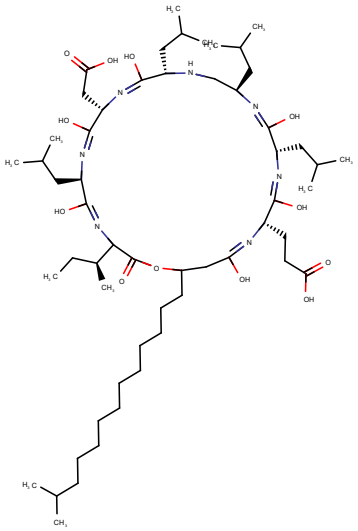
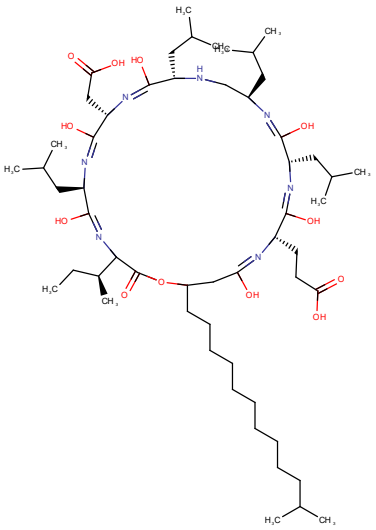
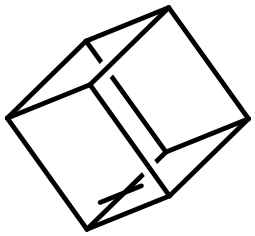
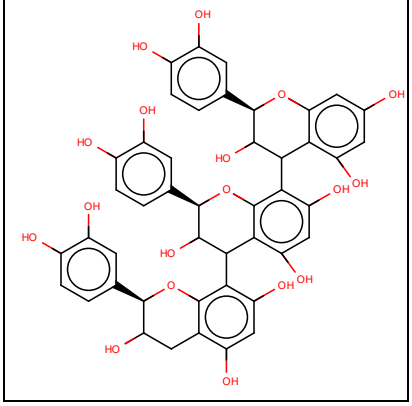
Figura 20. Gráfico de apalancamiento de la base de datos de productos naturales

Se encontraron 14 (2.71%) compuestos en la base de datos de productos naturales fuera del dominio de aplicabilidad para el modelo QSAR SVM-RF-GBM (Anexo 11). La Tabla 5 presenta la estructura y nombre de los 13 productos naturales fuera del dominio de aplicabilidad, se descartó un compuesto duplicado (ácido gálico) debido a que fue identificado en 2 artículos.

Se removieron los 13 productos naturales fuera del dominio de aplicabilidad: SCH 644343, SCH 644342, ácido hexahidroxidifénico (HHDP), cinchonain Ib, ácido gálico, 2,4,6-trihidroxibenzaldehído (THBAL), ácido 2,4,6-trihidroxibenzoico (THBA), ariakemicina A, ariakemicina B, pumilacidina C, pumilacidina E, cubeno y crofelemer. Los 502 compuestos restantes se emplearon para los análisis posteriores.

Tabla 5. Productos naturales fuera del dominio de aplicabilidad

SCH 644343	SCH 644342	HHDP
		
Cinchonain Ib	**Ácido gálico**	**THBAL**
		

THBA	Ariakemicina A	Ariakemicina B
		
Pumilacidina C	**Pumilacidina E**	**Cubeno**
		
	Crofelemer	
		

V.3.3 Distribución del logPapp de la base de datos de productos naturales

Se analizó nuevamente el logPapp predicho por el modelo SVM-RF-GBM, empleando los 502 productos naturales restantes. Se encontraron 338 (67.33%)

compuestos con una alta permeabilidad aparente (logPapp > -5); 113 (22.51%), moderada permeabilidad (logPapp < -5 y logPapp > -6); y 51 (10.16%), baja permeabilidad (logPapp < -6), lo cual se observa en la distribución del logPapp predicho (Figura 21). Asimismo, presenta una media = -4.93, min = -6.60 y max = -4.00.

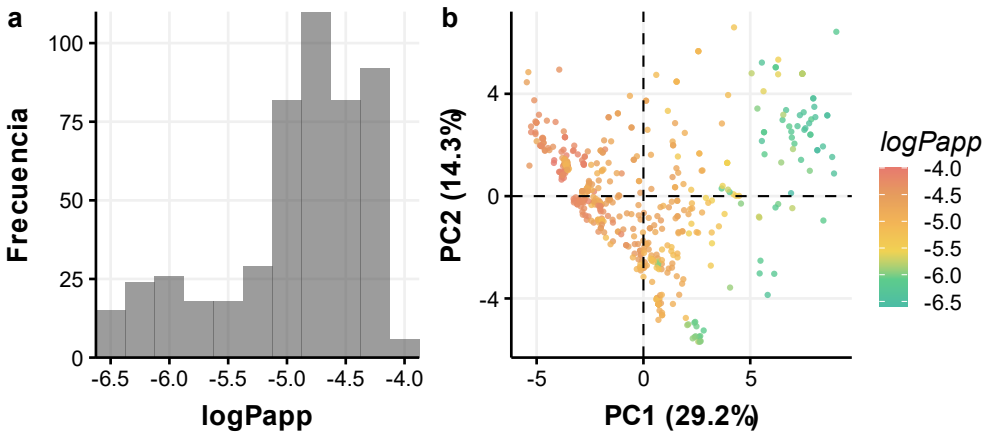


Figura 21. Análisis de datos de la base de datos de productos naturales. a, Distribución de logPapp de la base de datos de productos naturales, b, PCA de la base de datos de productos naturales

V.3.4 Asociación de la permeabilidad aparente y grupos químicos

Se relacionó el logPapp predicho con la vía metabólica de los productos naturales (Figura 22). Los compuestos de la vía metabólica de los terpenoides, policétidos, alcaloides, ácidos grasos, shikimatos y fenilpropanoides presentaron compuestos con alta, media y baja permeabilidad aparente. Por otor lado, los aminoácidos y péptidos presentaron una media permeabilidad aparente. Asimismo, se subdividieron en base a su grupo químico. Se encontraron 19 grupos químicos en los que predominan compuestos con una alta permeabilidad aparente (logPapp > -5): apocarotenoides, ácidos grasos, monoterpenoides, diterpenoides, sequiterpenoides, cumarinas y acil grasos, alcaloides de lisina y pseudoalcaloides, policétidos, fenilpropanoides, estilbenoides, ácidos fenólicos, lignanos, isoflavonoides, ácidos grasos y conjugados, alcaloides de triptófano, alcaloides de tirosina, y meroterpenoides y triterpenoides (Figura 23). Los esteroides, alcaloides

de ornitina y flavonoides presentaron valores de permeabilidad baja, media y alta. Las xantonas presentaron una baja permeabilidad aparente.

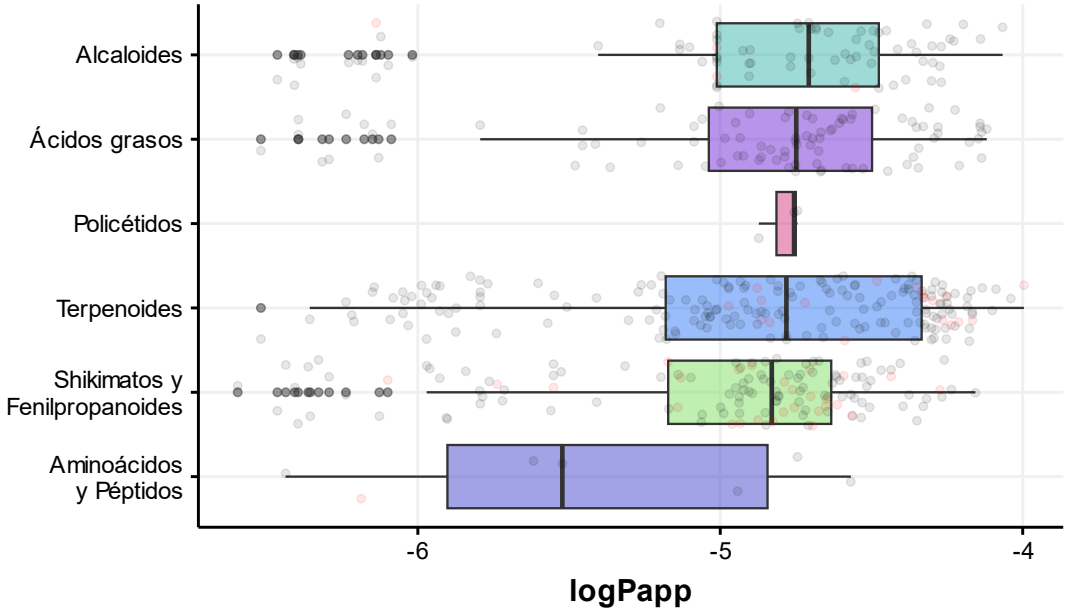


Figura 22. Distribución del logPapp predicho de la base de datos productos naturales por vía metabólica de los grupos químicos. Presencia de enlace glucosídico (puntos rojos), ausencia de enlace glucosídico (puntos negros).

Los 338 productos naturales con alta permeabilidad aparente se relacionaron con su grupo químico y actividad farmacológica. Se encontraron 124 terpenoides con alta permeabilidad aparente que están asociados principalmente con una actividad citotóxica, leishmanicida, antiproliferativo, antioxidante y antibacteriano. Se encontraron 66 ácidos grados con una alta permeabilidad aparente, relacionados con una actividad antioxidante, citotóxica, inhibidor de la amida hidrolasa de ácidos grasos (FAAH), inmunomodulador y antibacteriano. Se encontraron 93 compuestos provenientes de la vía de los shikimatos y fenilpropanoides con una alta permeabilidad aparente y se relacionaron principalmente con la actividad antioxidante, citotóxica, estrogénica, antiinflamatoria y antiviral. Los alcaloides con alta permeabilidad aparente (n = 49) se asociaron con una actividad antiinflamatoria, antioxidante, citotóxica, antiparasitaria, antiviral, inmunoestimulante, genotóxica, hepatotóxica y analgésica. Los policétidos con alta

permeabilidad aparente (n = 3) se asociaron con una actividad laxante, antianémica y antibacteriana. Finalmente, el grupo de aminoácidos y péptidos con alta permeabilidad aparente (n = 3) se asoció con una actividad antiproliferativa y antibacteriana.

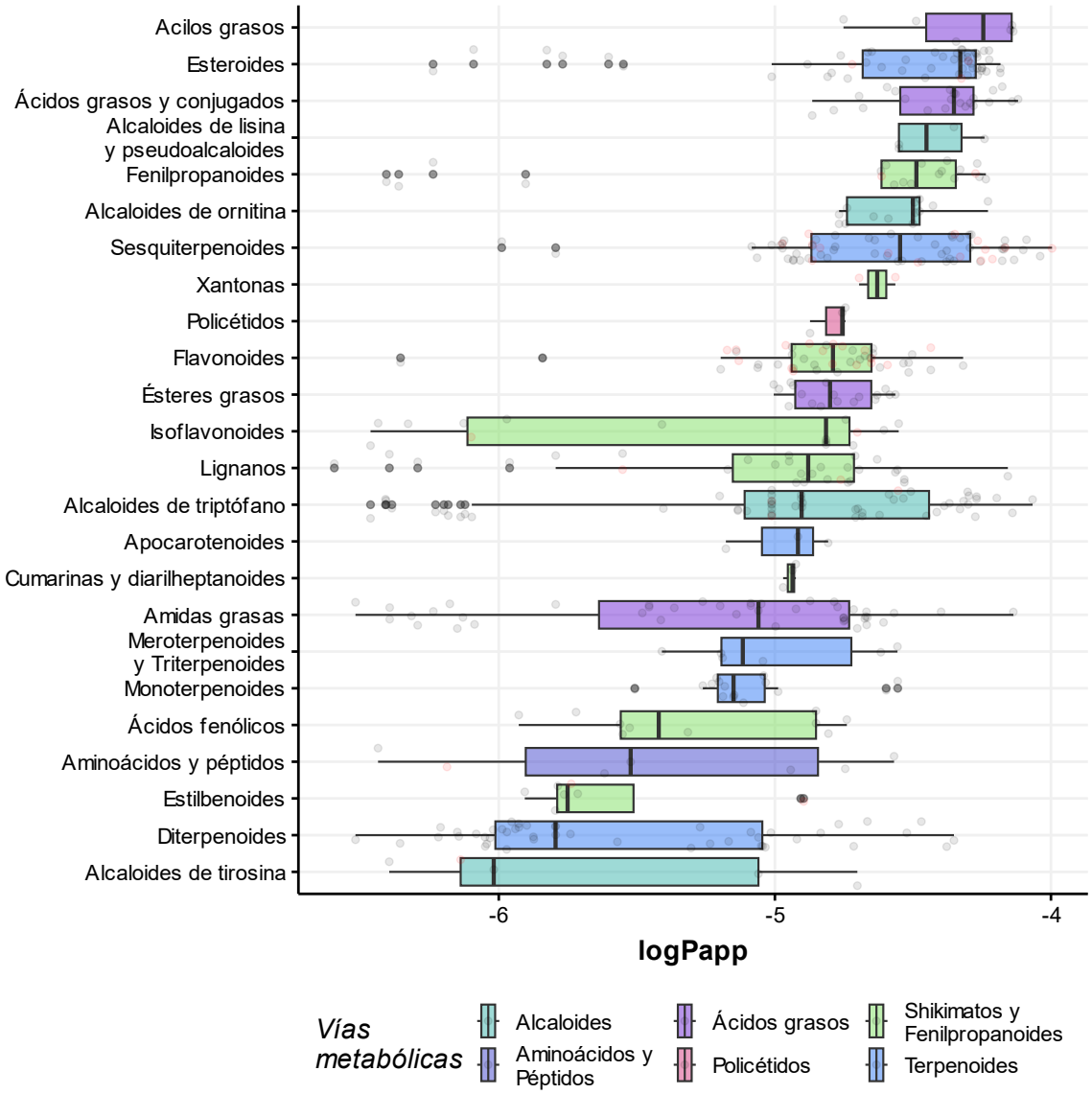


Figura 23. Distribución del logPapp predicho de la base de datos productos naturales por grupos químicos. Presencia de enlace glucosídico (puntos rojos), ausencia de enlace glucosídico (puntos negros).

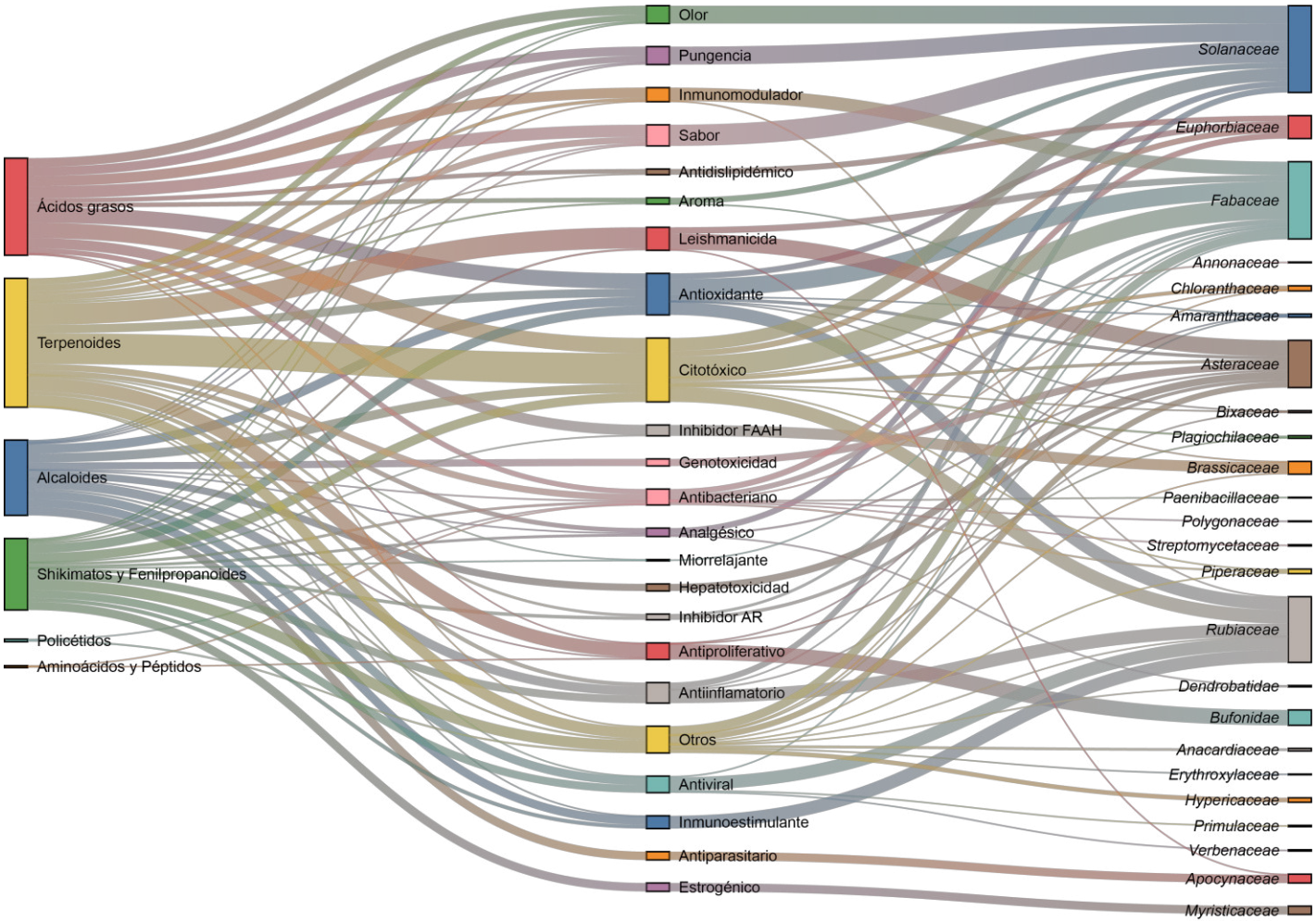


Figura 24. Diagrama de flujo de la vía metabólica, actividad farmacológica reportada y familia de la especie de la que se obtuvo el producto natural

V.3.5 Determinación de las reglas de Lipinski y Veber

Se dividió el conjunto de datos de productos naturales en 3 clases: baja permeabilidad (logPapp ≤ -6, n = 51), media permeabilidad (-6 < logPapp < -5, n = 113) y alta permeabilidad (logPapp ≥ -5, n = 338).[150]

Los compuestos del grupo de alta permeabilidad presentaron valores de MW, MlogP, HBD, HBA, TPSA y RBN menores que el del grupo de moderada y baja permeabilidad, lo cual se observa en la Tabla 6 en su valor de media, mediana y rango, y en la Figura 25, mediante su distribución y mediana. Asimismo, el grupo de alta permeabilidad presentó la menor cantidad de violaciones a las reglas de Lipinski y Veber: MW > 500 (0.59%), HBD > 5 (0%), HBA > 10 (0%), RBN > 10 (18.93%) y TPSA > 140 (0%); a excepción del MLogP > 4.15 (34.32%).

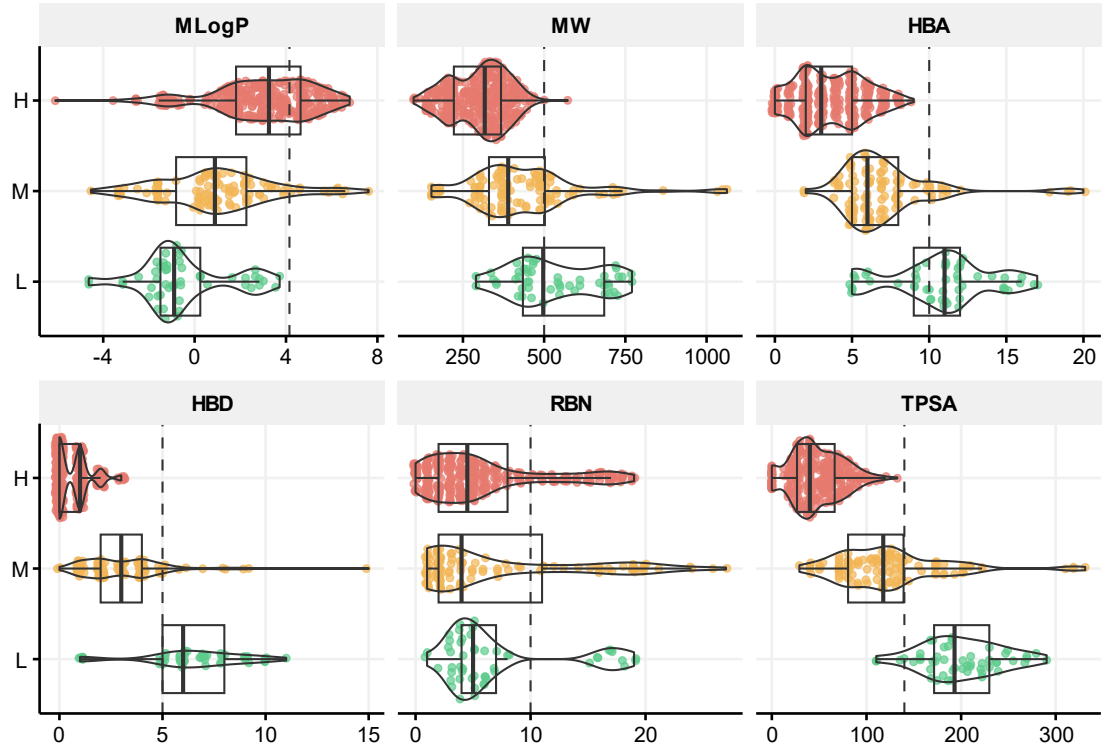


Figura 25. Gráfico de distribución de 6 descriptores moleculares de la base de datos de productos naturales del Perú. H: permeabilidad alta, M: permeabilidad media, L: permeabilidad baja.

Tabla 6. Resumen de la base de datos de productos naturales del Perú: logPapp y 6 descriptores moleculares

	Total	**L Class**	**M Class**	**H Class**
Número de compuestos	**502**	**51**	**113**	**338**
Caco-2 logPapp media (cm/s)	-4.93	-6.25	-5.39	-4.58
Caco-2 logPapp mediana (cm/s)	-4.78	-6.23	-5.23	-4.6
MW rango (g/mol)	98.04–1062.74	290.08–770.41	153.02–1062.74	98.04–572.44
MW media/mediana (g/mol)	354.41/344.18	534.25/497.17	432.07/389.32	301.31/317.25
MlogP rango	-6.1–7.62	-4.63–3.72	-4.54–7.62	-6.1–6.79
MlogP media/mediana	2.18/2.39	-0.45/-0.9	0.940.89	3.0/3.25
HBD rango	0–15	1–11	0–15	0–3
HBD media/mediana	1.82/1	6.37/6	3.16/3	0.68/1
HBA rango	0–20	5–17	2–20	0–9
HBA media/mediana	5.11/5	10.76/11	7.06/6	3.6/3
TPSA (Å)	0–331.14	110.38–290.43	29.1–331.14	0–132.5
TPSA media/mediana (Å)	79.87/61.67	200.39/193.1	124.4/117.73	46.8/40.13
RBN rango	0–27	1–19	1–27	0–19
RBN media/mediana	6.29/4	7.02/5	7.14/4	5.9/4.5
Lipinski (Ro5) [151] **Absorption Permeability**				
MW ≤ 500	56 (11.16%)	25 (49.02%)	29 (25.66%)	2 (0.59%)
ClogP5 ≤ 5 (MLogP ≤ 4.15)	126 (25.10%)	0 (0.00%)	10 (8.85%)	116 (34.32%)
HBD ≤ 5	45 (8.96%)	37 (72.55%)	8 (7.08%)	0 (0.00%)
HBA ≤ 10	41 (8.17%)	31 (60.78%)	10 (8.85%)	0 (0.00%)
Veber [152]				
RBN ≤ 10	106 (21.12%)	10 (19.61%)	32 (28.32%)	64 (18.93%)
TPSA ≤ 140	75 (14.94%)	47 (92.16%)	28 (24.78%)	0 (0.00%)

La Figura 26 y Anexo 12 muestran la cantidad y porcentaje de violaciones a las reglas de Lipinski y Veber por cada grupo. El grupo con alta permeabilidad aparente (n = 338) presentó 222 (65.7%) compuestos y 274 (81.1%) compuestos con 0 violaciones a las reglas de Lipinski y a las reglas de Veber, respectivamente. Además, 114 (33.7%) compuestos y 64 (18.9%) compuestos presentaron 1 violación a las reglas de Lipinski y a las reglas de Veber.

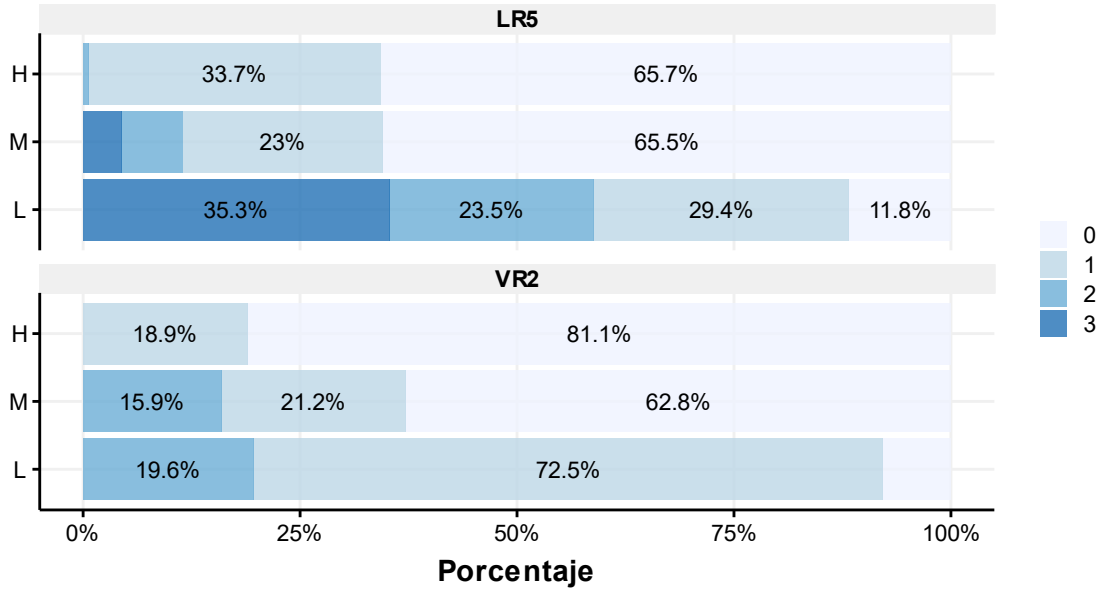


Figura 26. Gráfico de barras del número de violaciones a las Reglas de Lipinski o Veber de la base de datos de productos naturales del Perú. Izquierda: Reglas de Lipinski (LR5). Derecha: Reglas de Veber (VR2). H: permeabilidad alta, M: permeabilidad media, L: permeabilidad baja.

VI. DISCUSIÓN

VI.1 Desarrollo de modelos QSAR para predecir la permeabilidad aparente en células Caco-2

La permeabilidad aparente es una medida de la cantidad de compuesto transportado por tiempo en células Caco-2, que permite inferir parámetros farmacológicos como la absorción y biodisponibilidad.[14]

En este estudio, se desarrollaron 6 modelos QSAR para predecir la permeabilidad aparente en productos naturales de la biodiversidad del Perú. Se empleó un conjunto de datos de 1817 moléculas del conjunto de datos reportado por Wang[19]; sin embargo, la base de datos original contenía 1827 moléculas debido a que presentaban duplicados, lo cual ha podido sesgar sus resultados. En otros estudios QSAR, se emplearon 1017 y 442 compuestos,[13,18] lo cual impacta en la diversidad de compuestos que puede predecir correctamente.

Se seleccionaron finalmente 41 descriptores a partir de 7112 descriptores moleculares, lo que indica que existe una variedad de descriptores que no aportan información o presentan redundancia en la información; es decir, están altamente correlacionados.

VI.1.1 Principios de la OCDE para la validación de modelos QSAR

La OCDE establece 5 principios para facilitar la consideración de un modelo QSAR con fines regulatorios[47], debe presentarse la siguiente información:

1) Una variable respuesta definida. Según este principio, un modelo QSAR debe presentar una variable medible, la cual pueda modelarse, como un efecto fisicoquímico, biológico o ambiental.[47] En este estudio, se empleó como variable respuesta el logaritmo de la permeabilidad aparente (logPapp).
2) Un algoritmo no ambiguo. El algoritmo empleado fue SVM-RF-GBM, el cual se obtuvo por el método de conjunto por apilamiento usando las predicciones

de los modelos SVM, RF y GBM, las cuales se combinaron usando un modelo lineal.

3) Un dominio de aplicabilidad definido. Los modelos QSAR son modelos reduccionistas que están limitados a los tipos de estructuras químicas, propiedades fisicoquímicas y mecanismos de acción para los cuales pueden generar predicciones confiables.[47] En este estudio, se determinó el dominio de aplicabilidad mediante el gráfico de Williams para el conjunto de entrenamiento y prueba, y con valores de apalancamiento para los conjuntos anteriores y la base de datos de productos naturales.
4) Medidas apropiadas de bondad de ajuste, robustez y capacidad predictiva. Según el este principio, es necesario proporcionar información sobre el desempeño determinado por el conjunto de entrenamiento y la capacidad predictiva, determinado por el conjunto de prueba. [47] En este estudio, se emplearon el error cuadrático medio (RMSE) y el coeficiente de determinación (R^2) para evaluar el desempeño del conjunto de entrenamiento y prueba.
5) Una interpretación mecanicista, si es posible. Según el principio 5, se debe evaluar una evaluar la relación entre los descriptores empleados en el modelo y la variable respuesta predicha, y debe documentarse cualquier asociación.[47]

VI.1.2 Interpretación del modelo QSAR

En la comparación de modelos predictivos, se considera un mejor desempeño cuando el RMSE presenta valores bajos y el R^2 presenta valores altos; por el contrario, se considera que el modelo presenta un mal desempeño cuando tiene valores altos de RMSE y valores bajos de R^2. En este estudio, se construyeron 6 modelos QSAR para predecir la permeabilidad aparente en Caco-2 y se escogió al modelo SVM-RF-GBM debido a su mejor capacidad predictiva. Los modelos MLR y PLS presentaron resultados similares en RMSE y R^2. Los modelos SVM, RF, GBM y SVM-RF-GBM presentaron un mejor desempeño. Estos resultados sugieren que la permeabilidad aparente en Caco-2 es predicha por diferentes variables con

relaciones no lineales como se observa en el Anexo 7. El modelo SVM-RF-GBM presenta una capacidad predictiva similar al estudio realizado por Wang, ya que presento un mismo valor de R^2 y un valor RMSE ligeramente menor en este estudio (Tabla 2); a pesar de que se emplearon modelos diferentes.

Los resultados indican que la permeabilidad aparente tiene una explicación multivariada, ya que es predicha por 41 descriptores moleculares. Según la Tabla 3, los 10 mejores descriptores están relacionados con los puentes de hidrógeno, como maxHBint7, maxHBint5, maxHBint9, maxHBint3, Eta_D_epsiD y maxHBd, lo cual coincide con lo mencionado por Wang[19]; asimismo, hay descriptores relacionados con la lipofilicidad como AlogP.

Los índices de estado electrotopológico (E-state)[153] proporcionan información electrónica y topológica de los átomos de una molécula. Estos índices están asociados a determinados tipos de átomo. Se dividen en cuatro grupos: *E-state sums, Atom-type counts, E-state minimum* y *E-state maximum*.[153,154]

VI.1.3 Permeabilidad aparente y absorción intestinal

La absorción intestinal *in vivo* en humanos y logPapp en células Caco-2 presenta una alta correlación;[150] en consecuencia, la permeabilidad aparente en células Caco-2 es un buen predictor de la absorción oral de fármacos por transporte pasiva en humanos.[155]

La permeabilidad aparente puede clasificarse en 3 categorías: Baja, que presenta un logPapp < -6 y una baja absorción (0 – 20 %); Media, logPapp < -5 y logPapp > -6, moderada absorción (20 – 70 %); y Alta, logPapp > -5, alta absorción (70 – 100%).[150]

VI.2 Base de datos de productos naturales de la biodiversidad del Perú

Los recursos naturales han sido utilizados en la medicina tradicional para tratar y prevenir enfermedades durante miles de años. El Perú es uno de los países con mayor biodiversidad en el mundo, que se refleja en la variedad de organismos, como animales y plantas que son fuente de productos naturales con propiedades

terapéuticas[6,27,28]. Sin embargo, existe una falta de información sobre las propiedades farmacológicas y la seguridad de los productos naturales; por ello, surge la importancia de conocer las actividades farmacológicas de los productos naturales.[11,12]

La catalogación de las estructuras químicas obtenidas de los recursos naturales puede servir como punto de partida en el descubrimiento de fármacos que pueden dar paso a análogos estructurales o cabezas de serie.[156]. Entre 1981 y 2015, más del 50% de nuevos fármacos desarrollados, se obtuvieron de productos naturales.[157] Por ello, es importante la separación, purificación y elucidación de extractos de recursos naturales, para obtener el producto natural aislado.[158]

En otros países, se han planteado bases de datos de productos naturales como *The Universal Natural Product Database*[159], *The Natural Product Atlas*[160], $NuBBE_{DB}$[53] BIOFACQUIM[161], COCONUT[162] y SuperNatural 3.0[163], la cuales almacenan la información química del producto natural, como el nombre común y formato SMILES, y la información relacionada al organismo de procedencia, como nombre de la especie, género, familia, y actividad farmacológica asociada. En el Perú, este es el primer esfuerzo en plantear una base de datos de productos naturales, la cual ha recogido la información química del compuesto, la información del organismo de procedencia y de la actividad farmacológica atribuida.

VI.3 Predicción de la permeabilidad aparente en células Caco-2 usando la base de datos de productos naturales mediante los modelos QSAR

En este estudio, se predijo la permeabilidad aparente de 516 productos naturales provenientes de la biodiversidad del Perú empleando el modelo QSAR SVM-RF-GBM; se seleccionaron los 502 compuestos dentro del dominio de aplicabilidad para los análisis posteriores. Se encontró que el 67.33 % (n = 338) de los compuestos presentaban valores de logPapp superiores a -5, lo que indica una alta absorción intestinal[150]. Los compuestos con valores de logPapp entre -6 y -5 se clasificaron como de absorción intestinal media, representando el 22.51 % (n =

113). Por último, los compuestos con valores de logPapp inferiores a -6, se consideraron de absorción intestinal baja, que representan el 10.16 % (n = 51).

La Figura 22 mostró que los compuestos de la vía de alcaloides, ácidos grasos, policétidos, terpenoides, shikimatos y fenilpropanoides tienen principalmente un alto logPapp, lo que sugiere una alta absorción intestinal por transporte pasivo. En la vía de los alcaloides, se observó un alto logPapp en los alcaloides de lisina, pseudoalcaloides y alcaloides de ornitina. Los alcaloides de triptófano mostraron una distribución variada con compuestos de bajo, medio y alto logPapp. En la vía de los shikimatos y fenilpropanoides, los fenilpropanoides y xantonas presentaron un alto logPapp. Los flavonoides, isoflavonoides, lignanos y ácidos fenólicos presentaron una distribución diversa con compuestos de baja, media y alta permeabilidad aparente. Los estilbenos presentaron un medio logPapp. En la vía de los ácidos grasos, los ácidos grasos, ácidos grasos conjugados y ésteres grasos presentaron principalmente un alto logPapp. Las amidas grasas mostraron un bajo, medio y alto logPapp. Finalmente, en la vía de los terpenoides, los esteroides, sesquiterpenoides y apocarotenoides presentaron una alta permeabilidad aparente, mientras que los meroterpenoides, triterpenoides, monoterpenoides y diterpenoides presentaron compuestos con una permeabilidad aparente baja, media y alta.

Los compuestos fenólicos son el grupo más grande de metabolitos secundarios en las plantas, que incluye desde compuestos aromáticos simples hasta complejos. Estos compuestos se obtienen principalmente de la vía de los shikimatos y fenilpropanoides.[164] Los compuestos de la vía de los shikimatos y fenilpropanoides o compuestos fenólicos y los alcaloides presentaron una baja, media y alta permeabilidad aparente. Esta variabilidad en la permeabilidad aparente coincide con la diversidad química de estos grupos.

Los compuestos que tienen la capacidad de atravesar la barrera intestinal deben presentar un equilibrio entre la lipofilicidad y la solubilidad en agua. Esto se debe a que estos compuestos deben estar disueltos en un entorno acuoso para poder atravesar la bicapa lipídica de los enterocitos. Algunos ejemplos de productos naturales que mostraron una alta permeabilidad aparente son α-ionona, R-(-)-

linalool y metilcativato (Anexo 13). Estos compuestos se caracterizan por el predominio de regiones hidrofóbicas, como metilos, ciclo hexanos, alcanos y alquenos, lo que favorece su capacidad de atravesar la bicapa lipídica; asimismo, presentan pocos grupos polares como carbonilos, hidroxilos y ésteres, lo cual les permite formar puentes de hidrógeno con las moléculas de agua y tener cierta solubilidad en medio acuoso.[165]

Por otro lado, algunos compuestos que mostraron una baja permeabilidad aparente como betacianina, luteolina-7-O-glucósido y genistina (Anexo 14), se caracterizan por la presencia de abundantes grupos polares como carboxilo, hidroxilos, éteres y carbonilos, y por pocas regiones hidrofóbicas, lo que le otorga una baja lipofilicidad que impacta en su baja capacidad de atravesar la bicapa lipídica.[165]

VII. CONCLUSIONES

- Se desarrollaron 6 modelos QSAR para predecir la permeabilidad aparente en células Caco-2: MLR, PLS, SVM, RF, GBM y SVM-RF-GBM, donde este último demostró la mejor capacidad predictiva para el conjunto de entrenamiento y prueba.
- Se generó una base de datos de 516 productos naturales de la biodiversidad del Perú clasificados en 6 vías metabólicas: vía de los alcaloides, ácidos grasos, policétidos, terpenoides, shikimatos y fenilpropanoides, y aminoácidos y péptidos que fueron obtenidos de 59 especies que comprendían en su mayoría plantas angiospermas.
- Se predijo la permeabilidad aparente en células Caco-2 de 516 compuestos usando la base de datos de productos naturales mediante el modelo QSAR SVM-RF-GBM. Asimismo, se encontraron 502 compuestos dentro del dominio de aplicabilidad y se clasificaron en 3 categorías: alta permeabilidad, 338 compuestos; moderada permeabilidad, 113 compuestos; y baja permeabilidad, 51 compuestos.

VIII. RECOMENDACIONES

- Añadir descriptores moleculares y ampliar el número de compuestos del conjunto de entrenamiento para mejorar la capacidad predictiva del modelo
- Emplear un equipo de cómputo especializado para disminuir el tiempo de cálculo con algunos algoritmos
- Validar experimentalmente los resultados predichos por el modelo QSAR mediante el ensayo en la línea celular Caco-2
- Emplear otros modelos no lineales debido a que podrían tener una mejor capacidad predictiva para el conjunto de datos

IX. REFERENCIAS BIBLIOGRÁFICAS

1. FDA. Importing Human Drugs. FDA [Internet]. 3 de mayo de 2021 [citado 29 de enero de 2023]; Disponible en: https://www.fda.gov/industry/importing-fda-regulated-products/importing-human-drugs

2. Patrick GL. An introduction to medicinal chemistry. Fifth edition. Oxford: Oxford University Press; 2013. 789 p.

3. Alqahtani MS, Kazi M, Alsenaidy MA, Ahmad MZ. Advances in Oral Drug Delivery. Front Pharmacol. 19 de febrero de 2021;12:618411.

4. Lüllmann H, Mohr K, Hein L. Color atlas of pharmacology. Fifth edition. Stuttgart ; New York: Thieme; 2018.

5. Ekor M. The growing use of herbal medicines: issues relating to adverse reactions and challenges in monitoring safety. Front Pharmacol. 2014;4(177):10.

6. De-la-Cruz H, Vilcapoma G, Zevallos PA. Ethnobotanical study of medicinal plants used by the Andean people of Canta, Lima, Peru. J Ethnopharmacol. mayo de 2007;111(2):284-94.

7. Herrera-Añazco P, Taype-Rondan A, Ortiz PJ, Málaga G, del Carpio-Toia AM, Alvarez-Valdivia MG, et al. Use of medicinal plants in patients with chronic kidney disease from Peru. Complement Ther Med. diciembre de 2019;47:102215.

8. Acosta S, Meléndez C. Catálogo Florístico de Plantas Medicinales Peruanas [Internet]. Lima: Centro Nacional De Salud Intercultural, Instituto Nacional De Salud; 2013. 59 p. Disponible en: https://bvs.ins.gob.pe/insprint/CENSI/catalogo_floristico_plantas_medicinales.pdf

9. Bussmann RW. The Globalization of Traditional Medicine in Northern Peru: From Shamanism to Molecules [Internet]. Vol. 2013, Evidence-Based Complementary and Alternative Medicine. Hindawi; 2013 [citado 11 de noviembre de 2020]. p. e291903. Disponible en: https://www.hindawi.com/journals/ecam/2013/291903/

10. Gonzales de la Cruz M, Baldeón Malpartida S, Beltrán Santiago H, Jullian V, Bourdy G. Hot and cold: Medicinal plant uses in Quechua speaking communities in the high Andes (Callejón de Huaylas, Ancash, Perú). J Ethnopharmacol. septiembre de 2014;155(2):1093-117.

11. Ahmed I, Leach DN, Wohlmuth H, De Voss JJ, Blanchfield JT. Caco-2 Cell Permeability of Flavonoids and Saponins from Gynostemma pentaphyllum: the Immortal Herb. ACS Omega. 20 de agosto de 2020;5(34):21561-9.

12. Firenzuoli F, Gori L. Herbal Medicine Today: Clinical and Research Issues. Evid-Based Complement Altern Med ECAM. septiembre de 2007;4(Suppl 1):37-40.

13. Lanevskij K, Didziapetris R. Physicochemical QSAR Analysis of Passive Permeability Across Caco-2 Monolayers. J Pharm Sci. enero de 2019;108(1):78-86.

14. Hubatsch I, Ragnarsson EGE, Artursson P. Determination of drug permeability and prediction of drug absorption in Caco-2 monolayers. Nat Protoc. septiembre de 2007;2(9):2111-9.

15. Fredlund L, Winiwarter S, Hilgendorf C. In Vitro Intrinsic Permeability: A Transporter-Independent Measure of Caco-2 Cell Permeability in Drug Design and Development. Mol Pharm. mayo de 2017;14(5):1601-9.

16. Over B, Matsson P, Tyrchan C, Artursson P, Doak BC, Foley MA, et al. Structural and conformational determinants of macrocycle cell permeability. Nat Chem Biol. diciembre de 2016;12(12):1065-74.

17. Sherer EC, Verras A, Madeira M, Hagmann WK, Sheridan RP, Roberts D, et al. QSAR Prediction of Passive Permeability in the LLC-PK1 Cell Line: Trends in Molecular Properties and Cross-Prediction of Caco-2 Permeabilities. Mol Inform. abril de 2012;31(3-4):231-45.

18. Wang NN, Dong J, Deng YH, Zhu MF, Wen M, Yao ZJ, et al. ADME Properties Evaluation in Drug Discovery: Prediction of Caco-2 Cell Permeability Using a Combination of NSGA-II and Boosting. J Chem Inf Model. 25 de abril de 2016;56(4):763-73.

19. Wang Y, Chen X. QSPR model for Caco-2 cell permeability prediction using a combination of HQPSO and dual-RBF neural network. RSC Adv. 23 de noviembre de 2020;10(70):42938-52.

20. Groschwitz KR, Hogan SP. Intestinal Barrier Function: Molecular Regulation and Disease Pathogenesis. J Allergy Clin Immunol. julio de 2009;124(1):3-22.

21. Dahlgren D, Lennernäs H. Intestinal Permeability and Drug Absorption: Predictive Experimental, Computational and In Vivo Approaches. Pharmaceutics. 13 de agosto de 2019;11(8):411.

22. Volpe DA. Application of Method Suitability for Drug Permeability Classification. AAPS J. 2 de septiembre de 2010;12(4):670-8.

23. Lea T. Caco-2 Cell Line. En: Verhoeckx K, Cotter P, López-Expósito I, Kleiveland C, Lea T, Mackie A, et al., editores. The Impact of Food Bioactives on Health: in vitro and ex vivo models [Internet]. Cham: Springer International Publishing; 2015 [citado 9 de abril de 2021]. p. 103-11. Disponible en: https://doi.org/10.1007/978-3-319-16104-4_10

24. Ministerio del Ambiente. SEXTO INFORME NACIONAL SOBRE DIVERSIDAD BIOLÓGICA INFORME DE GESTIÓN [Internet]. Primera edición. Lima, Perú; 2019 [citado 14 de abril de 2021]. 105 p. Disponible en: https://cdn.www.gob.pe/uploads/document/file/360830/Informe_de_Gestion_final.pdf

25. Acuña VA. Predicción de la presión de vapor de esteres ftálicos empleados como plastificantes en función de su estructura molecular por descriptores moleculares [Internet] [Tesis de maestría]. [Lima, Perú]: Universidad Nacional Mayor de San Marcos; 2018 [citado 25 de febrero de 2021]. Disponible en: https://cybertesis.unmsm.edu.pe/handle/20.500.12672/9403

26. Rabanal J. Estudio in silico de la reactividad y propiedades fisicoquímicas de aductos de epóxido de eugenol y quinona metilada con glutatión, aminoácidos y poliaminas de Candida albicans [Internet] [Tesis de maestría]. [Lima, Perú]: Universidad Nacional Mayor de San Marcos; 2019 [citado 25 de febrero de 2021]. Disponible en: https://cybertesis.unmsm.edu.pe/bitstream/handle/20.500.12672/11189/Rabanal_sj.pdf?sequence=3&isAllowed=y

27. Bussmann RW, Sharon D. Traditional medicinal plant use in Northern Peru: tracking two thousand years of healing culture. J Ethnobiol Ethnomedicine. 7 de noviembre de 2006;2(1):47.

28. Cragg GM, Newman DJ. Natural products: A continuing source of novel drug leads. Biochim Biophys Acta BBA - Gen Subj. junio de 2013;1830(6):3670-95.

29. Ménard S, Cerf-Bensussan N, Heyman M. Multiple facets of intestinal permeability and epithelial handling of dietary antigens. Mucosal Immunol. mayo de 2010;3(3):247-59.

30. Xu Y, Shrestha N, Préat V, Beloqui A. An overview of in vitro, ex vivo and in vivo models for studying the transport of drugs across intestinal barriers. Adv Drug Deliv Rev. agosto de 2021;175:113795.

31. Bischoff SC, Barbara G, Buurman W, Ockhuizen T, Schulzke JD, Serino M, et al. Intestinal permeability – a new target for disease prevention and therapy. BMC Gastroenterol. diciembre de 2014;14(1):189.

32. Van de Waterbeemd H, Testa B. Drug bioavailability: estimation of solubility, permeability, absorption and bioavailability. 2nd, completely revised ed ed. Weinheim: Wiley-VCH; 2009. (Methods and principles in medicinal chemistry).

33. Dressman JB, Reppas C, editores. Oral Drug Absorption: Prediction and Assessment, Second Edition [Internet]. 2nd ed. CRC Press; 2016 [citado 9 de diciembre de 2021]. (Drugs and the pharmaceutical sciences). Disponible en: https://www.taylorfrancis.com/books/9781420077346

34. Artursson P, Palm K, Luthman K. Caco-2 monolayers in experimental and theoretical predictions of drug transport. Adv Drug Deliv Rev. diciembre de 2012;64:280-9.

35. Estudante M, Morais JG, Soveral G, Benet LZ. Intestinal drug transporters: An overview. Adv Drug Deliv Rev. octubre de 2013;65(10):1340-56.

36. Larregieu CA, Benet LZ. Distinguishing between the Permeability Relationships with Absorption and Metabolism To Improve BCS and BDDCS Predictions in Early Drug Discovery. Mol Pharm. 7 de abril de 2014;11(4):1335-44.

37. Ponce de León-Rodríguez M del C, Guyot JP, Laurent-Babot C. Intestinal *in vitro* cell culture models and their potential to study the effect of food components on intestinal inflammation. Crit Rev Food Sci Nutr. 16 de diciembre de 2019;59(22):3648-66.

38. Dastmalchi S, Hamzeh-Mivehroud M, Sokouti B. Quantitative structure - activity relationship: a practical approach. Boca Raton: CRC Press, Taylor & Francis Group; 2018. 102 p.

39. Neves BJ, Braga RC, Melo-Filho CC, Moreira-Filho JT, Muratov EN, Andrade CH. QSAR-Based Virtual Screening: Advances and Applications in Drug Discovery. Front Pharmacol. 13 de noviembre de 2018;9:1275.

40. Roy K, editor. Advances in QSAR Modeling [Internet]. Cham: Springer International Publishing; 2017 [citado 3 de mayo de 2020]. (Challenges and Advances in Computational Chemistry and Physics; vol. 24). Disponible en: http://link.springer.com/10.1007/978-3-319-56850-8

41. Danishuddin, Khan AU. Descriptors and their selection methods in QSAR analysis: paradigm for drug design. Drug Discov Today. agosto de 2016;21(8):1291-302.

42. Boehmke B, Greenwell BM. Hands-on machine learning with R. Boca Raton: CRC Press; 2019. (Chapman & Hall/CRC the R series).

43. Khan PM, Roy K. Current approaches for choosing feature selection and learning algorithms in quantitative structure–activity relationships (QSAR). Expert Opin Drug Discov. 2 de diciembre de 2018;13(12):1075-89.

44. Yu W, Liu T, Valdez R, Gwinn M, Khoury MJ. Application of support vector machine modeling for prediction of common diseases: the case of diabetes and pre-diabetes. BMC Med Inform Decis Mak. 22 de marzo de 2010;10(1):16.

45. Starmer J. The StatQuest illustrated guide to machine learning!!! Coppell, Texas: StatQuest; 2022.

46. Müller AC, Guido S. Introduction to machine learning with Python: a guide for data scientists. First edition. Sebastopol, CA: O'Reilly Media, Inc; 2016. 376 p.

47. OECD. Guidance Document on the Validation of (Quantitative) Structure-Activity Relationship [(Q)SAR] Models [Internet]. Paris, Francia; 2014. 154 p. Disponible en: https://www.oecd-ilibrary.org/content/publication/9789264085442-en

48. Roy K, Kar S, Das RN. A Primer on QSAR/QSPR Modeling [Internet]. Cham: Springer International Publishing; 2015 [citado 6 de diciembre de 2021]. (SpringerBriefs in Molecular Science). Disponible en: http://link.springer.com/10.1007/978-3-319-17281-1

49. Zhu L, Zhao J, Zhang Y, Zhou W, Yin L, Wang Y, et al. ADME properties evaluation in drug discovery: in silico prediction of blood–brain partitioning. Mol Divers. noviembre de 2018;22(4):979-90.

50. Shi Y. Support vector regression-based QSAR models for prediction of antioxidant activity of phenolic compounds. Sci Rep. diciembre de 2021;11(1):8806.

51. Bernardini S, Tiezzi A, Laghezza Masci V, Ovidi E. Natural products for human health: an historical overview of the drug discovery approaches. Nat Prod Res. 18 de agosto de 2018;32(16):1926-50.

52. Dias DA, Urban S, Roessner U. A Historical Overview of Natural Products in Drug Discovery. Metabolites. junio de 2012;2(2):303-36.

53. Pilon AC, Valli M, Dametto AC, Pinto MEF, Freire RT, Castro-Gamboa I, et al. NuBBEDB: an updated database to uncover chemical and biological information from Brazilian biodiversity. Sci Rep. 3 de agosto de 2017;7(1):7215.

54. Li J, Larregieu CA, Benet LZ. Classification of natural products as sources of drugs according to the biopharmaceutics drug disposition classification system (BDDCS). Chin J Nat Med. diciembre de 2016;14(12):888-97.

55. Kim HW, Wang M, Leber CA, Nothias LF, Reher R, Kang KB, et al. NPClassifier: A Deep Neural Network-Based Structural Classification Tool for Natural Products. J Nat Prod. 26 de noviembre de 2021;84(11):2795-807.

56. Habtemariam S. Introduction to plant secondary metabolites—From biosynthesis to chemistry and antidiabetic action. En: Medicinal Foods as Potential Therapies for Type-2 Diabetes and Associated Diseases [Internet]. Elsevier; 2019 [citado 19 de marzo de 2023]. p. 109-32. Disponible en: https://linkinghub.elsevier.com/retrieve/pii/B9780081029220000067

57. Walker PD, Weir ANM, Willis CL, Crump MP. Polyketide β-branching: diversity, mechanism and selectivity. Nat Prod Rep. 28 de abril de 2021;38(4):723-56.

58. Aslan I, Aslan M. Plasma Polyunsaturated Fatty Acids After Weight Loss Surgery. En: Metabolism and Pathophysiology of Bariatric Surgery [Internet]. Elsevier; 2017 [citado 22 de marzo de 2023]. p. 529-34. Disponible en: https://linkinghub.elsevier.com/retrieve/pii/B9780128040119000583

59. Scott S, Cahoon EB, Busta L. Variation on a theme: the structures and biosynthesis of specialized fatty acid natural products in plants. Plant J. 2022;111(4):954-65.

60. Kurek J. Alkaloids - Their Importance in Nature and Human Life [Internet]. Alkaloids - Their Importance in Nature and Human Life. IntechOpen; 2019 [citado 18 de marzo de 2023]. Disponible en: https://www.intechopen.com/chapters/undefined/chapters/66742

61. Faisal S, Badshah SL, Kubra B, Emwas AH, Jaremko M. Alkaloids as potential antivirals. A comprehensive review. Nat Prod Bioprospecting. 4 de enero de 2023;13(1):4.

62. Averesch NJH, Krömer JO. Metabolic Engineering of the Shikimate Pathway for Production of Aromatics and Derived Compounds—Present and Future Strain Construction Strategies. Front Bioeng Biotechnol. 26 de marzo de 2018;6:32.

63. Dixon RA, Achnine L, Kota P, Liu CJ, Reddy MSS, Wang L. The phenylpropanoid pathway and plant defence—a genomics perspective. Mol Plant Pathol. 2002;3(5):371-90.

64. Dewick PM. Medicinal natural products: a biosynthetic approach. 3rd edition. Chichester, West Sussex, United Kingdom: Wiley, A John Wiley and Sons, Ltd., Publication; 2009. 539 p.

65. Kubinyi H. From Narcosis to Hyperspace: The History of QSAR. Quant Struct-Act Relatsh. octubre de 2002;21(4):348-56.

66. Brown AC, Fraser TR. On the Connection between Chemical Constitution and Physiological Action; with special reference to the Physiological Action of the Salts of the Ammonium Bases derived from Strychnia, Brucia, Thebaia, Codeia, Morphia, and Nicotia. J Anat Physiol. 1868;2(2):224-42.

67. Hammett LP (Louis P. Physical organic chemistry : reaction rates, equilibria, and mechanisms. 1st ed. New York: McGraw-Hill Book Company, Inc.; 1940. (International chemical series).

68. Hansch C, Maloney PP, Fujita T, Muir RM. Correlation of Biological Activity of Phenoxyacetic Acids with Hammett Substituent Constants and Partition Coefficients. Nature. abril de 1962;194(4824):178-80.

69. ChemAxon. Molconvert: Molecule File Converter [Internet]. ChemAxon Ltd.; 2021. Disponible en: https://docs.chemaxon.com/display/docs/molconvert.md

70. Halgren TA, Nachbar RB. Merck molecular force field. IV. conformational energies and geometries for MMFF94. J Comput Chem. abril de 1996;17(5-6):587-615.

71. O'Boyle NM, Banck M, James CA, Morley C, Vandermeersch T, Hutchison GR. Open Babel: An open chemical toolbox. J Cheminformatics. diciembre de 2011;3(1):33.

72. Wang J, Wolf RM, Caldwell JW, Kollman PA, Case DA. Development and testing of a general amber force field. J Comput Chem. 15 de julio de 2004;25(9):1157-74.

73. Yap CW. PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints. J Comput Chem. 2011;32(7):1466-74.

74. Mauri A. alvaDesc: A Tool to Calculate and Analyze Molecular Descriptors and Fingerprints. En: Roy K, editor. Ecotoxicological QSARs [Internet]. New York, NY: Springer US; 2020 [citado 10 de marzo de 2021]. p. 801-20. (Methods in Pharmacology and Toxicology). Disponible en: http://link.springer.com/10.1007/978-1-0716-0150-1_32

75. Okuyama E, Umeyama K, Ohmori S, Yamazaki M, Satake M. Pharmacologically Active Components from a Peruvian Medicinal Plant Huira-Huira (*Culcitium canescens* H. & B.). Chem Pharm Bull (Tokyo). 1994;42(10):2183-6.

76. Fuchino H, Koide T, Takahashi M, Sekita S, Satake M. New Sesquiterpene Lactones from *Elephantopus mollis* and Their Leishmanicidal Activities. Planta Med. octubre de 2001;67(7):647-53.

77. Kang TH, Matsumoto K, Tohda M, Murakami Y, Takayama H, Kitajima M, et al. Pteropodine and isopteropodine positively modulate the function of rat muscarinic M1 and 5-HT2 receptors expressed in *Xenopus oocyte*. Eur J Pharmacol. mayo de 2002;444(1-2):39-45.

78. Tincusi BM, Jiménez IA, Bazzocchi IL, Moujir LM, Mamani ZA, Barroso JP, et al. Antimicrobial Terpenoids from the Oleoresin of the Peruvian Medicinal Plant *Copaifera paupera*. Planta Med. septiembre de 2002;68(9):808-12.

79. Hegde VR, Pu H, Patel M, Das PR, Butkiewicz N, Arreaza G, et al. Two Antiviral Compounds from the Plant Stylogne cauliflora as Inhibitors of HCV NS3 Protease. ChemInform [Internet]. 2 de diciembre de 2003 [citado 16 de

diciembre de 2021];34(48). Disponible en: https://onlinelibrary.wiley.com/doi/10.1002/chin.200348206

80. Hegde VR, Pu H, Patel M, Black T, Soriano A, Zhao W, et al. Two new bacterial DNA primase inhibitors from the plant *Polygonum cuspidatum*. Bioorg Med Chem Lett. mayo de 2004;14(9):2275-7.

81. Hegde VR, Pu H, Patel M, Das PR, Strizki J, Gullo VP, et al. Three new compounds from the plant *Lippia alva* as inhibitors of chemokine receptor 5 (CCR5). Bioorg Med Chem Lett. noviembre de 2004;14(21):5339-42.

82. Heitzman ME, Neto CC, Winiarz E, Vaisberg AJ, Hammond GB. Ethnobotany, phytochemistry and pharmacology of *Uncaria* (Rubiaceae). Phytochemistry. 1 de enero de 2005;66(1):5-29.

83. Aguayo L, Guzman L, Perez C, Aguayo L, Silva M, Becerra J, et al. Historical and Current Perspectives of Neuroactive Compounds Derived from Latin America. Mini-Rev Med Chem. 1 de septiembre de 2006;6(9):997-1008.

84. Rojas R, Bustamante B, Ventosilla P, Fernádez I, Caviedes L, Gilman RH, et al. Larvicidal, Antimycobacterial and Antifungal Compounds from the Bark of the Peruvian Plant *Swartzia polyphylla* DC. Chem Pharm Bull (Tokyo). 2006;54(2):278-9.

85. Aguiar CL, Baptista AS, Alencar SM, Haddad R, Eberlin MN. Analysis of isoflavonoids from leguminous plant extracts by RPHPLC/DAD and electrospray ionization mass spectrometry. Int J Food Sci Nutr. enero de 2007;58(2):116-24.

86. Castillo D, Arevalo J, Herrera F, Ruiz C, Rojas R, Rengifo E, et al. Spirolactone iridoids might be responsible for the antileishmanial activity of a Peruvian traditional remedy made with *Himatanthus sucuuba* (Apocynaceae). J Ethnopharmacol. junio de 2007;112(2):410-4.

87. Mesa-Siverio D, Machín RP, Estévez-Braun A, Ravelo ÁngelG, Lock O. Structure and estrogenic activity of new lignans from *Iryanthera lancifolia*. Bioorg Med Chem. 15 de marzo de 2008;16(6):3387-94.

88. Gonzales GF, Gonzales-Castañeda C. The Methyltetrahydro-β-Carbolines in Maca (*Lepidium meyenii*). Evid Based Complement Alternat Med. 2009;6(3):315-6.

89. Kawano M, Otsuka M, Umeyama K, Yamazaki M, Shiota T, Satake M, et al. Anti-inflammatory and analgesic components from “hierba santa,” a traditional medicine in Peru. J Nat Med. abril de 2009;63(2):147-58.

90. Aponte J, Yang H, Vaisberg A, Castillo D, Málaga E, Verástegui M, et al. Cytotoxic and Anti-infective Sesquiterpenes Present in *Plagiochila disticha*

(Plagiochilaceae) and *Ambrosia peruviana* (Asteraceae). Planta Med. mayo de 2010;76(07):705-7.

91. García Giménez D, García Prado E, Sáenz Rodríguez T, Fernández Arche A, De la Puerta R. Cytotoxic Effect of the Pentacyclic Oxindole Alkaloid Mitraphylline Isolated from *Uncaria tomentosa* Bark on Human Ewing's Sarcoma and Breast Cancer Cell Lines. Planta Med. enero de 2010;76(02):133-6.

92. Aponte J, Jin Z, Vaisberg A, Castillo D, Málaga E, Lewis W, et al. Cytotoxic and Anti-infective Phenolic Compounds Isolated from *Mikania decora* and *Cremastosperma microcarpum*. Planta Med. septiembre de 2011;77(14):1597-9.

93. Fuchino H, Kiuchi F, Yamanaka A, Obu A, Wada H, Mori-Yasumoto K, et al. New Leishmanicidal Stilbenes from a Peruvian Folk Medicine, *Lonchocarpus nicou*. Chem Pharm Bull (Tokyo). 2013;61(9):979-82.

94. Leuner O, Havlik J, Budesinsky M, Vrkoslav V, Chu J, Bradshaw TD, et al. Cytotoxic Constituents of *Pachyrhizus Tuberosus* from Peruvian Amazon. Nat Prod Commun. octubre de 2013;8(10):1934578X1300801.

95. Wu H, Kelley CJ, Pino-Figueroa A, Vu HD, Maher TJ. Macamides and their synthetic analogs: Evaluation of in vitro FAAH inhibition. Bioorg Med Chem. septiembre de 2013;21(17):5188-97.

96. Baldera-Aguayo PA. Phytochemical study of *Echinopsis peruviana*. Rev Soc Quím Perú. 2014;9.

97. Hajdu Z, Nicolussi S, Rau M, Lorántfy L, Forgo P, Hohmann J, et al. Identification of Endocannabinoid System-Modulating *N* -Alkylamides from *Heliopsis helianthoides* var. *scabra* and *Lepidium meyenii*. J Nat Prod. 25 de julio de 2014;77(7):1663-9.

98. Reina M, Ruiz-Mesia L, Ruiz-Mesia W, Sosa-Amay FE, Arevalo-Encinas L, González-Coloma A, et al. Antiparasitic Indole Alkaloids from *Aspidosperma desmanthum* and *A. spruceanum* from the Peruvian Amazonia. Nat Prod Commun. agosto de 2014;9(8):1934578X1400900.

99. Abderrahim F, Huanatico E, Segura R, Arribas S, Gonzalez MC, Condezo-Hoyos L. Physical features, phenolic compounds, betalains and total antioxidant capacity of coloured quinoa seeds (*Chenopodium quinoa* Willd.) from Peruvian Altiplano. Food Chem. septiembre de 2015;183:83-90.

100. Colegate SM, Boppré M, Monzón J, Betz JM. Pro-toxic dehydropyrrolizidine alkaloids in the traditional Andean herbal medicine "asmachilca". J Ethnopharmacol. agosto de 2015;172:179-94.

101. Esparza E, Hadzich A, Kofer W, Mithöfer A, Cosio EG. Bioactive maca (*Lepidium meyenii*) alkamides are a result of traditional Andean postharvest drying practices. Phytochemistry. agosto de 2015;116:138-48.

102. Girardi C, Fabre N, Paloque L, Ramadani AP, Benoit-Vical F, González-Aspajo G, et al. Evaluation of antiplasmodial and antileishmanial activities of herbal medicine *Pseudelephantopus spiralis* (Less.) Cronquist and isolated hirsutinolide-type sesquiterpenoids. J Ethnopharmacol. julio de 2015;170:167-74.

103. Patel K, Ruiz C, Calderon R, Marcelo M, Rojas R. Characterisation of volatile profiles in 50 native Peruvian chili pepper using solid phase microextraction–gas chromatography mass spectrometry (SPME–GCMS). Food Res Int. noviembre de 2016;89:471-5.

104. Schmeda-Hirschmann G, Quispe C, Arana GV, Theoduloz C, Urra FA, Cárdenas C. Antiproliferative activity and chemical composition of the venom from the Amazonian toad *Rhinella marina* (Anura: Bufonidae). Toxicon. octubre de 2016;121:119-29.

105. Boniface PK, Baptista Ferreira S, Roland Kaiser C. Current state of knowledge on the traditional uses, phytochemistry, and pharmacology of the genus *Hymenaea*. J Ethnopharmacol. julio de 2017;206:193-223.

106. Feuereisen MM, Zimmermann BF, Schulze-Kaysers N, Schieber A. Differentiation of Brazilian Peppertree (*Schinus terebinthifolius* Raddi) and Peruvian Peppertree (*Schinus molle* L.) Fruits by UHPLC–UV–MS Analysis of Their Anthocyanin and Biflavonoid Profiles. J Agric Food Chem. 5 de julio de 2017;65(26):5330-8.

107. Gálvez Ranilla L, Christopher A, Sarkar D, Shetty K, Chirinos R, Campos D. Phenolic Composition and Evaluation of the Antimicrobial Activity of Free and Bound Phenolic Fractions from a Peruvian Purple Corn (*Zea mays* L.) Accession: Antimicrobial activity of purple corn…. J Food Sci. diciembre de 2017;82(12):2968-76.

108. Guillen Quispe Y, Hwang S, Wang Z, Zuo G, Lim S. Screening In Vitro Targets Related to Diabetes in Herbal Extracts from Peru: Identification of Active Compounds in *Hypericum laricifolium* Juss. by Offline High-Performance Liquid Chromatography. Int J Mol Sci. 24 de noviembre de 2017;18(12):2512.

109. Linares-Otoya L, Linares-Otoya V, Armas-Mantilla L, Blanco-Olano C, Crüsemann M, Ganoza-Yupanqui M, et al. Diversity and Antimicrobial Potential of Predatory Bacteria from the Peruvian Coastline. Mar Drugs. 12 de octubre de 2017;15(10):308.

110. Quispe Y, Hwang S, Wang Z, Lim S. Screening of Peruvian Medicinal Plants for Tyrosinase Inhibitory Properties: Identification of Tyrosinase Inhibitors in *Hypericum laricifolium* Juss. Molecules. 4 de marzo de 2017;22(3):402.

111. Xu YM, Wijeratne EMK, Babyak AL, Marks HR, Brooks AD, Tewary P, et al. Withanolides from Aeroponically Grown *Physalis peruviana* and Their Selective Cytotoxicity to Prostate Cancer and Renal Carcinoma Cells. J Nat Prod. 28 de julio de 2017;80(7):1981-91.

112. Morales-Soriano E, Kebede B, Ugás R, Grauwet T, Van Loey A, Hendrickx M. Flavor characterization of native Peruvian chili peppers through integrated aroma fingerprinting and pungency profiling. Food Res Int. julio de 2018;109:250-9.

113. Stivers N, Islam A, Reyes-Reyes E, Casson L, Aponte J, Vaisberg A, et al. Plagiochiline A Inhibits Cytokinetic Abscission and Induces Cell Death. Molecules. 12 de junio de 2018;23(6):1418.

114. Wang S, Zhu F, Kakuda Y. Sacha inchi (*Plukenetia volubilis* L.): Nutritional composition, biological activity, and uses. Food Chem. noviembre de 2018;265:316-28.

115. Alves NSF, Setzer WN, da Silva JKR. The chemistry and biological activities of *Peperomia pellucida* (Piperaceae): A critical review. J Ethnopharmacol. marzo de 2019;232:90-102.

116. Carlos Castro J, Dylan Maddox J, Cobos M, Diana Paredes J, Jhoao Fasabi A, Vargas-Arana G, et al. Medicinal Plants of the Peruvian Amazon: Bioactive Phytochemicals, Mechanisms of Action, and Biosynthetic Pathways. En: Perveen S, Al-Taweel A, editores. Pharmacognosy - Medicinal Plants [Internet]. IntechOpen; 2019 [citado 16 de diciembre de 2021]. Disponible en: https://www.intechopen.com/books/pharmacognosy-medicinal-plants/medicinal-plants-of-the-peruvian-amazon-bioactive-phytochemicals-mechanisms-of-action-and-biosynthet

117. Han Y, Chi J, Zhang M, Zhang R, Fan S, Huang F, et al. Characterization of saponins and phenolic compounds: antioxidant activity and inhibitory effects on α-glucosidase in different varieties of colored quinoa (*Chenopodium quinoa* Willd). Biosci Biotechnol Biochem. 2 de noviembre de 2019;83(11):2128-39.

118. Hwang SH, Kim HY, Guillen Quispe YN, Wang Z, Zuo G, Lim SS. Aldose Reductase, Protein Glycation Inhibitory and Antioxidant of Peruvian Medicinal Plants: The Case of *Tanacetum parthenium* L. and Its Constituents. Molecules. 25 de mayo de 2019;24(10):2010.

119. Radice M, Tasambay A, Pérez A, Diéguez-Santana K, Sacchetti G, Buso P, et al. Ethnopharmacology, phytochemistry and pharmacology of the genus

Hedyosmum (Chlorantaceae): A review. J Ethnopharmacol. noviembre de 2019;244:111932.

120. Tauchen J, Huml L, Bortl L, Doskocil I, Jarosova V, Marsik P, et al. Screening of medicinal plants traditionally used in Peruvian Amazon for *in vitro* antioxidant and anticancer potential. Nat Prod Res. 17 de septiembre de 2019;33(18):2718-21.

121. Zhang Y, Zhang S, Fan W, Duan M, Han Y, Li H. Identification of volatile compounds and odour activity values in quinoa porridge by gas chromatography–mass spectrometry. J Sci Food Agric. junio de 2019;99(8):3957-66.

122. Zhong JL, Yan H, Xu HD, Muhammad N, Yan WD. Preparation from *Lepidium meyenii* Walpers using high-speed countercurrent chromatography and thermal stability of macamides in air at various temperatures. J Pharm Biomed Anal. febrero de 2019;164:768-76.

123. Cornwell W, FitzJohn R, Pennell M. taxonlookup: A dynamically-updating versioned taxonomic resource for vascular plants [Internet]. 2016. Disponible en: https://doi.org/10.5281/zenodo.839589

124. R Core Team. R: A Language and Environment for Statistical Computing [Internet]. Vienna, Austria: R Foundation for Statistical Computing; 2022. Disponible en: http://www.R-project.org/

125. RStudio Team. RStudio: Integrated Development Environment for R [Internet]. Boston, MA: RStudio, Inc.; 2022. Disponible en: http://www.rstudio.com/

126. Wickham H, Averick M, Bryan J, Chang W, McGowan LD, François R, et al. Welcome to the tidyverse. J Open Source Softw. 2019;4(43):1686.

127. Wickham H, François R, Henry L, Müller K. dplyr: A Grammar of Data Manipulation [Internet]. 2020. Disponible en: https://CRAN.R-project.org/package=dplyr

128. Wickham H, Hester J. readr: Read Rectangular Text Data [Internet]. 2021. Disponible en: https://CRAN.R-project.org/package=readr

129. Wickham H. stringr: Simple, Consistent Wrappers for Common String Operations [Internet]. 2019. Disponible en: https://CRAN.R-project.org/package=stringr

130. Henry L, Wickham H. purrr: Functional Programming Tools [Internet]. 2020. Disponible en: https://CRAN.R-project.org/package=purrr

131. Müller K, Wickham H. tibble: Simple Data Frames [Internet]. 2021. Disponible en: https://CRAN.R-project.org/package=tibble

132. Lê S, Josse J, Husson F. FactoMineR: A Package for Multivariate Analysis. J Stat Softw. 2008;25(1):1-18.

133. Kuhn M, Wickham H. recipes: Preprocessing and Feature Engineering Steps for Modeling [Internet]. 2022. Disponible en: https://CRAN.R-project.org/package=recipes

134. Gotti M, Kuhn M. applicable: A Compilation of Applicability Domain Methods [Internet]. 2022. Disponible en: https://CRAN.R-project.org/package=applicable

135. Kuhn M. caret: Classification and Regression Training [Internet]. 2022. Disponible en: https://CRAN.R-project.org/package=caret

136. Deane-Mayer ZA, Knowles JE. caretEnsemble: Ensembles of Caret Models [Internet]. 2019. Disponible en: https://CRAN.R-project.org/package=caretEnsemble

137. Mevik BH, Wehrens R, Liland KH. pls: Partial Least Squares and Principal Component Regression [Internet]. 2020. Disponible en: https://CRAN.R-project.org/package=pls

138. Meyer D, Dimitriadou E, Hornik K, Weingessel A, Leisch F. e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien [Internet]. 2020. Disponible en: https://CRAN.R-project.org/package=e1071

139. Liaw A, Wiener M. Classification and Regression by randomForest. R News. 2002;2(3):18-22.

140. Greenwell B, Boehmke B, Cunningham J, Developers GBM. gbm: Generalized Boosted Regression Models [Internet]. 2020. Disponible en: https://CRAN.R-project.org/package=gbm

141. Wickham H. ggplot2: Elegant Graphics for Data Analysis [Internet]. Springer-Verlag New York; 2016. Disponible en: https://ggplot2.tidyverse.org

142. Kassambara A, Mundt F. factoextra: Extract and Visualize the Results of Multivariate Data Analyses [Internet]. 2020. Disponible en: https://CRAN.R-project.org/package=factoextra

143. Auguie B. gridExtra: Miscellaneous Functions for «Grid» Graphics [Internet]. 2017. Disponible en: https://CRAN.R-project.org/package=gridExtra

144. Sarkar D. Lattice: Multivariate Data Visualization with R [Internet]. New York: Springer; 2008. Disponible en: http://lmdvr.r-forge.r-project.org

145. Kassambara A. ggpubr: «ggplot2» Based Publication Ready Plots [Internet]. 2020. Disponible en: https://CRAN.R-project.org/package=ggpubr

146. Taskesen E, Verver O. D3Blocks: A Python Package to create interactive d3js visualizations [Internet]. 2022. Disponible en: https://d3blocks.github.io/d3blocks

147. Van Rossum G. The Python Library Reference, release 3.8.2. Python Software Foundation; 2020.

148. ChemAxon. MarvinSketch [Internet]. ChemAxon Ltd.; 2021. Disponible en: http://www.chemaxon.com/products/marvin/marvinsketch/

149. Instituto de Investigaciones de la Amazonía Peruana. Centro de Alto Rendimiento Computacional de la Amazonia Peruana [Internet]. 2017 [citado 6 de diciembre de 2021]. Disponible en: http://www.iiap.gob.pe/web/manati.aspx/

150. Yee S. In Vitro Permeability Across Caco-2 Cells (Colonic) Can Predict In Vivo (Small Intestinal) Absorption in Man—Fact or Myth. Pharm Res. 1 de junio de 1997;14(6):763-6.

151. Lipinski CA, Lombardo F, Dominy BW, Feeney PJ. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settingsq. Adv Drug Deliv Rev. 2001;24.

152. Veber DF, Johnson SR, Cheng HY, Smith BR, Ward KW, Kopple KD. Molecular Properties That Influence the Oral Bioavailability of Drug Candidates. J Med Chem. 1 de junio de 2002;45(12):2615-23.

153. Hall LH, Mohney B, Kier LB. The Electrotopological State: An Atom Index for QSAR. Quant Struct-Act Relatsh. 1991;10(1):43-51.

154. Hall LH, Kier LB. Electrotopological State Indices for Atom Types: A Novel Combination of Electronic, Topological, and Valence State Information. J Chem Inf Comput Sci. 1 de noviembre de 1995;35(6):1039-45.

155. Pade V, Stavchansky S. Link between drug absorption solubility and permeability measurements in Caco-2 cells. J Pharm Sci. diciembre de 1998;87(12):1604-7.

156. Gómez-García A, Medina-Franco JL. Progress and Impact of Latin American Natural Product Databases. Biomolecules. septiembre de 2022;12(9):1202.

157. Newman DJ, Cragg GM. Natural Products as Sources of New Drugs from 1981 to 2014. J Nat Prod. 25 de marzo de 2016;79(3):629-61.

158. Zhang QW, Lin LG, Ye WC. Techniques for extraction and isolation of natural products: a comprehensive review. Chin Med. 17 de abril de 2018;13(1):20.

159. Gu J, Gui Y, Chen L, Yuan G, Lu HZ, Xu X. Use of Natural Products as Chemical Library for Drug Discovery and Network Pharmacology. PLOS ONE. 25 de abril de 2013;8(4):e62839.

160. van Santen JA, Poynton EF, Iskakova D, McMann E, Alsup TA, Clark TN, et al. The Natural Products Atlas 2.0: a database of microbially-derived natural products. Nucleic Acids Res. 7 de enero de 2022;50(D1):D1317-23.

161. Pilón-Jiménez BA, Saldívar-González FI, Díaz-Eufracio BI, Medina-Franco JL. BIOFACQUIM: A Mexican Compound Database of Natural Products. Biomolecules. 17 de enero de 2019;9(1):31.

162. Sorokina M, Merseburger P, Rajan K, Yirik MA, Steinbeck C. COCONUT online: Collection of Open Natural Products database. J Cheminformatics. 10 de enero de 2021;13(1):2.

163. Gallo K, Kemmler E, Goede A, Becker F, Dunkel M, Preissner R, et al. SuperNatural 3.0—a database of natural products and natural product-based derivatives. Nucleic Acids Res. 6 de enero de 2023;51(D1):D654-9.

164. Marchica A, Cotrozzi L, Detti R, Lorenzini G, Pellegrini E, Petersen M, et al. The Biosynthesis of Phenolic Compounds Is an Integrated Defence Mechanism to Prevent Ozone Injury in Salvia officinalis. Antioxidants. 14 de diciembre de 2020;9(12):1274.

165. Foye WO, Lemke TL, Williams DA, editores. Foye's principles of medicinal chemistry. 7th ed. Philadelphia: Wolters Kluwer Health/Lippincott Williams & Wilkins; 2013. 1500 p.

X. ANEXOS

Anexo 1. Operacionalización de variables

VARIABLE	TIPO DE VARIABLE	DEFINICIÓN CONCEPTUAL	DEFINICIÓN OPERACIONAL	DIMENSIONES	INDICADOR	UNIDAD DE MEDIDA	OBSERVACIONES
logPapp	Cuantitativa	Es el logaritmo de la cantidad de compuesto transportado por tiempo.	(1) Búsqueda bibliográfica del logPapp experimental (2) Predicción del logPapp mediante el modelo QSAR	-	-	adimensional	-logPapp es adimensional debido a que es un logaritmo.
Descriptores moleculares	Cuantitativa	Son las características químicas de una molécula expresada en forma numérica.	Propiedad estructural o fisicoquímica de una molécula calculada a partir de su información estructural	Se emplean miles de descriptores moleculares. Se pueden agrupar en 5 tipos: topológicos, geométricos, termodinámicos, electrónicos y constitucionales	-	Los descriptores moleculares presentan diferentes medidas.	-Se calculan computacionalmente. -Cada descriptor molecular tiene una unidad diferente o es adimensional.

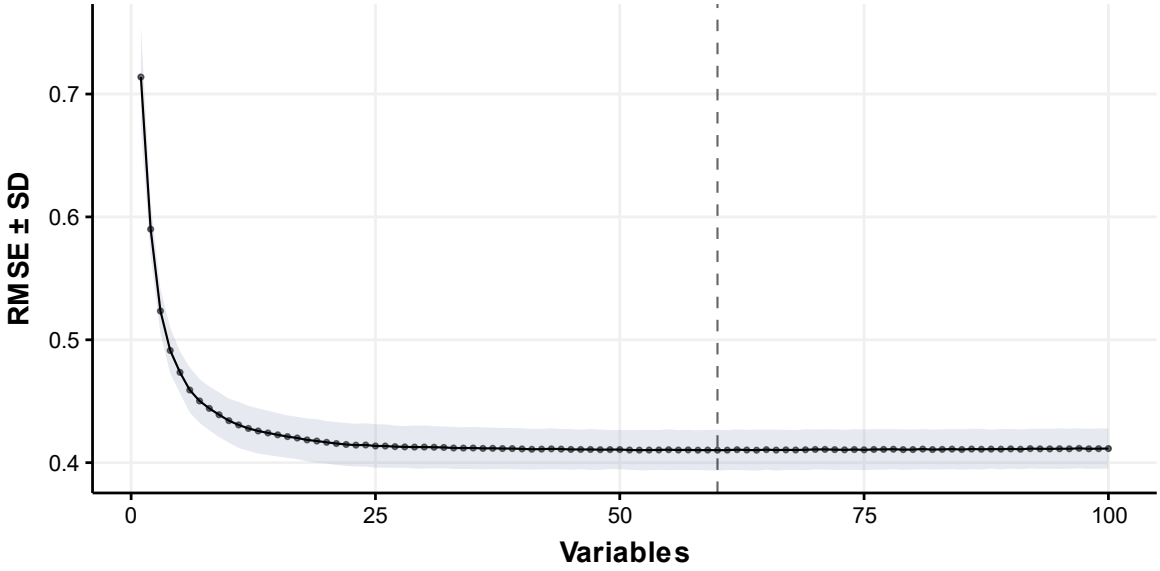
Anexo 2A. Ejemplo de un código escrito en lenguaje R.

```r
# Cargar librerías
library(tidyverse)
library(caret)
# Cargar datos del modelamiento
data_modeling <- read.xlsx("data_modeling.xlsx")

# División del conjunto de datos
index <- createDataPartition(data_modeling$logPapp, p = 0.8, list = FALSE)
Train <- data_modeling[,index]
Test <- data_modeling[,-index]

# Preprocesamiento
library(recipes)
preprocesamiento <- recipe(logPapp~., data = Train) %>%
  step_medianimpute() %>%
  step_nzv() %>%
  step_corr(threshold = 0.8) %>%
  prep()

Train_preprocesado <- bake(preprocesamiento)
Test_preprocesado <- bake(preprocesamiento, Test)

# Selección de características
# A. Selección recursiva de características (RFE)
rfe_train <- rfe(logPapp~., data = Train_preprocesado)

predictores_seleccionados_rfe <- predictors(rfe_train)

# B. Algoritmo genético (GA)
rfe_train <- gafs(y = Train_preprocesado[,"logPapp"],
                  x = Train_preprocesado[,predictores_seleccionados_rfe],
                  iters = 100)

predictores_seleccionados_ga <- rfe_train$optVariables
```

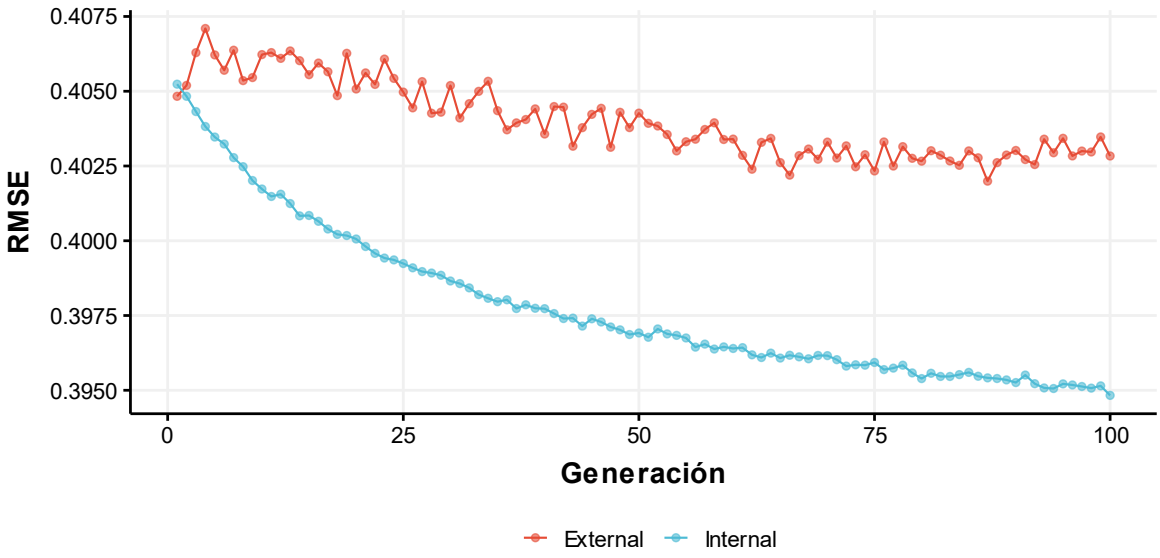
Anexo 2B. Ejemplo de un código escrito en lenguaje R.

```r
# Modelamiento
data_final_modelamiento <- Train_preprocesado %>%
  select(logPapp, predictores_seleccionados_ga)

# Validación cruzada
CV5 <- trainControl(method = "repeatedcv", number = 5, repeats = 5)

# Modelo random forest

RF <- train(logPapp~., data = data_final_modelamiento,
            method = "rf", # Random Forest
            trControl = CV5)

# Descriptores moleculares de base productos naturales
descriptores_NP <- read.xlsx("base_datos_productos_naturales.xlsx")

# Predecir logPapp de la base de datos de productos naturales
logPapp_predecido_NP <- predict(RF, descriptores_NP)
```

Anexo 3. Gráfico de la selección de características recursivas (RFE).



Anexo 4. Gráfico de la selección de características mediante algoritmo genético (GA-RF).



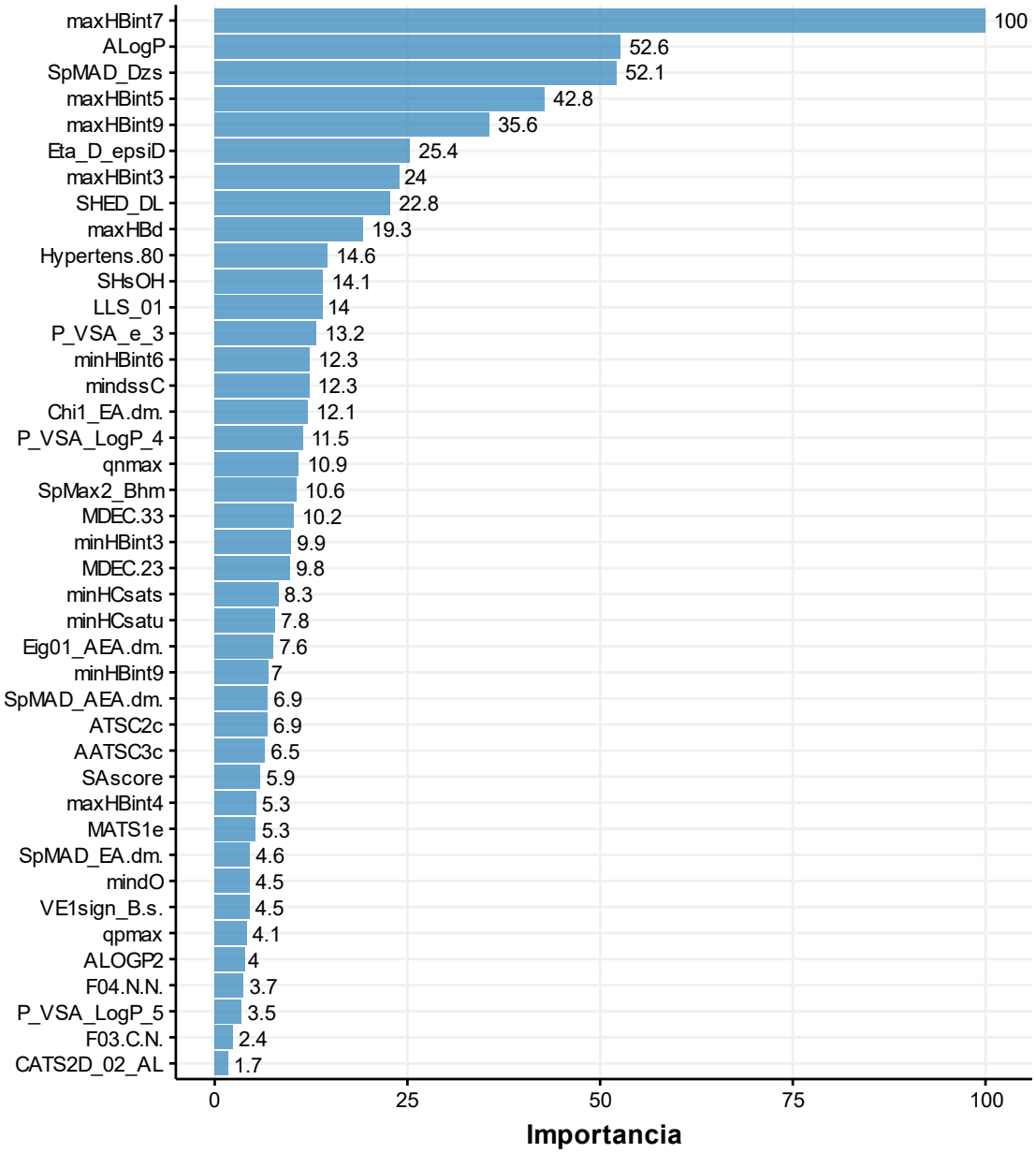
Anexo 5A. Valores del gráfico de RFE

Variables	RMSE	R^2	RMSESD	R^2SD	Variables	RMSE	R^2	RMSESD	R^2SD
1	0.71384	0.21654	0.04133	0.05458	**51**	0.41030	0.73067	0.01616	0.02473
2	0.58998	0.43293	0.02182	0.04483	**52**	0.41019	0.73094	0.01638	0.02493
3	0.52341	0.55223	0.01817	0.03726	**53**	0.41025	0.73087	0.01643	0.02538
4	0.49128	0.60625	0.01862	0.03574	**54**	0.41034	0.73078	0.01622	0.02481
5	0.47348	0.63578	0.01698	0.03383	**55**	0.41054	0.73058	0.01663	0.02516
6	0.45911	0.65564	0.01840	0.03181	**56**	0.41028	0.73111	0.01621	0.02501
7	0.45023	0.66952	0.01780	0.02949	**57**	0.41039	0.73093	0.01636	0.02493
8	0.44415	0.67910	0.01783	0.02970	**58**	0.41017	0.73128	0.01618	0.02492
9	0.43898	0.68596	0.01823	0.03006	**59**	0.41023	0.73131	0.01645	0.02485
10	0.43418	0.69336	0.01792	0.02876	**60**	0.41013	0.73146	0.01660	0.02528
11	0.43069	0.69875	0.01876	0.02920	**61**	0.41019	0.73151	0.01620	0.02491
12	0.42786	0.70261	0.01829	0.02782	**62**	0.41055	0.73098	0.01654	0.02539
13	0.42580	0.70591	0.01841	0.02811	**63**	0.41029	0.73136	0.01638	0.02516
14	0.42418	0.70859	0.01801	0.02733	**64**	0.41015	0.73167	0.01674	0.02556
15	0.42265	0.71055	0.01766	0.02701	**65**	0.41058	0.73122	0.01631	0.02495
16	0.42122	0.71286	0.01730	0.02628	**66**	0.41021	0.73174	0.01669	0.02546
17	0.42006	0.71466	0.01722	0.02630	**67**	0.41040	0.73157	0.01634	0.02491
18	0.41853	0.71662	0.01742	0.02654	**68**	0.41029	0.73173	0.01624	0.02496
19	0.41761	0.71805	0.01774	0.02687	**69**	0.41050	0.73136	0.01668	0.02554
20	0.41654	0.71976	0.01732	0.02611	**70**	0.41073	0.73118	0.01625	0.02483
21	0.41558	0.72095	0.01748	0.02608	**71**	0.41077	0.73128	0.01643	0.02487
22	0.41480	0.72227	0.01750	0.02572	**72**	0.41060	0.73149	0.01654	0.02527
23	0.41424	0.72311	0.01736	0.02582	**73**	0.41046	0.73178	0.01626	0.02495
24	0.41436	0.72277	0.01767	0.02623	**74**	0.41066	0.73157	0.01657	0.02521
25	0.41357	0.72413	0.01771	0.02625	**75**	0.41045	0.73182	0.01654	0.02532
26	0.41350	0.72427	0.01762	0.02593	**76**	0.41073	0.73150	0.01649	0.02520
27	0.41304	0.72490	0.01724	0.02548	**77**	0.41079	0.73152	0.01651	0.02497
28	0.41283	0.72526	0.01695	0.02517	**78**	0.41091	0.73132	0.01627	0.02468
29	0.41267	0.72561	0.01750	0.02565	**79**	0.41060	0.73183	0.01648	0.02520
30	0.41270	0.72559	0.01759	0.02575	**80**	0.41067	0.73180	0.01628	0.02501
31	0.41255	0.72585	0.01717	0.02548	**81**	0.41104	0.73124	0.01631	0.02507
32	0.41238	0.72630	0.01702	0.02539	**82**	0.41065	0.73194	0.01628	0.02485
33	0.41192	0.72689	0.01710	0.02524	**83**	0.41082	0.73176	0.01631	0.02481
34	0.41186	0.72712	0.01699	0.02523	**84**	0.41101	0.73139	0.01633	0.02520
35	0.41193	0.72723	0.01715	0.02552	**85**	0.41071	0.73192	0.01646	0.02534
36	0.41161	0.72765	0.01713	0.02551	**86**	0.41108	0.73145	0.01649	0.02519
37	0.41168	0.72777	0.01683	0.02501	**87**	0.41090	0.73171	0.01655	0.02550
38	0.41149	0.72814	0.01674	0.02516	**88**	0.41101	0.73168	0.01591	0.02479
39	0.41145	0.72805	0.01673	0.02507	**89**	0.41096	0.73178	0.01645	0.02532
40	0.41119	0.72864	0.01678	0.02507	**90**	0.41123	0.73135	0.01617	0.02496
41	0.41089	0.72918	0.01656	0.02480	**91**	0.41089	0.73196	0.01618	0.02492
42	0.41101	0.72894	0.01650	0.02504	**92**	0.41140	0.73130	0.01612	0.02522
43	0.41125	0.72868	0.01670	0.02536	**93**	0.41126	0.73140	0.01636	0.02521
44	0.41103	0.72918	0.01641	0.02516	**94**	0.41123	0.73155	0.01633	0.02540
45	0.41076	0.72955	0.01658	0.02526	**95**	0.41135	0.73136	0.01647	0.02553
46	0.41073	0.72973	0.01611	0.02458	**96**	0.41129	0.73150	0.01634	0.02556
47	0.41074	0.72985	0.01666	0.02514	**97**	0.41162	0.73102	0.01628	0.02515
48	0.41056	0.73003	0.01642	0.02533	**98**	0.41127	0.73166	0.01634	0.02509
49	0.41064	0.73011	0.01586	0.02433	**99**	0.41139	0.73144	0.01647	0.02538
50	0.41066	0.73024	0.01595	0.02448	**100**	0.41146	0.73142	0.01621	0.02513

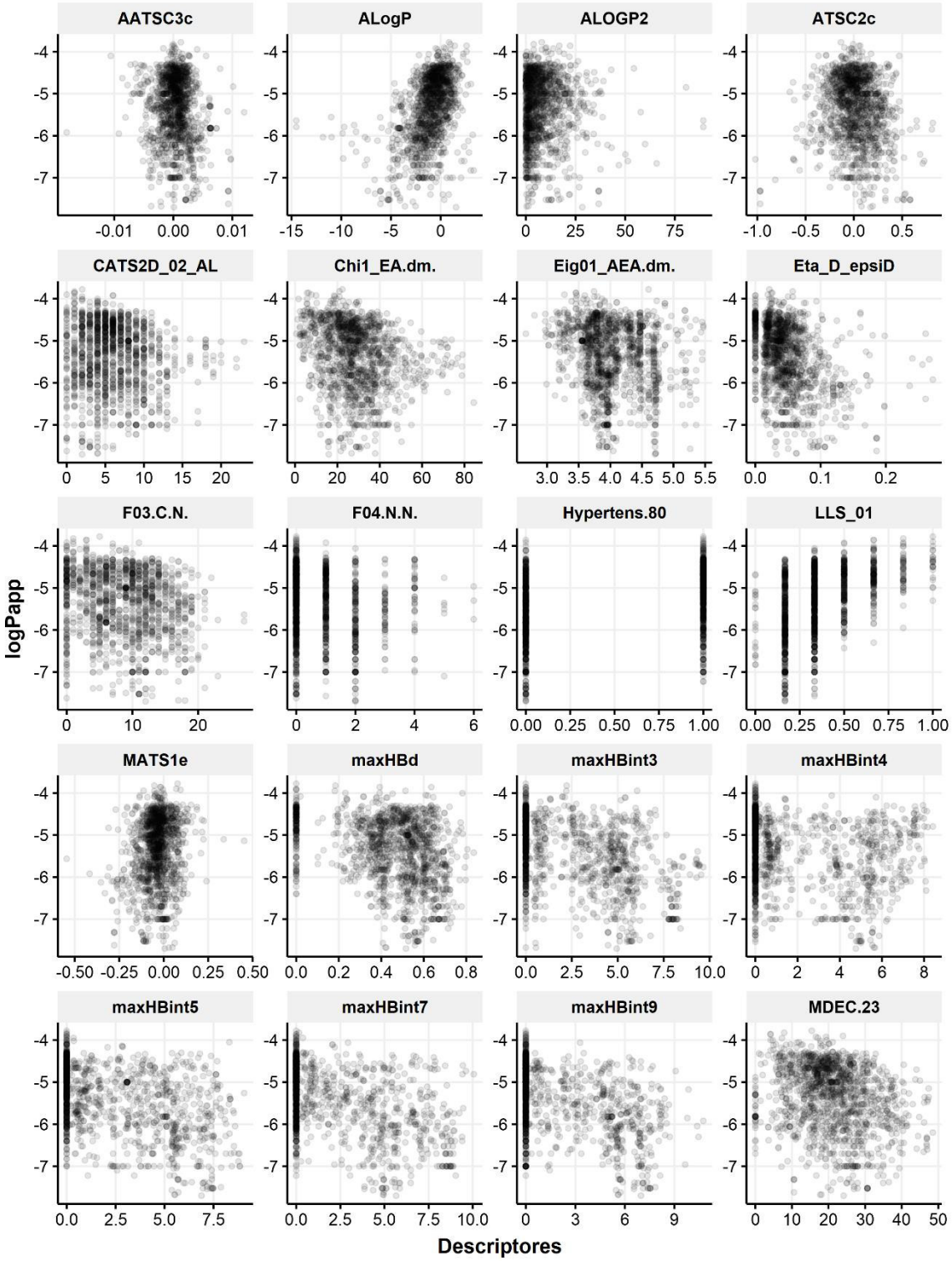
Anexo 5B. Valores del gráfico de RFE

Variables	RMSE	R^2	RMSESD	R^2SD	Variables	RMSE	R^2	RMSESD	R^2SD
101	0.41130	0.73179	0.01617	0.02505	**151**	0.41358	0.73002	0.01603	0.02493
102	0.41149	0.73142	0.01652	0.02531	**152**	0.41399	0.72942	0.01623	0.02528
103	0.41162	0.73146	0.01641	0.02520	**153**	0.41432	0.72896	0.01618	0.02526
104	0.41114	0.73209	0.01632	0.02503	**154**	0.41403	0.72940	0.01607	0.02490
105	0.41165	0.73133	0.01661	0.02550	**155**	0.41395	0.72951	0.01627	0.02538
106	0.41188	0.73103	0.01613	0.02501	**156**	0.41405	0.72942	0.01609	0.02497
107	0.41187	0.73116	0.01636	0.02530	**157**	0.41380	0.72980	0.01595	0.02483
108	0.41177	0.73118	0.01644	0.02548	**158**	0.41392	0.72971	0.01608	0.02499
109	0.41175	0.73137	0.01648	0.02556	**159**	0.41384	0.72985	0.01639	0.02523
110	0.41168	0.73155	0.01620	0.02473	**160**	0.41391	0.72976	0.01609	0.02494
111	0.41185	0.73130	0.01630	0.02487	**161**	0.41421	0.72936	0.01597	0.02486
112	0.41211	0.73097	0.01615	0.02483	**162**	0.41433	0.72921	0.01636	0.02540
113	0.41201	0.73126	0.01595	0.02478	**163**	0.41412	0.72954	0.01616	0.02517
114	0.41221	0.73092	0.01618	0.02501	**164**	0.41406	0.72968	0.01575	0.02471
115	0.41256	0.73036	0.01611	0.02523	**165**	0.41403	0.72970	0.01568	0.02441
116	0.41210	0.73114	0.01673	0.02544	**166**	0.41415	0.72971	0.01591	0.02483
117	0.41211	0.73108	0.01618	0.02507	**167**	0.41453	0.72892	0.01610	0.02535
118	0.41229	0.73096	0.01613	0.02497	**168**	0.41456	0.72892	0.01620	0.02545
119	0.41239	0.73090	0.01620	0.02467	**169**	0.41445	0.72923	0.01602	0.02486
120	0.41246	0.73069	0.01628	0.02494	**170**	0.41441	0.72927	0.01627	0.02517
121	0.41264	0.73047	0.01615	0.02492	**171**	0.41439	0.72929	0.01570	0.02463
122	0.41233	0.73109	0.01640	0.02524	**172**	0.41409	0.72982	0.01614	0.02497
123	0.41276	0.73038	0.01633	0.02532	**173**	0.41440	0.72947	0.01590	0.02468
124	0.41249	0.73086	0.01622	0.02512	**174**	0.41455	0.72916	0.01601	0.02476
125	0.41244	0.73099	0.01636	0.02524	**175**	0.41465	0.72903	0.01628	0.02528
126	0.41270	0.73050	0.01639	0.02527	**176**	0.41472	0.72908	0.01591	0.02483
127	0.41277	0.73051	0.01626	0.02523	**177**	0.41488	0.72867	0.01628	0.02506
128	0.41277	0.73057	0.01637	0.02533	**178**	0.41467	0.72914	0.01641	0.02523
129	0.41271	0.73069	0.01618	0.02494	**179**	0.41491	0.72881	0.01601	0.02508
130	0.41281	0.73062	0.01607	0.02515	**180**	0.41461	0.72919	0.01599	0.02509
131	0.41290	0.73060	0.01605	0.02469	**181**	0.41491	0.72881	0.01597	0.02476
132	0.41285	0.73055	0.01633	0.02493	**182**	0.41451	0.72946	0.01623	0.02494
133	0.41307	0.73026	0.01621	0.02524	**183**	0.41489	0.72890	0.01598	0.02492
134	0.41296	0.73041	0.01664	0.02564	**184**	0.41469	0.72922	0.01635	0.02534
135	0.41311	0.73024	0.01622	0.02497	**185**	0.41475	0.72921	0.01621	0.02511
136	0.41319	0.73023	0.01634	0.02503	**186**	0.41468	0.72936	0.01619	0.02537
137	0.41271	0.73091	0.01639	0.02534	**187**	0.41476	0.72923	0.01624	0.02520
138	0.41294	0.73062	0.01641	0.02521	**188**	0.41514	0.72855	0.01601	0.02483
139	0.41352	0.72977	0.01622	0.02526	**189**	0.41500	0.72890	0.01574	0.02443
140	0.41335	0.73020	0.01620	0.02454	**190**	0.41510	0.72882	0.01583	0.02454
141	0.41316	0.73029	0.01641	0.02514	**191**	0.41464	0.72948	0.01595	0.02492
142	0.41327	0.73022	0.01623	0.02520	**192**	0.41525	0.72856	0.01580	0.02474
143	0.41357	0.72986	0.01593	0.02474	**193**	0.41511	0.72881	0.01601	0.02499
144	0.41358	0.72973	0.01655	0.02554	**194**	0.41505	0.72886	0.01587	0.02471
145	0.41358	0.73003	0.01610	0.02512	**195**	0.41524	0.72862	0.01640	0.02548
146	0.41369	0.72974	0.01604	0.02464	**196**	0.41502	0.72902	0.01624	0.02500
147	0.41349	0.73012	0.01612	0.02519	**197**	0.41508	0.72891	0.01613	0.02512
148	0.41366	0.72984	0.01613	0.02499	**198**	0.41537	0.72847	0.01589	0.02501
149	0.41375	0.72975	0.01614	0.02506	**199**	0.41543	0.72847	0.01604	0.02520
150	0.41369	0.72973	0.01628	0.02538	**200**	0.41537	0.72858	0.01626	0.02519
					521	0.41953	0.72533	0.01613	0.02562

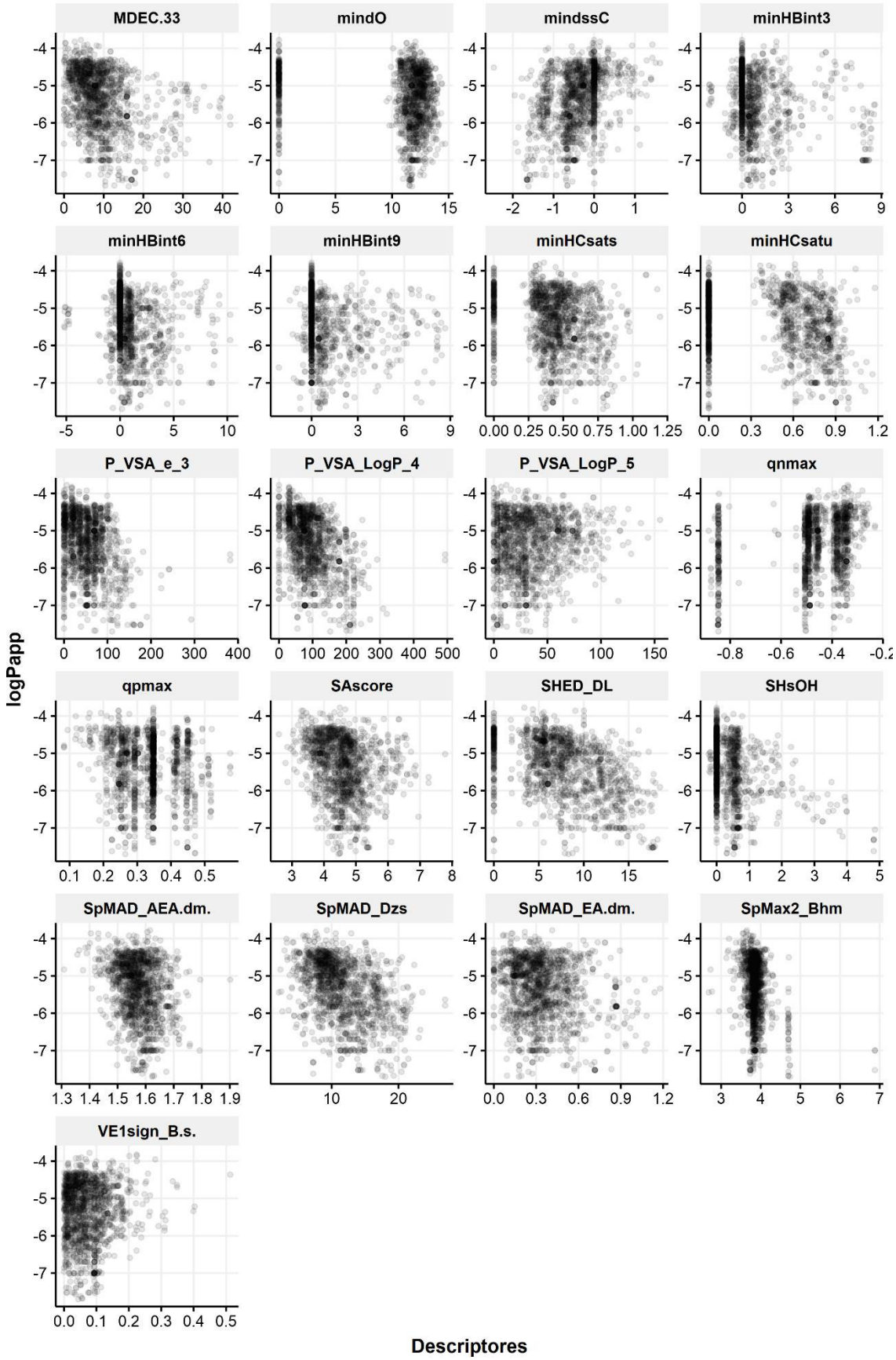
Anexo 6. Gráfico de importancia relativa de la variable según el modelo SVM-RF-GBM



Anexo 7A. Gráfico de dispersión entre logPapp y los descriptores seleccionados



Anexo 7B. Gráfico de dispersión de correlación entre logPapp y los descriptores seleccionados



Anexo 8A. Tabla de clasificación metabólica de los 516 productos naturales de la biodiversidad del Perú

Vía metabólica	Super clase	Clase	N
Alcaloides	Alcaloides de lisina y pseudoalcaloides	Capsaicinas y Capsaicinoides	2
Alcaloides	Alcaloides de lisina y pseudoalcaloides	Alcaloides de piperidina	2
Alcaloides	Alcaloides de ornitina	Alcaloides de pirrolizidina	11
Alcaloides	Alcaloides de ornitina	Alcaloides de tropano	2
Alcaloides	Alcaloides de triptófano	Tipo de aspidosperma	16
Alcaloides	Alcaloides de triptófano	Alcaloides de carbolina	3
Alcaloides	Alcaloides de triptófano	Alcaloides de cinchona	1
Alcaloides	Alcaloides de triptófano	Tipo corynanthe	10
Alcaloides	Alcaloides de triptófano	Alcaloides de pirroloquinolina	1
Alcaloides	Alcaloides de triptófano	Alcaloides de indol simples	14
Alcaloides	Alcaloides de triptófano	Alcaloides de oxindol simple	7
Alcaloides	Alcaloides de triptófano	Tipo Estricnos	1
Alcaloides	Alcaloides de triptófano	Alcaloides similares a la yohimbina	3
Alcaloides	Alcaloides de tirosina	Alcaloides de betalaína	1
Alcaloides	Alcaloides de tirosina	Alcaloides de isoquinolina	1
Alcaloides	Alcaloides de tirosina	Feniletilaminas	2
Alcaloides	Alcaloides de tirosina	Alcaloides de tetrahidroisoquinolina	1
Aminoácidos y Péptidos	Aminoácidos y Péptidos	Tipo de ariakemicina	2
Aminoácidos y Péptidos	Aminoácidos y Péptidos	Glucosinolatos	1
Aminoácidos y Péptidos	Aminoácidos y Péptidos	Lipopéptidos	3
Aminoácidos y Péptidos	Aminoácidos y Péptidos	N-acilaminas	5
Ácidos grasos	Ácidos Grasos y Conjugados	Ácidos grasos insaturados	23
Ácidos grasos	Acilos grasos	Alcoholes grasos	1
Ácidos grasos	Acilos grasos	Aldehídos grasos	1
Ácidos grasos	Acilos grasos	Hidrocarburos	3
Ácidos grasos	Acilos grasos	Hidrocarburos oxigenados	1
Ácidos grasos	Amidas grasas	Capsaicinas y Capsaicinoides	2
Ácidos grasos	Amidas grasas	N-acilaminas	39
Ácidos grasos	Amidas grasas	N-acil etanolaminas (endocannabinoides)	1
Ácidos grasos	Amidas grasas	Amidas primarias	1
Ácidos grasos	Ésteres grasos	Monoésteres de cera	18
Policétidos	Policétidos	Acil floroglucinoles	1
Policétidos	Policétidos	Bisnaftalenos	1
Policétidos	Policétidos	Derivados de ftaluro	2

Anexo 8*B. Tabla de clasificación metabólica de los 516 productos naturales de la biodiversidad del Perú*

Vía metabólica	Super clase	Clase	N
Shikimatos y fenilpropanoides	Cumarinas y Diarilheptanoides	Furocumarinas	1
Shikimatos y fenilpropanoides	Cumarinas y Diarilheptanoides	Diarilheptanoides lineales	1
Shikimatos y fenilpropanoides	Cumarinas y Diarilheptanoides	Cumarinas simples	1
Shikimatos y fenilpropanoides	Flavonoides	Antocianidinas	7
Shikimatos y fenilpropanoides	Flavonoides	Chalcones	3
Shikimatos y fenilpropanoides	Flavonoides	Dihidroflavonoles	3
Shikimatos y fenilpropanoides	Flavonoides	Flavan-3-oles	5
Shikimatos y fenilpropanoides	Flavonoides	Flavanonas	3
Shikimatos y fenilpropanoides	Flavonoides	Flavans	1
Shikimatos y fenilpropanoides	Flavonoides	Flavonas	6
Shikimatos y fenilpropanoides	Flavonoides	Flavonoles	18
Shikimatos y fenilpropanoides	Flavonoides	Proantocianinas	3
Shikimatos y fenilpropanoides	Isoflavonoides	Isoflavanonas	2
Shikimatos y fenilpropanoides	Isoflavonoides	Isoflavonas	5
Shikimatos y fenilpropanoides	Isoflavonoides	Rotenoides	8
Shikimatos y fenilpropanoides	Lignanos	Lignanos de arilnaftaleno y ariltetralina	3
Shikimatos y fenilpropanoides	Lignanos	Lignanos de dibencilbutano	8
Shikimatos y fenilpropanoides	Lignanos	Lignanos de dibencilbutirolactona	6
Shikimatos y fenilpropanoides	Lignanos	Lignanos furanoideos	4
Shikimatos y fenilpropanoides	Lignanos	Lignanos furofuranoide	2
Shikimatos y fenilpropanoides	Lignanos	Lignanos menores	3
Shikimatos y fenilpropanoides	Lignanos	Neolignanos	4
Shikimatos y fenilpropanoides	Ácidos fenólicos (C6-C1)	Galotaninos	3
Shikimatos y fenilpropanoides	Ácidos fenólicos (C6-C1)	Ácidos fenólicos simples	12
Shikimatos y fenilpropanoides	Fenilpropanoides (C6-C3)	Ácidos cinámicos y derivados	20
Shikimatos y fenilpropanoides	Estilbenoides	Estilbenos monoméricos	8
Shikimatos y fenilpropanoides	Xantonas	Xantonas vegetales	2
Terpenoides	Esteroides	Bufadienolidos	21
Terpenoides	Esteroides	Esteroides ergostano	26
Terpenoides	Esteroides	Esteroides espirostano	2
Terpenoides	Esteroides	Esteroides de estigmastano	3

Anexo 8*C. Tabla de clasificación metabólica de los 516 productos naturales de la biodiversidad del Perú*

Vía metabólica	Super clase	Clase	N
Terpenoides	Apocarotenoides	Apocarotenoides	3
Terpenoides	Diterpenoides	Diterpenoides colensano y clerodano	5
Terpenoides	Diterpenoides	Diterpenoides Halimane	3
Terpenoides	Diterpenoides	Diterpenoides Kaurane y Phyllocladane	3
Terpenoides	Diterpenoides	Diterpenoides labdanos	26
Terpenoides	Diterpenoides	Diterpenoides norkaurano	1
Terpenoides	Diterpenoides	Diterpenoides de norlabdano	6
Terpenoides	Diterpenoides	Diterpenoides de fitano	1
Terpenoides	Diterpenoides	Diterpenoides Pimarane e Isopimarane	1
Terpenoides	Meroterpenoides y Triterpenoides	Meroterpenoides de prenilquinona	4
Terpenoides	Meroterpenoides y Triterpenoides	Triterpenoides ursano y taraxastano	2
Terpenoides	Monoterpenoides	Monoterpenoides acíclicos	3
Terpenoides	Monoterpenoides	Monoterpenoides iridoides	3
Terpenoides	Monoterpenoides	Monoterpenoides monocíclicos	7
Terpenoides	Monoterpenoides	Monoterpenoides de pinano	3
Terpenoides	Sesquiterpenoides	Sesquiterpenoides de aristolane	1
Terpenoides	Sesquiterpenoides	Sesquiterpenoides de aromadendrano	6
Terpenoides	Sesquiterpenoides	Bisabolano sesquiterpenoides	5
Terpenoides	Sesquiterpenoides	Sesquiterpenoides de bourbonano	1
Terpenoides	Sesquiterpenoides	Sesquiterpenoides de cadinano	8
Terpenoides	Sesquiterpenoides	Sesquiterpenoides de cariofilano	5
Terpenoides	Sesquiterpenoides	Sesquiterpenoides de Copaane	1
Terpenoides	Sesquiterpenoides	Sesquiterpenoides cicloeudesmanos	7
Terpenoides	Sesquiterpenoides	Sesquiterpenoides ciclofarnesano	1
Terpenoides	Sesquiterpenoides	Sesquiterpenoides Daucane	1
Terpenoides	Sesquiterpenoides	Tipo ectocarpeno	1
Terpenoides	Sesquiterpenoides	Elemane sesquiterpenoides	1
Terpenoides	Sesquiterpenoides	Sesquiterpenoides de eudesmane	2
Terpenoides	Sesquiterpenoides	Sesquiterpenoides de germacrane	20
Terpenoides	Sesquiterpenoides	Sesquiterpenoides de guayano	3
Terpenoides	Sesquiterpenoides	Himachalane sesquiterpenoides	1
Terpenoides	Sesquiterpenoides	Sesquiterpenoides de humulano	1
Terpenoides	Sesquiterpenoides	Longifolano sesquiterpenoides	1
Terpenoides	Sesquiterpenoides	Sesquiterpenoides pseudoguaianos	2

Anexo 9A. Tabla de clasificación taxonómica del organismo de origen de los 516 productos naturales de la biodiversidad del Perú

Reino	División/Filo	Orden	Familia	Género	Especie	N
Animal	Amphibia	Anura	Bufonidae	*Rhinella*	*Rhinella marina*	29
Animal	Amphibia	Anura	Dendrobatidae	*Epipedobates*	*Epipedobates tricolor*	2
Monera	Actinomycetes	Streptomycetales	Streptomycetaceae	*Streptomyces*	*Streptomyces sp.*	2
Monera	Bacilli	Bacillales	Paenibacillaceae	*Paenibacillus*	*Paenibacillus sp.*	3
Monera	Cytophagia	Cytophagales	Flammeovirgaceae	*Rapidithrix*	*Rapidithrix thailandica*	2
Vegetal	Angiosperms	Asterales	Asteraceae	*Ambrosia*	*Ambrosia peruviana*	2
Vegetal	Angiosperms	Asterales	Asteraceae	*Aristeguietia*	*Aristeguietia gayana* (Wedd.) R.M. King & H. Rob	11
Vegetal	Angiosperms	Asterales	Asteraceae	*Culcitium*	*Culcitium canescens* H. & B.	3
Vegetal	Angiosperms	Asterales	Asteraceae	*Elephantopus*	*Elephantopus mollis* H.B.K.	16
Vegetal	Angiosperms	Asterales	Asteraceae	*Mikania*	*Mikania decora*	4
Vegetal	Angiosperms	Asterales	Asteraceae	*Pseudelephantopus*	*Pseudelephantopus spiralis* (Less.) Cronquist	7
Vegetal	Angiosperms	Asterales	Asteraceae	*Tanacetum*	*Tanacetum parthenium* L.	7
Vegetal	Angiosperms	Brassicales	Brassicaceae	*Lepidium*	*Lepidium meyenii*	45
Vegetal	Angiosperms	Caryophyllales	Amaranthaceae	*Chenopodium*	*Chenopodium murale*	1
Vegetal	Angiosperms	Caryophyllales	Amaranthaceae	*Chenopodium*	*Chenopodium quinoa*	7
Vegetal	Angiosperms	Caryophyllales	Amaranthaceae	*Dysphania*	*Dysphania ambrosioides*	1
Vegetal	Angiosperms	Caryophyllales	Cactaceae	*Echinopsis*	*Echinopsis pachanoi*	1
Vegetal	Angiosperms	Caryophyllales	Cactaceae	*Echinopsis*	*Echinopsis peruviana*	1
Vegetal	Angiosperms	Caryophyllales	Polygonaceae	*Polygonum*	*Polygonum cuspidatum.*	3
Vegetal	Angiosperms	Chloranthales	Chloranthaceae	*Hedyosmum*	*Hedyosmum angustifolium*	9
Vegetal	Angiosperms	Ericales	Primulaceae	*Stylogyne*	*Stylogyne cauliflora*	5

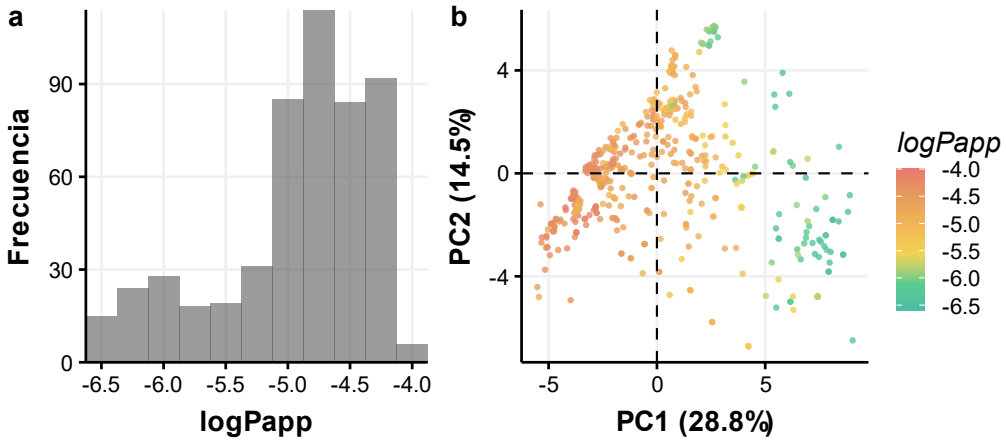
Anexo 9B. Tabla de clasificación taxonómica del organismo de origen de los 516 productos naturales de la biodiversidad del Perú

Reino	División/Filo	Orden	Familia	Género	Especie	N
Vegetal	Angiosperms	Fabales	Fabaceae	*Copaifera*	*Copaifera paupera* (Herzog) Dwyer	17
Vegetal	Angiosperms	Fabales	Fabaceae	*Glycine*	*Glycine max* L.	4
Vegetal	Angiosperms	Fabales	Fabaceae	*Hymenaea*	*Hymenaea courbaril* L.	45
Vegetal	Angiosperms	Fabales	Fabaceae	*Hymenaea*	*Hymenaea courbaril* L., H. *stigonocarpa*	4
Vegetal	Angiosperms	Fabales	Fabaceae	*Hymenaea*	*Hymenaea oblongifolia* H.	5
Vegetal	Angiosperms	Fabales	Fabaceae	*Hymenaea*	*Hymenaea parvifolia* H.	2
Vegetal	Angiosperms	Fabales	Fabaceae	*Hymenaea*	*Hymenaea stigonocarpa*	4
Vegetal	Angiosperms	Fabales	Fabaceae	*Lonchocarpus*	*Lonchocarpus nicou*	8
Vegetal	Angiosperms	Fabales	Fabaceae	*Pachyrhizus*	*Pachyrhizus tuberosus*	7
Vegetal	Angiosperms	Fabales	Fabaceae	*Swartzia*	*Swartzia polyphylla* DC	3
Vegetal	Angiosperms	Gentianales	Apocynaceae	*Aspidosperma*	*Aspidosperma desmanthum*	15
Vegetal	Angiosperms	Gentianales	Apocynaceae	*Aspidosperma*	*Aspidosperma spruceanum*	8
Vegetal	Angiosperms	Gentianales	Apocynaceae	*Himatanthus*	*Himatanthus sucuuba*	2
Vegetal	Angiosperms	Gentianales	Loganiaceae	*Strychnos*	*Strychnos toxifera*	1
Vegetal	Angiosperms	Gentianales	Rubiaceae	*Cinchona*	*Cinchona sp.*	1
Vegetal	Angiosperms	Gentianales	Rubiaceae	*Psychotria*	*Psychotria viridis*	1
Vegetal	Angiosperms	Gentianales	Rubiaceae	*Uncaria*	*Uncaria sp.*	29
Vegetal	Angiosperms	Gentianales	Rubiaceae	*Uncaria*	*Uncaria tomentosa*	5
Vegetal	Angiosperms	Lamiales	Verbenaceae	*Lippia*	*Lippia alva*	3
Vegetal	Angiosperms	Laurales	Lauraceae	*Aniba*	*Aniba rosaeodora*	1
Vegetal	Angiosperms	Magnoliales	Annonaceae	*Annona*	*Annona montana*	1

Anexo 9C. Tabla de clasificación taxonómica del organismo de origen de los 516 productos naturales de la biodiversidad del Perú

Reino	División/Filo	Orden	Familia	Género	Especie	N
Vegetal	Angiosperms	Magnoliales	Annonaceae	*Cremastosperma*	*Cremastosperma microcarpum*	1
Vegetal	Angiosperms	Magnoliales	Myristicaceae	*Iryanthera*	*Iryanthera lancifolia*	17
Vegetal	Angiosperms	Malpighiales	Erythroxylaceae	*Erythroxylum*	*Erythroxylum coca*	1
Vegetal	Angiosperms	Malpighiales	Euphorbiaceae	*Croton*	*Croton lechleri* Müll. Arg.	2
Vegetal	Angiosperms	Malpighiales	Euphorbiaceae	*Plukenetia*	*Plukenetia volubilis* L	14
Vegetal	Angiosperms	Malpighiales	Hypericaceae	*Hypericum*	*Hypericum laricifolium* Juss.	16
Vegetal	Angiosperms	Malvales	Bixaceae	*Bixa*	*Bixa orellana*	1
Vegetal	Angiosperms	Piperales	Piperaceae	*Peperomia*	*Peperomia pellucida* (L.) Kunth	37
Vegetal	Angiosperms	Poales	Poaceae	*Zea*	*Zea mays* L.	10
Vegetal	Angiosperms	Sapindales	Anacardiaceae	*Schinus*	*Schinus molle*	4
Vegetal	Angiosperms	Solanales	Solanaceae	*Capsicum*	*Capsicum annuum*	7
Vegetal	Angiosperms	Solanales	Solanaceae	*Capsicum*	*Capsicum baccatum*	15
Vegetal	Angiosperms	Solanales	Solanaceae	*Capsicum*	*Capsicum chinense*	10
Vegetal	Angiosperms	Solanales	Solanaceae	*Capsicum*	*Capsicum species*	14
Vegetal	Angiosperms	Solanales	Solanaceae	*Cestrum*	*Cestrum auriculatum*	3
Vegetal	Angiosperms	Solanales	Solanaceae	*Cestrum*	*Cestrum hediundinum*	8
Vegetal	Angiosperms	Solanales	Solanaceae	*Physalis*	*Physalis peruviana* L.	25
Vegetal	Bryophytes	Jungermanniales	Plagiochilaceae	*Plagiochila*	*Plagiochila disticha*	4

Anexo 10. Análisis de datos de los 516 productos naturales. a, Distribución de logPapp de la base de datos de productos naturales, b, PCA de la base de datos de productos naturales



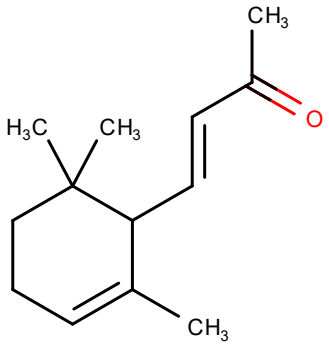
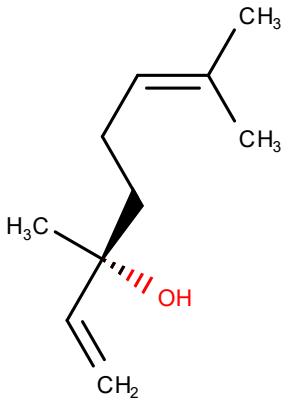
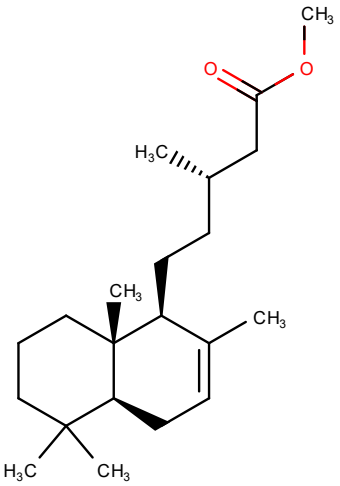
Anexo 11. Apalancamientos de los compuestos fuera del dominio de aplicabilidad en el conjunto de entrenamiento, prueba y base de datos de productos naturales

ID	h	Grupo	logPapp	ID	h	Grupo	logPapp
13	0.187	Train	-7.52	211	0.131	Test	-5.03
64	0.178	Train	-7.00	40	0.114	NP	-6.41
65	0.187	Train	-7.00	41	0.115	NP	-6.39
336	0.149	Train	-5.92	42	0.103	NP	-6.18
404	0.226	Train	-5.79	71	0.173	NP	-6.10
487	0.226	Train	-5.64	87	0.102	NP	-5.30
518	0.091	Train	-5.56	158	0.089	NP	-5.14
577	0.133	Train	-5.44	159	0.129	NP	-5.36
586	0.098	Train	-5.42	317	0.102	NP	-5.71
798	0.112	Train	-5.10	318	0.101	NP	-5.74
799	0.094	Train	-5.10	321	0.248	NP	-5.80
901	0.110	Train	-4.96	322	0.245	NP	-5.79
997	0.115	Train	-4.83	330	0.102	NP	-5.30
42	0.642	Test	-6.39	358	0.157	NP	-4.77
54	0.138	Test	-6.18	426	0.184	NP	-5.99
132	0.146	Test	-5.58				

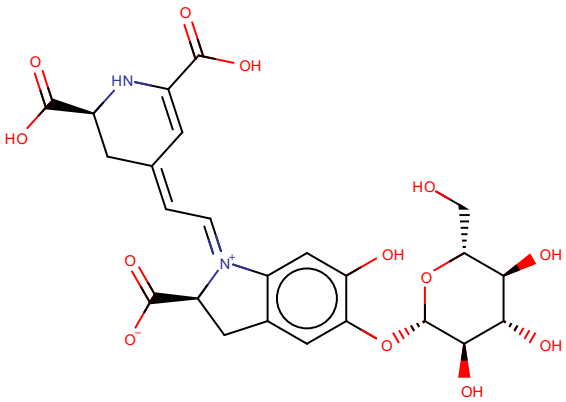
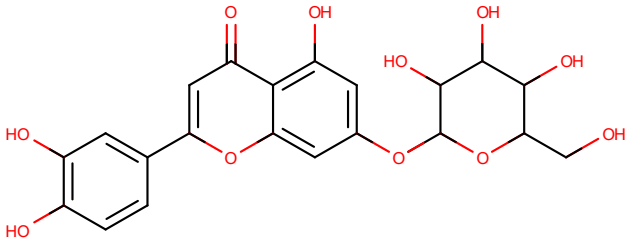
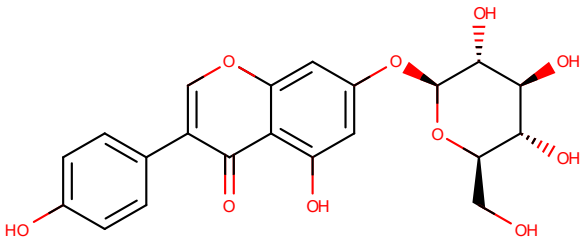
Anexo 12. Porcentaje de violaciones las reglas de Lipinski o Veber de la base de datos de productos naturales de la biodiversidad del Perú

Pappclass	Rule	Valor	Total	Porcentaje (%)	Grupo
L	0	6	51	11.77	LR5
L	1	15	51	29.41	LR5
L	2	12	51	23.53	LR5
L	3	18	51	35.29	LR5
M	0	74	113	65.49	LR5
M	1	26	113	23.01	LR5
M	2	8	113	7.08	LR5
M	3	5	113	4.43	LR5
H	0	222	338	65.68	LR5
H	1	114	338	33.73	LR5
H	2	2	338	0.59	LR5
L	0	4	51	7.84	VR2
L	1	37	51	72.55	VR2
L	2	10	51	19.61	VR2
M	0	71	113	62.83	VR2
M	1	24	113	21.24	VR2
M	2	18	113	15.93	VR2
H	0	274	338	81.07	VR2
H	1	64	338	18.94	VR2

Anexo 13. Productos naturales con alto LogPapp predicho

α-ionona	R-(-)-Linalool	metilcativato
		
MW = 192.15 MLogP = 3 HBA = 1 HBD = 9 RBN = 2 TPSA = 17.07 LogPapp = - 4.00 Clase = H	MW = 154.14 MLogP = 2.64 HBA = 1 HBD = 1 RBN = 4 TPSA = 20.23 LogPapp = - 4.18 Clase = H	MW = 320.27 MLogP = 4.96 HBA = 2 HBD = 0 RBN = 6 TPSA = 26.30 LogPapp = - 4.22 Clase = H
• Terpenoide, apocarotenoide • *Capsicum annuum* • Sabor, olor y pungencia • Glucósido: No	• Terpenoide, monoterpenoide • *Aniba rosaeodora* • Sedativo, anticonvulsivo, antidepresivo, antiinflamatorio. • Glucósido: No	• Terpenoide, dipeterpenoide • *Hymenaea stigonocarpa* • Antioxidante • Glucósido: No

Anexo 14. Productos naturales con bajo LogPapp predicho

Betacianina	
	MW = 549.14 MLogP = - 1.89 HBA = 15 HBD = 6 RBN = 7 TPSA = 255.04 LogPapp = - 6.60 Clase = L Alcaloides derivados de tirosina *Chenopodium quinoa* (“Quinoa”) Antioxidante Glucósido: Sí
Luteolina-7-O-glucósido	
	MW = 448.10 MLogP = - 1.28 HBA = 11 HBD = 7 RBN = 4 TPSA = 190.28 LogPapp = - 6.46 Clase = L Shikimatos y fenilpropanoides, flavonoides *Tanacetum parthenium* (“Santa Maria”) Antioxidante Glucósido: Sí
Genistina	
	MW = 432.11 MLogP = - 0.56 HBA = 10 HBD = 6 RBN = 4 TPSA = 170.05 LogPapp = - 6.40 Clase = L Shikimatos y fenilpropanoides, isoflavonoides *Glycine max* (“Soja”) Antioxidante Glucósido: Sí