

Diversidad estructural y espacio químico de flavonoides reportados en recursos vegetales latinoamericanos: estudio químico-informático

Edwin Fernando Basantes Basantes*
<https://orcid.org/0000-0002-2262-0222>
ef.basantesb@uea.edu.ec
Universidad Estatal Amazónica
Pastaza, Ecuador

Marcelo Ramos-Flores
<https://orcid.org/0000-0001-9240-8748>
jmramos@esPOCH.edu.ec
Escuela Superior Politécnica de Chimborazo
Riobamba, Ecuador

Orley Alejandro Acosta Ordoñez
<https://orcid.org/0009-0003-7687-9672>
orley.acosta@esPOCH.edu.ec
Escuela Superior Politécnica de Chimborazo
Riobamba, Ecuador

Víctor Hugo Rodríguez Masabanda
<https://orcid.org/0009-0002-3719-3961>
vhrodriguez1@espe.edu.ec
Universidad de las Fuerzas Armadas-ESPE
Sangolquí, Ecuador

*Autor de correspondencia: ef.basantesb@uea.edu.ec

Recibido: (23/05/2026), Aceptado: (09/07/2026)

Resumen. Los flavonoides constituyen un grupo de metabolitos vegetales de alto interés por su diversidad estructural y por la influencia que sus modificaciones químicas ejercen sobre sus propiedades moleculares. En este estudio se caracterizó la diversidad estructural y el espacio químico de flavonoides reportados en recursos vegetales latinoamericanos mediante la integración y depuración de datos abiertos provenientes de *COCONUT* y *PubChem*. La metodología se sustentó en la construcción de un corpus regional, su depuración estricta por consistencia química y procedencia biológica, la integración nominal con *PubChem* y el análisis descriptivo y multivariado de descriptores fisicoquímicos. Los resultados mostraron un conjunto estructuralmente heterogéneo, con predominio de flavonoides metoxilados y glucosilados, alta trazabilidad química entre bases y una organización discernible del espacio químico. En conjunto, el estudio aporta una ruta reproducible para explorar metabolitos naturales latinoamericanos mediante herramientas quimiinformáticas.

Palabras clave: flavonoides, espacio químico, quimiinformática, productos naturales.

Structural Diversity and Chemical Space of Flavonoids Reported in Latin American Plant Resources: A Cheminformatics Study

Abstract. Flavonoids constitute a group of plant metabolites of high interest due to their structural diversity and the influence that their chemical modifications exert on their molecular properties. This study characterized the structural diversity and chemical space of flavonoids reported in Latin American plant resources through the integration and curation of open data from *COCONUT* and *PubChem*. The methodology was based on the construction of a regional corpus, its strict curation according to chemical consistency and biological origin, nominal integration with *PubChem*, and descriptive and multivariate analysis of physicochemical descriptors. The results showed a structurally heterogeneous set, with a predominance of methoxylated and glycosylated flavonoids, high chemical traceability between databases, and a discernible organization of the chemical space. Overall, this study provides a reproducible pathway for exploring Latin American natural metabolites using cheminformatics tools.

Keywords: flavonoids, chemical space, cheminformatics, natural products.

I. INTRODUCCIÓN

Los productos naturales continúan siendo una fuente relevante de diversidad molecular para la investigación química, farmacéutica y biotecnológica. En paralelo, el crecimiento de las bases de datos abiertas y de las herramientas de quimioinformática ha ampliado la posibilidad de organizar, comparar y explorar grandes conjuntos de compuestos con criterios estructurales y fisicoquímicos. En este escenario, recursos como *Collection of Open Natural Products (COCONUT)* han contribuido a reunir estructuras, anotaciones y metadatos de productos naturales en entornos reutilizables, mientras que los estudios de espacio químico han mostrado que este tipo de compuestos ocupa regiones amplias y distintivas dentro del universo molecular, con interés directo para análisis comparativos y procesos de priorización [1], [2], [3], [4].

Dentro de ese universo, los flavonoides ocupan un lugar particularmente relevante por su diversidad estructural, su amplia distribución en plantas y la influencia que ejercen sus modificaciones químicas sobre el comportamiento molecular. La presencia de grupos hidroxilo, metoxi y azúcares, así como la variación entre flavonas, flavonoles, flavanonas, chalconas e isoflavonas, modifica propiedades como la polaridad, la lipofilicidad, la capacidad de formar puentes de hidrógeno y, en consecuencia, la posición que estos compuestos pueden ocupar en el espacio químico. Por ello, su estudio no se agota en la identificación nominal de clases, sino que requiere enfoques capaces de integrar clasificación estructural, descriptores fisicoquímicos y trazabilidad de datos en una misma matriz analítica [3], [5].

Sin embargo, la disponibilidad de información abierta no elimina un problema metodológico importante. Los compuestos de origen natural suelen encontrarse dispersos en diferentes repositorios, con niveles variables de anotación, sinónimos múltiples, metadatos heterogéneos y diferencias en la forma de representar sus propiedades. Esta situación dificulta la construcción de subconjuntos regionales comparables y limita la caracterización sistemática de grupos específicos cuando se trabaja con fuentes abiertas. En el caso de los recursos vegetales latinoamericanos, esta limitación adquiere especial interés, ya que la riqueza química potencial de la región contrasta con la necesidad de disponer de corpus depurados y analíticamente consistentes para su exploración quimioinformática [1], [2], [4].

En este contexto, el estudio tuvo como objetivo caracterizar la diversidad estructural y el espacio químico de flavonoides reportados en recursos vegetales latinoamericanos mediante la integración, depuración y análisis quimioinformático de datos abiertos obtenidos de *COCONUT* y *PubChem*. El trabajo buscó pasar de la simple recuperación de registros a la construcción de una base analítica depurada y trazable, capaz de describir patrones estructurales y fisicoquímicos dentro de un subconjunto regional de metabolitos naturales. En este marco, el carácter latinoamericano del corpus se asumió como un criterio operacional sustentado en la colección regional de *COCONUT* y no como una validación taxonómica exhaustiva de cada especie asociada.

El artículo se estructura de la siguiente manera. En primer lugar, se describe la metodología empleada para la construcción del corpus, su depuración y la integración de las fuentes abiertas utilizadas. A continuación, se presentan los resultados y su discusión integrada, con énfasis en la conformación del conjunto final, su composición estructural, sus propiedades fisicoquímicas y la representación del espacio químico. Finalmente, se exponen las conclusiones derivadas del aporte realizado y sus posibles proyecciones en estudios posteriores.

II. FUNDAMENTOS TEÓRICOS

El análisis actual de productos naturales se apoya cada vez más en bases de datos abiertas y herramientas de quimioinformática capaces de organizar grandes volúmenes de información estructural y fisicoquímica. En este contexto, *COCONUT* ha ampliado el acceso a compuestos naturales anotados y ha facilitado la construcción de corpus reutilizables para estudios comparativos, mientras que otras bases regionales han mostrado el valor de estructurar repertorios latinoamericanos para hacer visible esta diversidad y fortalecer su exploración química [1], [2], [4], [6], [7], [8]. De forma complementaria, *PubChem* aporta una infraestructura útil para la trazabilidad, normalización e integración de información química, aunque la ambigüedad de sinónimos sigue siendo un reto relevante en el cruce entre bases abiertas [9], [10].

Dentro de este marco, la noción de espacio químico permite representar conjuntos de compuestos a partir de sus propiedades estructurales y fisicoquímicas, en lugar de limitarse a clasificaciones nomi-

nales aisladas [3], [11]. Esta perspectiva resulta especialmente pertinente para los flavonoides por su amplia variabilidad estructural y por la influencia de modificaciones como hidroxilación, metoxilación y glicosilación sobre la polaridad, la lipofilicidad y la capacidad de interacción molecular [5], [12], [13], [14].

A nivel regional, América Latina representa un escenario de especial interés por su diversidad botánica y por el potencial químico asociado a sus metabolitos naturales. Sin embargo, la información correspondiente a compuestos de origen latinoamericano ha estado dispersa en iniciativas parciales, bases temáticas o repertorios nacionales. Desde esta perspectiva, la caracterización de flavonoides reportados en recursos vegetales latinoamericanos resulta pertinente tanto por su valor químico como por su aporte metodológico, al integrar diversidad estructural, trazabilidad de datos y análisis del espacio químico en un mismo enfoque de trabajo.

III. METODOLOGÍA

A. Diseño del estudio y procesamiento de la información

La investigación se desarrolló mediante un enfoque cuantitativo sustentado en el procesamiento de datos químicos abiertos. El trabajo se organizó en cuatro fases: construcción del corpus inicial en *COCONUT*, depuración estricta del conjunto recuperado, integración complementaria con *PubChem* y conformación de la base analítica final para el análisis descriptivo y multivariado.

Como fuente principal se utilizó el archivo masivo *coconut_complete-10-2024.csv.zip*, que contiene 695 133 registros, cada uno correspondiente a un compuesto natural individual descrito en *COCONUT* junto con su información estructural, clasificación química, propiedades fisicoquímicas y metadatos asociados. A partir de este archivo se generaron tablas sucesivas para la selección, depuración, cruce con *PubChem* e integración final del corpus, conservando archivos intermedios para verificar la consistencia y trazabilidad del procedimiento.

De manera complementaria, se empleó *PubChem* para recuperar la correspondencia nominal de los compuestos con sus respectivos identificadores *Compound Identifier (CID)* e incorporar variables adicionales cuando la resolución fue inequívoca. El procesamiento se realizó en *Python* 3.13.5 y hojas de cálculo, utilizando principalmente *pandas* 2.2.3, *scikit-learn* 1.8.0, *matplotlib* 3.10.8 y *scipy* 1.17.0. Los archivos intermedios, el *script* principal y los materiales de reproducibilidad se depositaron en *Zenodo*: <https://doi.org/10.5281/zenodo.21269644>.

B. Construcción del corpus inicial en COCONUT

En la primera fase se importó el archivo completo de *COCONUT* y se verificó su estructura general. Sobre los 695 133 registros iniciales se aplicó un primer filtro mediante la variable *collections*, conservando únicamente las entradas asociadas a "*Latin America dataset*", con lo que el conjunto se redujo a 12 029 registros.

Posteriormente, se aplicó un segundo filtro químico a partir de las variables *chemical_class*, *chemical_sub_class* y *np_classifier_class*. Se consideraron elegibles los registros de la colección "*Latin America dataset*" con clasificación flavonoidal explícita o subclase estructuralmente congruente en al menos una de esas variables, y con información mínima de identificación química y clasificación estructural. Cuando existieron discrepancias entre columnas, se priorizó el criterio más específico disponible; si la adscripción seguía siendo ambigua o no podía sostenerse con suficiente consistencia, el registro se excluyó del corpus flavonoidal. De este procedimiento se obtuvo un corpus inicial de 1 443 registros candidatos.

C. Depuración del corpus

Una vez conformado el corpus inicial, se realizó una depuración estricta orientada a conservar únicamente los registros químicamente consistentes y con información biológica utilizable. En un primer nivel, se revisó nuevamente la coherencia de la clasificación química y se excluyeron 93 registros cuya adscripción no resultó suficientemente congruente con el núcleo flavonoidal definido para la investigación, con lo que el conjunto pasó a 1 350 registros.

En un segundo nivel, se examinó la información de procedencia biológica registrada en los campos *organisms* y *primary_organism_candidate*. Para esta etapa se exigió la presencia de información utilizable

en al menos uno de esos campos y se conservaron únicamente las entradas con al menos una asociación biológica interpretable como recurso vegetal. En consecuencia, se eliminaron 452 registros por ausencia total de organismo asociado y 2 registros adicionales por presentar una procedencia ambigua, inespecífica o insuficiente para sostener razonablemente su relación con un recurso vegetal. En esta fase no se utilizaron porcentajes de coincidencia ni umbrales probabilísticos, ya que la decisión de conservación o exclusión dependió de la suficiencia documental de cada registro. Como resultado, el corpus depurado quedó conformado por 896 registros.

Durante esta etapa no se fusionaron registros ni se eliminaron variantes estructurales anotadas como entradas independientes en la base original. Cada registro aceptado se mantuvo con su identificador propio de *COCONUT* y sus variables originales, con el fin de preservar la trazabilidad de la información fuente. La delimitación regional del conjunto no se realizó mediante validación taxonómica externa individual de todas las especies, sino a partir de la pertenencia previa de los registros a la colección "*Latin America dataset*" de *COCONUT* y de la presencia de una asociación biológica utilizable en sus metadatos.

D. Preparación del cruce con PubChem

Con el fin de complementar la trazabilidad química del corpus depurado, se preparó un cruce nominal con *PubChem* a partir de los 896 registros retenidos. Para ello se extrajo la variable *name*, se normalizaron espacios redundantes, se controlaron celdas vacías y se eliminaron repeticiones exactas, obteniéndose una lista de 725 nombres únicos.

La lista se exportó en formato de texto plano, con un nombre por línea y sin encabezado, para asegurar su compatibilidad con el *Identifier Exchange Service* de *PubChem*. La conversión se realizó empleando *Synonyms* como tipo de entrada y *CIDs* como tipo de salida, con el operador *Same CID*. La salida devolvió 751 filas, debido a que algunos nombres produjeron más de una coincidencia. La correspondencia aceptable se definió exclusivamente como coincidencia nominal exacta sobre la lista normalizada, sin aplicar umbrales de similitud textual ni estrategias aproximadas de reasignación.

E. Integración final de COCONUT y PubChem

La salida obtenida desde *PubChem* se integró con la base depurada de *COCONUT* mediante un cruce por nombre normalizado. Antes de la integración, los nombres se homogeneizaron mediante ajustes básicos de formato, especialmente en espacios y variaciones menores de escritura, con el fin de reducir desajustes no químicos. El cruce se realizó por coincidencia nominal exacta sobre la lista ya normalizada, sin aplicar reglas automáticas adicionales de resolución para nombres ambiguos ni procedimientos heurísticos de reasignación.

Cuando un nombre devolvió un solo CID, la correspondencia se aceptó como válida y se incorporó directamente a la base final. En cambio, cuando un mismo nombre generó más de un CID, el registro se conservó con una marca de ambigüedad. Los nombres sin resolución en *PubChem* se mantuvieron en la base con trazabilidad exclusiva de *COCONUT*, y los registros sin nombre utilizable quedaron señalados como no elegibles para el cruce nominal. Como resultado, la base final conservó los 896 registros del corpus depurado: 664 con correspondencia única en *PubChem*, 15 ambiguos, 45 sin resolución pese a contar con nombre utilizable, 171 sin nombre suficiente para el cruce y 1 con desajuste de formato. Además del CID, se incorporaron desde *PubChem* los descriptores fisicoquímicos disponibles cuando la correspondencia fue inequívoca. Paralelamente, se conservaron las variables originales de *COCONUT* vinculadas con identificación química, clasificación estructural, propiedades fisicoquímicas y procedencia biológica, obteniéndose así una base final integrada de 35 variables.

F. Organización de la base analítica y análisis de datos

A partir de la base final integrada se realizó un análisis descriptivo del corpus mediante frecuencias absolutas y relativas de las categorías registradas en *chemical_sub_class* y *np_classifier_class*, así como medidas de tendencia central y dispersión para las variables continuas.

Para la caracterización del espacio químico se seleccionaron seis descriptores fisicoquímicos continuos disponibles de manera completa en los 896 registros: peso molecular, coeficiente de partición octanol/agua (AlogP), área de superficie polar topológica, número de enlaces rotatorios, aceptores de puentes de hidrógeno y donadores de puentes de hidrógeno. Estas variables, procedentes de *COCONUT*, se

emplearon como núcleo analítico del estudio debido a su disponibilidad integral en todo el corpus. Los descriptores incorporados desde *PubChem* se conservaron como capa de enriquecimiento y verificación, pero no se utilizaron como base principal del análisis multivariado.

La pertinencia de aplicar una técnica de reducción de dimensionalidad se evaluó mediante la matriz de correlaciones entre los seis descriptores, el índice KMO y la prueba de Bartlett. La matriz mostró correlaciones altas entre peso molecular, área de superficie polar topológica, aceptores y donadores de puentes de hidrógeno, así como relaciones inversas de AlogP con los descriptores de polaridad. En el corpus final, el KMO global fue de 0,681 y la prueba de Bartlett resultó significativa ($\chi^2 = 9281,03$; gl = 15; $p < 0,001$), lo que respaldó la aplicación de PCA. Este método se seleccionó porque permitió resumir la variabilidad conjunta de un conjunto reducido de descriptores continuos y obtener una representación bidimensional interpretable del espacio químico, sin requerir la complejidad adicional de enfoques no lineales como *t-SNE* o UMAP.

Antes de la reducción de dimensionalidad se verificó la ausencia de valores faltantes y los datos se estandarizaron mediante transformación z. Sobre la matriz estandarizada se aplicó PCA, trabajando con los dos primeros componentes principales para la visualización e interpretación del espacio químico. Finalmente, durante la organización de resultados se mantuvieron los estados de integración con *PubChem* para distinguir entre compuestos con correspondencia única, registros ambiguos y entradas sustentadas exclusivamente en *COCONUT*.

IV. RESULTADOS Y DISCUSIÓN

A. Conformación del corpus e integración de fuentes

El proceso de selección y depuración permitió reducir progresivamente el universo inicial hasta obtener un corpus analíticamente depurado y coherente con los criterios del estudio. A partir de los 695 133 registros contenidos en la base completa de *COCONUT*, el filtrado por la colección “*Latin America dataset*” redujo el conjunto a 12 029 entradas. Posteriormente, la aplicación del filtro químico sobre las variables de clasificación estructural permitió identificar 1 443 registros candidatos pertenecientes al núcleo flavonoidal. Finalmente, tras la depuración estricta por consistencia química y por disponibilidad de asociación biológica utilizable, el corpus quedó conformado por 896 registros. La Figura 1 resume este flujo de construcción y depuración del corpus.

En términos comparativos, la depuración redujo la riqueza de categorías de *chemical_sub_class* de 50 a 36 y la de *np_classifier_class* de 35 a 18. Paralelamente, los índices de Shannon descendieron de 2,268 a 2,194 y de 2,145 a 1,863, mientras que los de Simpson pasaron de 0,800 a 0,795 y de 0,811 a 0,776, respectivamente. En conjunto, estos resultados muestran que la depuración estrechó el corpus y redujo categorías marginales o ambiguas sin eliminar su heterogeneidad estructural principal.

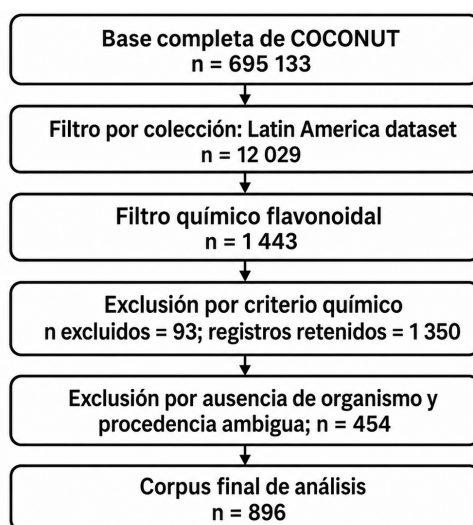


Fig. 1. Flujo de construcción y depuración del corpus.

La reducción observada no debe interpretarse como una pérdida de información, sino como un paso necesario para garantizar que el conjunto final respondiera efectivamente a los objetivos del estudio. En bases abiertas de productos naturales, el volumen inicial suele ser amplio, pero no todos los registros presentan el mismo nivel de completitud, consistencia química o utilidad analítica. En ese sentido, la necesidad de filtrar, depurar y reorganizar la información antes del análisis es coherente con lo reportado en estudios quimioinformáticos basados en bases de productos naturales de América Latina, donde la curación del contenido constituye una etapa decisiva para obtener corpus comparables y estructuralmente interpretables [7], [15].

Una vez consolidado el corpus depurado, se realizó su integración nominal con *PubChem*. La Tabla 1 resume el estado final de este cruce y muestra que, de los 896 registros retenidos, 664 presentaron una correspondencia única con *CID* en *PubChem*, lo que representó el 74,1 % del total. En contraste, 15 registros (1,7 %) devolvieron más de un *CID* y, por tanto, quedaron clasificados como ambiguos; 45 registros (5,0 %) no obtuvieron coincidencia en *PubChem*; 171 registros (19,1 %) no dispusieron de un nombre suficientemente utilizable para el cruce; y 1 registro (0,1 %) presentó un desajuste de formato. En términos generales, estos resultados indican que la integración con *PubChem* fue alta, aunque no completa.

Tabla 1. Estado de integración del corpus final con *PubChem*.

Estado de integración	n	%
Correspondencia única en <i>PubChem</i>	664	74,1
Correspondencia ambigua	15	1,7
Sin correspondencia en <i>PubChem</i>	45	5,0
Sin nombre utilizable para el cruce	171	19,1
Desajuste de formato	1	0,1
Total	896	100,0

Este comportamiento confirma que la integración con *PubChem* fue alta, aunque no total. La existencia de coincidencias únicas en casi tres cuartas partes del corpus respalda la trazabilidad química del conjunto; no obstante, la presencia de nombres ambiguos o no resueltos también pone de manifiesto un problema conocido en la integración entre bases químicas, particularmente cuando se trabaja con sinónimos, nombres no sistemáticos o variantes nominales heterogéneas. Esa dificultad ha sido reconocida tanto en estudios sobre ambigüedad de identificadores químicos no sistemáticos como en los propios procedimientos de filtrado de sinónimos y organización de nombres implementados en *PubChem* [9], [10]. En consecuencia, la articulación entre *COCONUT* y *PubChem* no funcionó como una sustitución entre fuentes, sino como una integración complementaria: la primera permitió delimitar y depurar el corpus regional, mientras que la segunda reforzó su trazabilidad química en los casos con resolución inequívoca.

B. Composición estructural del corpus

Una vez consolidada la base final, se examinó su composición estructural. Desde la perspectiva de *chemical_sub_class*, el conjunto estuvo dominado por *O-methylated flavonoids*, con 334 registros (37,3 %), seguido por *Flavonoid glycosides*, con 207 registros (23,1 %), y por *Flavones*, con 65 registros (7,3 %). A estas categorías les siguieron, ya con proporciones menores, *Cinnamylphenols* (4,5 %), *Flavans* (3,5 %) y *Pyranoflavonoids* (3,3 %). Esta distribución evidenció que el corpus no estuvo compuesto principalmente por flavonoides simples o poco modificados, sino por una fracción ampliamente enriquecida en derivados metoxilados y formas glicosiladas.

La clasificación según *np_classifier_class* reforzó esta tendencia. En este nivel, las clases más frecuentes fueron *Flavonols*, con 312 registros (34,8 %), *Flavones*, con 248 registros (27,7 %), *Chalcones*, con 102 registros (11,4 %), y *Flavanones*, con 88 registros (9,8 %). Aunque también aparecieron otras categorías, como *Isoflavones*, *Dihydroflavonols*, *Pterocarpan* y *Rotenoids*, su peso relativo fue claramente menor. En conjunto, estos resultados muestran que el corpus retuvo una base estructural ampliamente flavonólica y flavónica, pero acompañada por una diversidad apreciable de subgrupos menos abundantes. Esta distribución confirma que la diversidad del corpus no se explica solo por la clase base del flavonoide, sino también por el peso relativo de las modificaciones estructurales asociadas a metoxilación y glicosilación.

Este patrón resulta químicamente significativo, ya que el predominio de flavonoides *O*-metilados sugiere que la metilación constituye una vía importante de diversificación dentro del conjunto analizado. Además, la literatura ha señalado que esta modificación puede alterar la lipofilicidad, la estabilidad metabólica y la permeabilidad de los flavonoides [12], [13]. De manera complementaria, la elevada representación de glucósidos flavonoidales indica una contribución importante de compuestos con azúcares unidos a la estructura básica, lo que ayuda a explicar la heterogeneidad posterior del conjunto, pues la glicosilación suele incrementar la solubilidad y la estabilidad respecto de los aglicones [14], [16].

Por tanto, la composición observada no solo describe qué tipos de flavonoides estuvieron presentes en el corpus, sino que también anticipa una distribución amplia de propiedades fisicoquímicas. En otras palabras, la diversidad del conjunto no dependió únicamente del esqueleto flavonoidal de base, sino también del tipo de modificación estructural predominante, especialmente metoxilación y glicosilación. En este estudio, la diversidad estructural se examinó principalmente a partir de la clasificación química y de la distribución de subclases, por lo que su lectura se mantuvo en un plano comparativo-descriptivo más que en términos de índices cuantitativos de diversidad molecular.

C. Perfil fisicoquímico del corpus

Los principales descriptores fisicoquímicos empleados en la caracterización del corpus se resumen en la Tabla 2. En conjunto, los 896 registros mostraron un peso molecular medio de 394,07 Da, con una mediana de 365,40 Da y un valor máximo de 1171,03 Da, lo que evidenció una dispersión considerable. El comportamiento de la superficie polar topológica también fue amplio, con una media de 124,22 Å² y valores extremos que alcanzaron 467,17 Å². Del mismo modo, el número de aceptores y donadores de puentes de hidrógeno presentó una variabilidad marcada, con medias de 8,09 y 3,44, respectivamente. En contraste, la lipofilicidad estimada mediante el coeficiente de partición octanol/agua (AlogP) mostró un valor medio moderado de 2,37, aunque con un rango amplio entre -5,01 y 6,88.

Tabla 2. Estadísticos descriptivos de los descriptores fisicoquímicos del corpus final.

Descriptor	n	Media	Mediana	Desviación estándar	Mínimo	Máximo
Peso molecular (Da)	896	394,07	365,40	105,89	202,25	1171,03
AlogP	896	2,37	2,83	1,84	-5,01	6,88
Área de superficie polar topológica (Å ²)	896	124,22	105,82	62,44	0,00	467,17
Número de enlaces rotatorios	896	4,20	4,00	2,09	0,00	14,00
Aceptores de puentes de hidrógeno	896	8,09	7,00	3,44	1,00	29,00
Donadores de puentes de hidrógeno	896	3,44	3,00	2,90	0,00	16,00

Estos resultados confirman que el corpus no se distribuyó alrededor de un único perfil molecular, sino que reunió metabolitos con grados muy distintos de polaridad, funcionalización y tamaño. Esa amplitud fue coherente con la coexistencia de flavonoides metoxilados, glucósidos y subclases menos frecuentes, lo que desplazó la matriz tanto hacia compuestos relativamente compactos y menos polares como hacia estructuras de mayor peso molecular y mayor capacidad de interacción intermolecular.

Desde una perspectiva interpretativa, la dispersión del peso molecular y de la superficie polar resulta particularmente consistente con la importancia relativa de los glucósidos y derivados polioxigenados dentro del corpus. Al mismo tiempo, la amplitud observada en el coeficiente de partición octanol/agua (AlogP) concuerda con el predominio de compuestos metoxilados y con la coexistencia de estructuras menos polares dentro del conjunto. En otras palabras, la variabilidad fisicoquímica observada fue una consecuencia directa de la diversidad estructural retenida en el corpus, más que una propiedad secundaria del procedimiento analítico.

Este comportamiento también refuerza la utilidad del enfoque quimiinformático adoptado. En bases de productos naturales, la coexistencia de moléculas con propiedades contrastantes suele traducirse en coberturas amplias del espacio químico, lo que hace especialmente pertinentes las estrategias de reducción de dimensionalidad y representación multivariada. De hecho, trabajos previos sobre bases latinoamericanas de productos naturales y sobre quimiinformática aplicada a metabolitos naturales han destacado precisamente el valor de este tipo de análisis para organizar, comparar y visualizar corpus estructuralmente heterogéneos [6], [15].

D. Organización del espacio químico

La Figura 2 muestra la proyección del corpus en el espacio químico obtenido mediante análisis de componentes principales (PCA). Los dos primeros componentes explicaron conjuntamente el 92,55 % de la variabilidad total, lo que indicó una representación bidimensional altamente informativa del conjunto. El componente principal 1 (CP1) explicó el 74,44 % de la varianza y estuvo dominado principalmente por el peso molecular, la superficie polar topológica y la capacidad de formar puentes de hidrógeno. El componente principal 2 (CP2), con 18,11 % de la varianza, estuvo más influido por el número de enlaces rotatorios y, en menor medida, por la lipofilicidad.

El análisis de sensibilidad mostró que la estructura general del espacio químico se mantuvo prácticamente estable cuando el corpus se restringió a los registros con correspondencia única en *PubChem*, ya que la varianza acumulada de los dos primeros componentes pasó de 92,55 % a 92,48 %. En cambio, al excluir los glucósidos flavonoidales, la varianza acumulada descendió a 81,17 %, lo que indica que esta subclase contribuye de manera importante a la organización global del espacio químico. Por ello, la estructura observada no depende de forma crítica de la conservación de registros ambiguos de *PubChem*, pero sí está influida por el peso estructural de los flavonoides glicosilados dentro del corpus final.

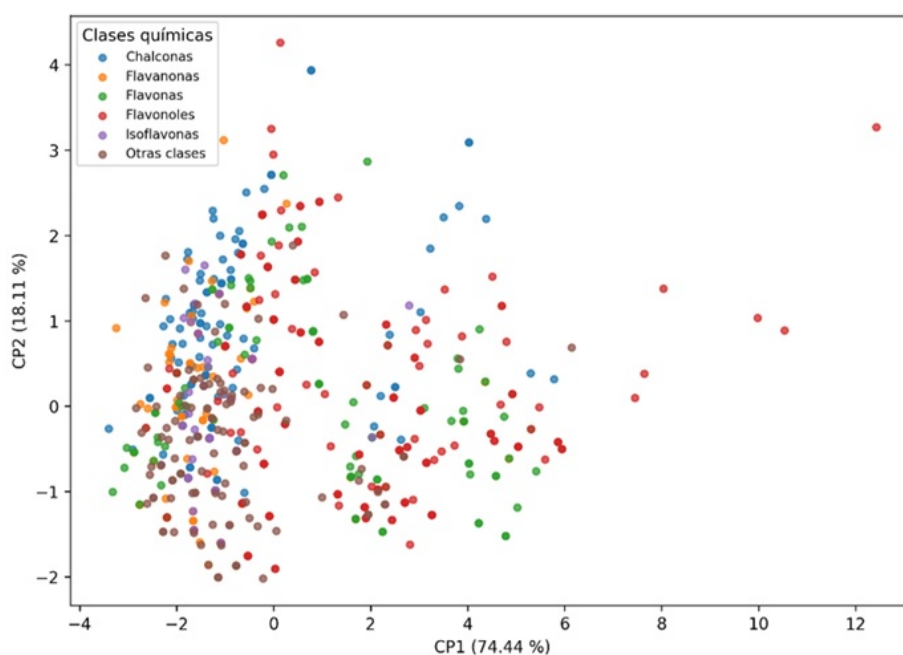


Fig. 2. Proyección del espacio químico mediante PCA.

La distribución de los puntos no mostró una separación tajante entre grupos mutuamente excluyentes, pero sí una organización gradual del corpus. En un extremo del espacio se concentraron preferentemente compuestos con mayor tamaño molecular y mayor polaridad, patrón compatible con estructuras glicosiladas o más intensamente funcionalizadas. En sentido opuesto, se ubicaron compuestos relativamente menos polares o más compactos. Esta disposición sugiere que el espacio químico del conjunto no fue aleatorio, sino que estuvo ordenado por descriptores fisicoquímicos con una fuerte capacidad discriminante.

Cuando la proyección del análisis de componentes principales se interpretó junto con la composición estructural del corpus, se observó que los flavonoles tendieron a ocupar regiones amplias del espacio químico, mientras que chalconas, flavanonas e isoflavonas se distribuyeron en zonas más acotadas o parcialmente superpuestas. Ello indica que la diferenciación del corpus no dependió exclusivamente de la clase química nominal, sino del balance entre tamaño molecular, polaridad y flexibilidad estructural. En consecuencia, el espacio químico no reflejó solo etiquetas de clasificación, sino combinaciones reales de propiedades moleculares [16], [17].

La elevada varianza explicada por los dos primeros componentes respalda la robustez descriptiva de esta representación. En estudios de quimioinformática aplicada a productos naturales, el uso de PCA

ha sido recurrente precisamente porque permite sintetizar la cobertura del espacio químico y detectar patrones de agrupamiento o solapamiento entre familias moleculares. Por ello, el comportamiento observado en este estudio se alinea con una lógica ya consolidada en la exploración computacional de bases de metabolitos naturales y, al mismo tiempo, muestra que el corpus de flavonoides latinoamericano retenido conserva una organización química interpretable en un espacio reducido de dimensiones [7], [11].

No obstante, la exploración del espacio químico se restringió a una representación mediante PCA y a su interpretación general, sin incorporar análisis adicionales de densidad, cobertura, similitud o agrupamientos. En consecuencia, la lectura del espacio químico se mantuvo en un plano descriptivo y constituye un punto de partida para futuras ampliaciones analíticas.

Entre las principales limitaciones del estudio se encuentran la dependencia de bases abiertas para la construcción del corpus, la integración nominal incompleta con *PubChem*, la existencia de registros sin *CID*, la ausencia de validación taxonómica exhaustiva de todas las especies asociadas, el sesgo regional derivado de la documentación disponible en *COCONUT* y la ausencia de validación experimental de los compuestos incluidos. Asimismo, el efecto de decisiones metodológicas específicas, como los filtros de depuración o la exclusión de determinados subconjuntos estructurales, no se evaluó mediante análisis de sensibilidad formal más amplio que el aquí presentado. Estas condiciones no invalidan el enfoque aplicado, pero sí delimitan el alcance interpretativo de los resultados y orientan futuras ampliaciones del estudio.

CONCLUSIONES

El principal aporte del estudio consistió en mostrar que la integración de bases químicas abiertas permite construir un corpus regional de flavonoides suficientemente consistente para su análisis estructural y quimioinformático. Más allá de la simple recuperación de registros, el trabajo convirtió un universo amplio y heterogéneo en una base analítica organizada, trazable y utilizable, lo que confirma la viabilidad de emplear repositorios abiertos como soporte para investigaciones químicas reproducibles enfocadas en recursos vegetales latinoamericanos.

Asimismo, el estudio evidenció que el conjunto analizado no responde a una composición uniforme, sino a una diversidad estructural real, modelada en buena medida por modificaciones como la metoxilación y la glicosilación. En consecuencia, la caracterización del espacio químico permitió reconocer que el repertorio flavonoidal asociado a la colección latinoamericana de *COCONUT* presenta una organización interna discernible y, por tanto, puede examinarse con criterios comparativos más sólidos que los derivados de una clasificación nominal aislada.

Desde el punto de vista metodológico, la articulación entre *COCONUT* y *PubChem* mostró que ambas fuentes cumplen funciones complementarias dentro de un mismo flujo de trabajo. Mientras *COCONUT* resultó decisiva para delimitar y estructurar el corpus regional, *PubChem* fortaleció su trazabilidad química en una proporción importante de los registros. Esta convergencia confirma que el valor de las bases abiertas no radica únicamente en su cobertura individual, sino en la posibilidad de combinarlas de manera crítica para mejorar la calidad analítica de los conjuntos de datos.

En términos de proyección, la base final obtenida constituye un punto de partida útil para procesos posteriores de priorización quimioinformática, exploración de relaciones estructura-propiedad, identificación temprana de compuestos conocidos y selección de compuestos con potencial de interés farmacológico, nutracéutico o quimiotaxonómico. Futuros estudios podrían ampliar este enfoque mediante validación taxonómica más fina, resolución estructural de casos ambiguos e incorporación de análisis de similitud, distancia química, diversidad molecular o cobertura del espacio químico. De este modo, el aporte del estudio no se limita a la caracterización del corpus construido, sino que propone una ruta metodológica aplicable a otros grupos de metabolitos naturales de interés regional.

REFERENCIAS

- [1] M. Sorokina, P. Merseburger, K. Rajan, M. A. Yirik, and C. Steinbeck, "Coconut online: Collection of open natural products database," *Journal of Cheminformatics*, vol. 13, no. 1, p. 2, 2021.
- [2] V. Chandrasekhar *et al.*, "Coconut 2.0: a comprehensive overhaul and curation of the collection of open natural products database," *Nucleic Acids Research*, vol. 53, no. D1, pp. D634–D643, 2025.
- [3] Y. Chen, M. Garcia De Lomana, N. O. Friedrich, and J. Kirchmair, "Characterization of the chemical space of known and readily obtainable natural products," *Journal of Chemical Information and Modeling*, vol. 58, no. 8, pp. 1518–1532, 2018.
- [4] T. Zeng, J. Li, and R. Wu, "Natural product databases for drug discovery: Features and applications," *Pharmaceutical Science Advances*, vol. 2, p. 100050, 2024.
- [5] M. C. Dias, D. C. G. A. Pinto, and A. M. S. Silva, "Plant flavonoids: Chemical characteristics and biological activity," *Molecules*, vol. 26, no. 17, p. 5377, 2021.
- [6] J. L. Medina-Franco and F. I. Saldívar-González, "Cheminformatics to characterize pharmacologically active natural products," *Biomolecules*, vol. 10, no. 11, p. 1566, 2020.
- [7] A. Gómez-García *et al.*, "Navigating the chemical space and chemical multiverse of a unified latin american natural product database: Lanapdb," *Pharmaceuticals*, vol. 16, no. 10, p. 1388, 2023.
- [8] A. Gómez-García and J. L. Medina-Franco, "Progress and impact of latin american natural product databases," *Biomolecules*, vol. 12, no. 9, p. 1202, 2022.
- [9] S. Kim, B. Yu, Q. Li, and E. E. Bolton, "Pubchem synonym filtering process using crowdsourcing," *Journal of Cheminformatics*, vol. 16, no. 1, p. 69, 2024.
- [10] S. Kim *et al.*, "Pubchem 2019 update: improved access to chemical data," *Nucleic Acids Research*, vol. 47, no. Database issue, pp. D1102–D1109, 2019.
- [11] J. Rosén, J. Gottfries, S. Muresan, A. Backlund, and T. I. Oprea, "Novel chemical space exploration via natural products," *Journal of Medicinal Chemistry*, vol. 52, no. 7, pp. 1953–1962, 2009.
- [12] Y. Fang, W. Cao, M. Xia, S. Pan, and X. Xu, "Study of structure and permeability relationship of flavonoids in caco-2 cells," *Nutrients*, vol. 9, no. 12, p. 1301, 2017.
- [13] M. Sajid, C. N. Channakesavula, S. R. Stone, and P. Kaur, "Synthetic biology towards improved flavonoid pharmacokinetics," *Biomolecules*, vol. 11, no. 5, p. 754, 2021.
- [14] L. Xie *et al.*, "Comparison of flavonoid o-glycoside, c-glycoside and their aglycones on antioxidant capacity and metabolism during in vitro digestion and in vivo," *Foods*, vol. 11, no. 6, p. 882, 2022.
- [15] B. A. Pilon-Jiménez, F. I. Saldívar-González, B. I. Díaz-Eufracio, and J. L. Medina-Franco, "Biofacquim: A mexican compound database of natural products," *Biomolecules*, vol. 9, no. 1, p. 31, 2019.
- [16] L. Hu, Y. Luo, J. Yang, and C. Cheng, "Botanical flavonoids: Efficacy, absorption, metabolism and advanced pharmaceutical technology for improving bioavailability," *Molecules*, vol. 30, no. 5, p. 1184, 2025.
- [17] S. Kim, "Getting the most out of pubchem for virtual screening," *Expert Opinion on Drug Discovery*, vol. 11, no. 9, pp. 843–855, 2016.