Universidad Central "Marta Abreu" de Las Villas

Facultad de Matemática, Física y Computación

Nueva codificación tridimensional de la estructura química de moléculas orgánicas

Tesis para optar por el título de Máster en Ciencias de la Computación

Autor

Ing. César Raúl García Jacas

Tutores

DrC. Yovani Marrero Ponce

DrC. Liesner Acevedo Martínez

Santa Clara, 2013

Dedicatoria

A toda mi familia por su apoyo incondicional

A mi mamá, hermanas y abuela por estar siempre a mi lado

A todas las personas que luchan por sus sueños sin importar los obstáculos

Agradecimientos

Ante todo quisiera agradecer a la persona más especial de este mundo, a la que me dio la posibilidad de vivir, y que pesar de las dificultades siempre fue delicada y a la vez exigente como un padre: a mi madre, que a pesar de no estar aquí sé que su pensamiento y amor está conmigo presente. A mi abuela María por estar a mi lado siempre y ser mi guía durante toda mi vida. A las niñas más lindas que existen para mí, mis hermanas Pilar e Ileanita, que siempre han sido fuente de inspiración y ejemplo. A Julio y Aimara, mi familia habanera, por su ayuda y apoyo incondicional. A mi novia Lisset, por su ternura y amor especial; y a su familia por acogerme como uno más.

A todas las personas que han contribuido y ayudado a realizar este trabajo: a Reinaldo Luna y Alexis René por cubrir mis turnos de clases cuando viajaba para Villa Clara, a Trinchet y Vladir por ser los primeros en impulsarme en esta tarea, a Téllez, Tonysé y mi gran amigo Longendri y su esposa Mónica por ser excepcionales personas con las cuales siempre puedo contar, a mi familia del departamento de Bioinformática de la UCI por su sincera amistad, a Stephen por sus sugerencias y críticas oportunas en el desarrollo de este trabajo, y a todos los profesores de la Maestría por el conocimiento aportado durante los cursos recibidos.

No por último es el menos importante, a mi tutor, el DrC. Yovani Marrero, por adentrarme y conducirme con sabiduría en el mágico mundo de la química informática, por su guía, enseñanza, motivación, ayuda constante y visión científica exacta. En general a todas las personas que de una forma u otra pusieron su granito de arena para ayudarme, tanto en este trabajo como en mi formación personal y profesional. A todos ustedes les estoy eternamente agradecido y de corazón les digo:

muchas gracias

Resumen

El presente trabajo describe nuevos métodos libres de alineamiento, basados en las formas lineal, bilineal, cuadrática, trilineal y cuatrilineal, utilizando la k^{th} matriz espacial de similitud-disimilitud. También son discutidos esquemas de generalización para el cálculo de la distancia espacial inter-atómica, mediante el uso de varias métricas. Por otro lado, son introducidos formalismos de normalización para la matriz espacial de similitud-disimilitud basados en los esquemas simple estocástico, doble estocástico y de probabilidad mutua. Además, con el fin de generalizar el uso de la combinación lineal de índices atómicos para alcanzar definiciones globales, fueron aplicados una serie de operadores de agregación. Análisis de variabilidad basado en entropía de Shannon, revelan que los índices propuestos tienen mejor comportamiento que los índices calculados por varios software utilizados en estudios quimio-informáticos. Un análisis de componentes principales muestra que los índices 3D basados en las formas algebraicas, codifican información estructural ortogonal a la capturada por los índices del software DRAGON. Finalmente, con el fin de obtener una visión sobre la contribución de los nuevos índices, fue realizado un estudio QSAR/QSPR para la afinidad de acoplamiento a la corticosteroide-binding globulin usando la base de datos de esteroides de Cramer. De esta forma se obtuvieron parámetros estadísticos de comparables a superiores de los índices QuBiLs MIDAS respecto a establecidas metodologías 3D-QSAR reportadas en la literatura, usando tanto las 31 estructuras como conjunto de entrenamiento, como dividiendo la data en conjunto de entrenamiento (1-21) y en conjunto de prueba (22-31).

Índice general

Introducción

En el siglo XVIII, con los trabajos del científico alemán Georg Ferdinand Helm, surge la química matemática [1]. Esta disciplina está relacionada con la aplicación de métodos computacionales para solucionar problemas químicos, con un énfasis particular en la manipulación de la información de la estructura química [2]. Con este fin son utilizados diferentes parámetros numéricos, denominados índices o descriptores moleculares (DMs). Formalmente se puede definir un descriptor molecular como: "*el resultado final de un procedimiento lógico y matemático que transforma información química codificada dentro de una representación simbólica de una molécula en un número de utilidad o el resultado de algún experimento estandarizado*" [3]. El término *utilidad* empleado en la definición anterior tiene un doble sentido: este significa que el número puede brindar una visión más profunda en la interpretación de propiedades moleculares y/o puede formar parte de un modelo que posibilite la predicción de alguna propiedad de interés en nuevas moléculas [3].

Los descriptores moleculares pueden ser obtenidos a partir de diferentes teorías, tales como: química-cuántica, teoría de información, química orgánica, teoría de grafos, entre otras; y han sido ampliamente utilizados en diversos estudios o campos científicos, entre los que se pueden mencionar los estudios QSAR/QSPR[1], estudios de similitud/disimilitud, el diseño de fármacos y la toxicología. Atendiendo a la naturaleza de su definición, los descriptores moleculares pueden ser clasificados en 0D o de conteo (que tienen en consideración la fórmula química), 1D o basados en fragmentos químicos de las moléculas, 2D o basados en rasgos topológicos de la estructura molecular, y en 3D o geométricos.

En la actualidad, la cantidad de índices moleculares reportados superan los 3000 [3]. A pesar de ello, la búsqueda de nuevos descriptores sigue siendo de interés científico, en aras de mejorar la diversidad de estos y posiblemente recoger información estructural no codificada adecuadamente por los índices definidos hasta la fecha. Entre las limitaciones existentes de los índices moleculares actuales, se pueden mencionar

[1] Quantitative Structure-Activity Relationships/Quantitative Structure-Property Relationships

las siguientes: 1) no existe un descriptor (variable) capaz de capturar y por lo tanto codificar toda la información química extrínseca e intrínseca de la estructura molecular, 2) aún existen propiedades moleculares no adecuadamente descritas por los índices definidos, 3) varios descriptores necesitan redefinirse a formas más simples o generalizadas en aras de disminuir el costo computacional, sin comprometer su calidad y al mismo tiempo aumentar su versatilidad, 4) una gran cantidad de índices moleculares solamente tienen una definición global, es decir, no pueden ser utilizados para estudios donde sean considerados fragmentos o tipos de átomos de la estructura química.

La definición local, es una de las características deseables de un descriptor molecular [3], debido a que muchas de las propiedades o incluso actividades biológicas dependen más de los rasgos estructurales de determinadas zonas moleculares que de la molécula como un todo. Es importante resaltar además, que los índices moleculares de definición local reportados hasta el momento para obtener índices globales o por tipos de átomos, solamente utilizan un enfoque aditivo (la suma de las partes es igual al total) [4–7].

Una de las principales aplicaciones de los descriptores moleculares es en el descubimiento de nuevos fármacos. Esta actividad por lo general involucra métodos para estudios QSAR/QSPR, los cuales se enfocan en determinar correlaciones entre los índices moleculares y la actividad farmacológica de las estructuras químicas. Estos métodos pueden ser categorizados en QSAR bidimensionales (2D-QSAR) y tridimensionales (3D-QSAR). Los procedimientos 2D-QSAR comúnmente utilizan información química derivada de la constitución e información topológica de las moléculas (descriptores 0D-2D) [8]. Por su parte los métodos 3D-QSAR, consideran propiedades físico-químicas de los ligandos y su conformación bioactiva [9–11], empleando para ello los llamados índices geométricos (3D), denominados así, por proporcionar información estructural relacionada con la representación tridimensional de la estructura molecular.

Los descriptores geométricos pueden ser definidos siguiendo varias estrategias. Una de ellas es describir las moléculas por medio de los campos de interacción molecular, los cuales son determinados mediante la energía de interacción de las estructuras respecto a otros átomos conocidos como exploratorios *(atom probes)*. Estas evaluaciones de la energía de interacción es realizada en miles de puntos ubicados en una malla *(grid)*, donde las moléculas son embebidas [3, 12–14]. De hecho, el valor de energía en cada punto p de la malla depende de la orientación relativa del compuesto respecto a esta. Esta orientación es realizada mediante diferentes reglas de alineamiento de las moléculas respecto a la estructura de un receptor. Otra manera de definir índices 3D, es mediante la representación matricial de la geometría molecular [3]. En dicha representación, cada entrada de la matriz utilizada representa la distancia entre dos núcleos atómicos. Esta distancia, siempre es calculada a partir de las coordenadas Cartesianas (x, y, z) de los átomos y haciendo

uso únicamente de la distancia Euclidiana.

A pesar de que con los índices geométricos se han obtenidos buenos resultados en la predicción de propiedades o actividad biológica de diferentes estructuras químicas, los mismos tienen algunos inconvenientes o elementos que pueden ser mejorados. Puede mencionarse así, 1) que generalmente el alineamiento no es una característica deseable por no contarse siempre con información relacionada con la estructura ligando-receptor; 2) que la separación de los puntos en la grid en la cual el conjunto de datos es embebido, no es una dimensión estándar y por lo tanto varía según la metodología [13,15–17]; 3) que solamente utilizan la métrica Euclidiana para el cálculo de la distancia entre los átomos de una molécula; y 4) que todos los índices geométricos definidos, únicamente consideran relaciones entre dos átomos [3] y no entre n núcleos atómicos ($n > 2$).

En los reportes [4, 5, 18–22], Marrero-Ponce y colaboradores, introdujeron nuevos conjuntos de descriptores moleculares de relevancia para estudios QSAR/QSPR y diseño racional de fármacos, mediante los cuales describen características concernientes a la topología (2D) de las moléculas. Estos métodos codifican información de la estructura molecular por medio de las formas algebraicas lineal, bilineal y cuadrática utilizando matrices de densidad-electrónica grafo-teórica. Estos índices han sido extendidos a considerar características geométricas de moléculas de pequeño y mediano tamaño, basados en el enfoque de quiralidad que pueden presentar las estructuras químicas [21,23]. Sin embargo, y al igual que otros índices reportados, únicamente tienen en cuenta relaciones entre pares de átomos. En este caso específico, sólo aplican conceptos relacionados con el álgebra lineal (formas lineal, bilineal y cuadrática).

Por todo lo anteriormente planteado se tiene como **problema científico** el siguiente: *no se conocen índices geométricos independientes de alineamiento y que no estén basados en la utilización de mallas, que consideren relaciones n-dimensionales entre los átomos de una molécula y que utilicen otras medidas diferentes a la distancia Euclidiana como invariante respecto a la rotación y traslación de la estructura molecular.*

Este problema se desglosa en las siguientes **preguntas de investigación**:

1. ¿La utilización de otras métricas diferentes a la Euclidiana para el cálculo de la distancia entre pares de átomos mejora la calidad de los índices geométricos obtenidos respecto a los reportados?

2. ¿Cómo establecer relaciones entre más de dos átomos para calcular índices moleculares con características distintas a los definidos en la literatura?

3. ¿Qué efecto provoca aplicar las propuestas presentadas en la realización de estudios QSAR/QSPR?

Para darle solución al problema científico se planteó el siguiente **objetivo general** de investigación que consiste en: *desarrollar nuevos índices moleculares, mediante el establecimiento de relaciones entre dos o más átomos y haciendo uso de conceptos del álgebra lineal y multilineal, que doten a los investigadores en esta rama de otras alternativas para enfrentar los problemas complejos existentes, especialmente en el área de la predicción de actividad biológica de compuestos orgánicos.*

Este objetivo general fue desglosado en los siguientes **objetivos específicos**:

1. Proponer nuevos índices moleculares geométricos a partir de relaciones entre dos átomos que consideren métricas diferentes a la Euclidiana como medidas invariantes a la rotación y traslación de la estructura molecular.

2. Definir nuevos índices moleculares geométricos que consideren relaciones n-dimensionales entre los núcleos atómicos de una molécula.

3. Desarrollar un software denominado ToMoCoMD-CARDD QuBiLs MIDAS (acrónimo de Topological Molecular Computer Design - Computer Aided Rational Drug Desing) para automatizar el cálculo de los descriptores propuestos.

4. Evaluar los índices propuestos respecto a los reportados en la literatura acorde a la calidad de la información que estos capturan y a la utilidad en la predicción de la actividad biológica en compuestos orgánicos.

Después de haber evaluado el marco teórico se formularon las siguientes **hipótesis de investigación** como respuestas a las preguntas de investigación:

- H1: Incluir nuevas métricas para el cálculo de las distancias inter-atómicas, aumenta la variabilidad de los índices propuestos respecto a los reportados en la literatura y mejora el poder predictivo de los modelos obtenidos para la predicción de la actividad biológica.

- H2: Incorporar relaciones entre más de dos átomos en el cálculo de índices moleculares, mejora la codificación de la información de las estructuras químicas, obteniendo índices colineales y ortogonales respecto a los reportados en la literatura.

Esta investigación permite obtener resultados cuya **novedad científica** radica en la obtención de nuevos índices geométricos totales y locales a partir de los conceptos de formas lineales, bilineales y cuadráticas, haciendo uso de métricas diferentes a la Euclidiana para el cálculo de distancias inter-atómicas. Además, se

proponen varios operadores de agregación sobre los índices atómicos (denominados también como LOVIs: siglas en inglés de LOcal Vertex InvariantS) que generalizan la obtención de descriptores moleculares como una combinación lineal de los mismos. Por otra parte, se introducen por primera vez descriptores geométricos de dimensiones superiores mediante la aplicación de conceptos del álgebra multilineal estableciendo relaciones entre tres y cuatro átomos.

Este trabajo posee como **valor práctico** el desarrollo del software ToMoCoMD-CARDD QuBiLs MIDAS, donde se encuentra implementada toda la teoría correspondiente al cómputo de los nuevos índices moleculares. Además, tiene incorporado un módulo denominado "Estructura" que contribuye al curado y limpieza de las bases de datos de las estructuras químicas que pueden ser analizadas.

Como **valor metodológico** este trabajo aporta la aplicación de una serie de procedimientos para evaluar la calidad de nuevos descriptores moleculares, tales como, análisis de variabilidad basado en entropía de Shannon, análisis de componentes principales y estudios QSAR/QSPR. Además, el uso de operadores de agregación como generalización de la combinación lineal de las contribuciones atómicas puede utilizarse en descriptores moleculares cuyo cálculo puede ser descompuesto en índices atómicos.

La **tesis** está **estructurada** en cuatro capítulos. En el Capítulo 1 se tratan de manera general, los fundamentos de las formas lineales, bilineales y cuadráticas, así como la definición de formas n-dimensionales. También se presenta el marco experimental estadístico para validar los resultados obtenidos. Seguidamente, en el Capítulo 2, son definidos los procedimientos para el cálculo de los nuevos índices moleculares, detallando de esta forma el concepto de vector molecular y los enfoques matriciales que son utilizados. En el Capítulo 3 se aborda el diseño del software implementado y las pruebas de rendimiento realizadas a este. Finalmente en el Capítulo 4, se muestran los resultados de las diferentes validaciones y pruebas estadísticas realizadas a los índices propuestos. Este documento culmina con las Conclusiones, Recomendaciones, Referencias Bibliográficas y Anexos.

Capítulo 1

Fundamento teórico

En este capítulo son tratados los conceptos y definiciones relacionadas con los descriptores geométricos reportados en la literatura, así como su clasificación y características más relevantes. Son expuestas además, las diferentes variantes para representar matemáticamente la geometría molecular. Finalmente, se abordan las técnicas estadísticas utilizadas para validar los resultados de los índices geométricos que se proponen.

1.1. Descriptores geométricos

Los descriptores geométricos son definidos por diferentes vías pero siempre partiendo de la estructura tridimensional de la molécula [2]. En sentido general, los índices geométricos se calculan ya sea tomando como punto de partida alguna geometría molecular optimizada por métodos de química computacional, o a partir de coordenadas cristalográficas. Los índices geométricos pueden ser clasificados en dependientes de alineamiento y en libres de alineamiento [24]. Los primeros son desarrollados usando información concerniente de la geometría del receptor, mientras que los segundos utilizan poco o ningún conocimiento sobre este aspecto.

Dentro de los índices dependientes de alineamiento se encuentran metodologías como CoMFA (del ingés - Comparative Molecular Field Analysis) [13], la cual compara los campos de energía potencial de un conjunto de moléculas y busca diferencias y similitudes que puedan estar correlacionadas con las diferencias y similitudes en los valores de las propiedades consideradas. El primer paso de CoMFA, consiste en la selección de un grupo de compuestos que tengan un farmacóforo común, para la generación de estructuras tridimensionales de conformación razonable y para su alineamiento. Otra metodología no libre de alineamiento es Compass [12], quien en la primera etapa realiza la generación de conformaciones de baja energía para cada

molécula, y la selección de un confórmero como el más probable a ser biológicamente activo; todos estos confórmeros son alineados respecto al farmacóforo identificado o a una subestructura común para todas las moléculas en la data. Con características similares a los dos anteriores, puede mencionarse además el procedimiento CoMSIA (del inglés - Comparative Molecular Similarity Indices Analysis) [16], que mide la similitud de moléculas sobre la base de sus propiedades físico-químicas. En este último caso, el alineamiento molecular es realizado partiendo de una orientación aleatoria de dos moléculas.

Como puede notarse, los métodos dependientes del receptor se caracterizan por requerir del alineamiento de todas las moléculas del conjunto de datos, ya sea entre ellas mismas o con respecto a un componente de referencia o farmacóforo. Típicamente las moléculas son alineadas realizando un solapamiento de unidades estructurales comunes, constituyendo este proceso un paso inherente y generalmente crítico para estas metodologías [25–27]. El proceso de alineamiento es adecuado para conjuntos de datos que son estructuralmente cercanos, y es mucho más difícil de aplicar a datas diversas. Además, y como principal desventaja de estos métodos es que siempre requieren conocimiento sobre la estructura del complejo ligando-receptor, una condición que desafortunadamente no siempre se cumple.

Otra característica que no solamente presentan los métodos dependientes de alineamiento sino también algunos métodos libres de alineamiento, es que son técnicas basadas en malla *(grid)*. Una malla es un arreglo 3D de $N_x \times N_y \times N_z$ puntos, que no es más que una cuadrícula con N_x puntos a lo largo del eje de las X, N_y puntos a lo largo del eje de las Y y N_z puntos a lo largo del eje de las Z, donde cada punto p es caracterizado por las coordenadas Cartesianas (x, y, z) en el espacio 3D. La malla debe ser seleccionada para embeber todos los átomos de todos los compuestos del conjunto de datos, o al menos cubrir determinadas regiones de interés. La densidad de la malla debe ser adecuada para muestrear las energías potenciales de los campos escalares teóricamente continuos.

Basados en esta característica, se pueden encontrar algunos procedimientos dependientes de alineamiento, tales como GRID [28], el cual es utilizado para detectar sitios de unión favorables en una molécula de estructura conocida. En este método, una molécula pequeña, por ejemplo el agua *(probe)*, es usada para generar los valores de energía de interacción en todos los puntos de la malla. Típicamente el espaciamiento que este procedimiento utiliza es de 0,5$\mathring{A}$. En el caso de CoMFA [13], es confeccionada una malla alrededor de los compuestos seleccionados. La distancia de los puntos en la malla son arbitrariamente escogidos, estableciendo 2$\mathring{A}$ por defecto; teniendo en consideración que tamaños pequeños conllevará a un número demasiado grande de puntos. Por su parte la metodología CoMSIA [16], utiliza una malla espaciada regularmente para valores entre 1,1$\mathring{A}$ y 2$\mathring{A}$; donde los puntos corresponden a valores de similitud entre las

moléculas y los átomos de exploración *(probe atoms)*.

Por otro lado, existen procedimientos que no siguen reglas de alineamiento, pero que su funcionamiento es basado en la utilización de una malla. De esta manera se tiene por ejemplo, el método GRIND [17], que está derivado de los campos de interacción molecular (MIF) y que por defecto utiliza $0{,}5\mathring{A}$ para la distancia entre los puntos de la malla. También se encuentra el procedimiento conocido como Análisis de Campos de Voronoi (Voronoi Field Analysis) [14], cuyo objetivo es el de reducir el gran número de valores de energía de interacción asignados a los puntos establecidos en la malla. Estos valores se asignan a cada poliedra de Voronoi [29], en la cual el espacio molecular superpuesto es descompuesto. Luego de esta descomposición, es definida una malla conteniendo exactamente la superficie molecular expandida, con puntos espaciados a $0{,}3\mathring{A}$.

Por lo tanto y acorde a los ejemplos de métodos basados en esta técnica, puede apreciarse que el hándicap de estas metodologías consiste en la selección adecuada de las distancias en las separaciones de la malla, lo cual requiere por lo general de un estudio estadístico.

Uno de los enfoques para superar el inconveniente de alineamiento molecular, es mediante el uso de un modelo 3D a partir de las coordenadas Cartesianas, como un medio para establecer medidas invariantes a la rotación y traslación de la estructura química. El enfoque MS-WHIM [30] (y también el WHIM [31]) soluciona el problema de alineamiento molecular mediante el cálculo de parámetros estadísticos a partir de una matiz de puntuación obtenida de un análisis de componentes principales. Métodos basados en autocorrelación de ciertas propiedades moleculares representan otros tipos de enfoques no sensibles al alineamiento [32]. Finalmente, los descriptores GRid-INdependent (GRIND) [17], 3D MoRSE (del inglés - Molecule Representation of Structures based on Electron diffraction) [33], GETAWAY (del inglés - GEometry, Topology, and Atom-Weights AssemblY) [34], RDF (del inglés - Radial Distribution Functions), entre otros, constituyen también ejemplos significativos de este tipo de índices.

Por otro lado, durante las últimas dos décadas, existió una tendencia a extender los índices topológicos a tener en consideración la representación 3D de la molécula, mediante la inclusión de información geométrica. Entre tales índices se pueden mencionar los índices 3D-Wiener [35], 3D-Balaban [36], Gravitacional [37], 3D-Petitjean [38], entre otros [3]. Estos descriptores son libres de alineamiento, fáciles y rápidos de calcular. Algunos de estos índices tienen como característica común que solamente utilizan la matriz de distancia geométrica o sus derivaciones (ver Sección 1.2), o siempre emplean la distancia Euclidiana como única métrica para establecer una medida invariante a la rotación y traslación de la estructura molecular.

1.2. Representación de la geometría molecular

Las representaciones matriciales, de forma general, constituyen una de las fuentes más importantes de índices moleculares, puesto que proporcionan una descripción numérica de los grafos moleculares. La descripción numérica de la estructura de los compuestos químicos, es esencial para la manipulación computacional de las moléculas y el cálculo de los índices que de ellas se derivan.

Para la caracterización de la topología molecular han sido propuestas varias matrices, entre las que se destacan la matriz Laplaciana, la matriz de Detour [39], la matriz de distancia-valencia [40], la matriz de resistencia-distancia [41], la matriz de conductancia eléctrica [41], entre otras. Sin embargo, estas matrices no son útiles para representar los aspectos geométricos de las estructuras químicas.

Por tal razón, una de las formas de representación de la estructura tridimensional de una molécula, es mediante la matriz molecular M [3] (ver Figura 1.1), la cual es una matriz rectangular de dimensión $A \times 3$, donde las filas representan los átomos de la molécula, y las columnas las coordenadas cartesianas (x, y, z). Como puede observarse, la matriz molecular M no contiene información relacionada de los átomos adyacentes, razón por la cual es usualmente ampliada y denotada como M' (ver Figura 1.1). Esta matriz M' es obtenida por la unión de la matriz molecular y la tabla de conectividad, donde la primera columna denota el tipo de átomo y las últimas cuatro contienen las etiquetas de los átomos conectados al *i-ésimo* átomo.

$$\mathbf{M} = \begin{vmatrix} x_1 & y_1 & z_1 \\ x_2 & y_2 & z_2 \\ \dots & \dots & \dots \\ x_A & y_A & z_A \end{vmatrix} \qquad \mathbf{M'} = \begin{vmatrix} at.\,1 & x_1 & y_1 & z_1 & c_{11} & c_{12} & c_{13} & c_{14} \\ at.\,2 & x_2 & y_2 & z_2 & c_{21} & c_{22} & c_{23} & c_{24} \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ at.\,A & x_A & y_A & z_A & c_{A1} & c_{A2} & c_{A3} & c_{A4} \end{vmatrix}$$

Figura 1.1: Matriz molecular utilizada pra representar la estructura tridimensional de una molécula

Una alternativa a la representación de la estructura química mediante la matriz molecular, es la matriz de coordenadas internas Z [3] (ver Figura 1.2), donde se tienen en consideración la posición relativa de cada átomo respecto a los otros átomos en la molécula. Estas coordenadas son: las distancias de los enlaces, los ángulos de enlaces, y los ángulos de torsión. La distancia de los enlaces, r_{st}, es la distancia inter-atómica entre dos núcleos atómicos enlazados (generalmente expresada en Angstrom); los ángulos de enlaces, ϑ_{stv}, es el ángulo plano entre tres átomos conectados (s, t, v) dentro de una molécula; y los ángulos de torsión, ω_{stvz}, es el ángulo diedro entre cuatro átomos conectados (s, t, v, z).

$$Z = \begin{vmatrix} at.1 & & & & 0 & 0 & 0 \\ at.2 & r_{12} & & & 1 & 0 & 0 \\ at.3 & r_{23} & \vartheta_{321} & & 2 & 1 & 0 \\ at.4 & r_{34} & \vartheta_{432} & \omega_{4321} & 3 & 2 & 1 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ at.A & r_{As} & \vartheta_{Ast} & \omega_{Astv} & s & t & v \end{vmatrix}$$

Figura 1.2: Matriz de coordenadas internas utilizada pra representar la estructura tridimensional de una molécula

La matriz molecular M y la matriz Z son el punto de partida del cálculo de varios descriptores 3D moleculares, tales como: descriptores químico-cuánticos [42], de interacción molecular [43], y de superficie molecular [30, 44]; así como de los descriptores EVA [45], WHIM [31], GETAWAY [34], CoMMA [15] y Compass [12].

Otra forma de representar la estructura tridimensional de una molécula y que también constituye una fuente común para el cálculo de descriptores 3D, es la conocida matriz geométrica [3]. La matriz geométrica (G), o matriz de distancia geométrica de una molécula, es obtenida a partir de la matriz molecular M, y es una matriz simétrica cuadrada $A \times A$ donde cada entrada r_{st} es la distancia geométrica calculada como la distancia Euclidiana entre el átomo s y t, y donde además los elementos de la diagonal son siempre cero (ver Figura 1.3).

$$G \equiv \begin{vmatrix} 0 & r_{12} & \dots & r_{1A} \\ r_{21} & 0 & \dots & r_{2A} \\ \dots & \dots & \dots & \dots \\ r_{A1} & r_{A2} & \dots & 0 \end{vmatrix}$$

Figura 1.3: Matriz de distancias geométrica utilizada pra representar la estructura tridimensional de una molécula

Al igual que la matriz molecular, la matriz de distancia geométrica contiene información sobre la conformación y configuración de la estructura química, sin embargo, no contiene información acerca de la conectividad de los átomos. Por lo tanto, para varias aplicaciones, la matriz G es acompañada por una tabla de conectividad, donde para cada átomo, se encuentran representados los números de identificación de los núcleos atómicos adyacentes. A partir de la matriz de distancia geométrica, son derivadas otras representaciones matriciales, tales como la matriz de adyacencia de longitud de enlace ponderada [36], la

matriz geométrica de vecindad [46], las matrices cociente de distancia geométrica/distancia topológica [47], la matriz combinada distancia-distancia [3], entre otras.

1.2.1. Distancias inter-atómicas

Como ha sido expuesto anteriormente, la matriz de distancia geométrica solamente utiliza la distancia Euclidiana para codificar interacciones no covalentes dentro de una molécula. Hoy en día, el concepto de distancia usualmente significa algún grado de cercanía entre dos objetos físicos o ideas, mientras que el término de métrica es generalmente utilizado como un estándar para una medida [48].

Formalmente, sea N un conjunto de elementos, y una función D: $N \times N \rightarrow \mathbb{R}$ con las siguientes propiedades, para todo a, b, $c \epsilon N$:

1. $D(a, b) \geq 0$
2. $D(a, b) = D(b, a)$
3. $D(a, a) = 0$
4. $D(a, b) \leq D(a, c) + D(c, b)$

Si D cumple las propiedades 1-3, es denominada *métrica* sobre N, mientras si D cumple con los axiomas del 1-4, entonces es nombrada como *métrica de distancia*. Además, un coeficiente de distancia el cual tiene sólo las últimas tres propiedades es nombrado como *pseudométrica*, y si no tiene la cuarta propiedad *no es métrica*. Aunque una función de distancia particular tenga las cuatro propiedades, esto no es suficiente para implicar que las distancias involucradas puedan ser embebidas en un espacio Euclidiano para cualquier dimensión [48].

Para calcular la distancia entre dos átomos, son consideradas las coordenadas Cartesianas (x, y, z) de cada núcleo atómico. Estas coordenadas son variables continuas, siendo medidas con la distancia más común para este tipo de variables, que es la distancia Euclidiana. Es importante resaltar, que hasta el momento, la distancia Euclidiana ha sido considerada prácticamente como la métrica exclusiva para establecer una relación entre pares de átomos (distancia inter-atómica) en el cálculo de descriptores moleculares 3D, a pesar de que no existe postulado teórico que sostenga a esta métrica como la más adecuada.

Realmente, la definición general de distancia depende del espacio y la métrica. Por lo tanto, si una molécula está en el espacio Euclidiano, es posible generalizar la distancia entre los átomo i y j a través de la distancia de Minkowski [49]:

$$d_{ij} = (|x_i - x_j|^p + |y_i - y_j|^p + |z_i - z_j|^p)^{\frac{1}{p}} \tag{1.1}$$

donde, x, y y z representan las coordenadas Cartesianas, y p es la norma de la distancia de Minkowski (e.j., $p = 1$ es la distancia de Manhattan, $p = 2$ es la bien conocida distancia Euclidiana, y $p = \infty$ es la distancia de Chebyshev). Puede notarse que una *"matriz de distancia de Minkowski"*, con elementos definidos en la ecuación 1.1, sería el caso más general de la *matriz de distancia geométrica* ($p = 2$) utilizada hasta el momento.

Además, existen un conjunto de métricas las cuales han sido usadas satisfactoriamente en algoritmos de aprendizaje automatizado y en estudios de similitud [50–52]. Así por ejemplo se tiene: la distancia de Canberra [53], la cual ha sido utilizada en aprendizaje automatizado para biología computacional [54] y en detección de intrusos en seguridad informática [55]; la distancia de Bhattacharyya [56], que mide la similitud de dos distribuciones de probabilidad continuas o discretas y ha sido utilizada en problemas de extracción de características [57]; la distancia Lance-Williams o Bray-Curtis, que es utilizada para cuantificar la disimilitud composicional entre dos sitios diferentes y ha sido empleada en problemas de ordenamiento ecológico [58, 59]; y la métrica Separación Angular [60], que mide la similitud entre dos vectores mediante el coseno del ángulo formado entre ellos, y es comúnmente utilizada en recuperación de información [61], donde por ejemplo, un documento está caracterizado por un vector donde el valor de cada dimensión corresponde al número de veces que un término aparece en el documento.

Estas métricas mencionadas con anterioridad, junto con otras reportadas en la literatura, pudieran ser usadas para calcular distancias inter-atómica, y de esta forma servir como esquemas de generalización para la distancia geométrica de todos los pares de átomos i, j de una molécula.

1.2.2. Esquemas de normalización

Como se puede apreciar, las matrices constituyen la herramienta matemática más común para codificar información estructural de moléculas [3], teniendo especial interés las relacionadas con la geometría molecular. Sin embargo, es inusual utilizar transformaciones probabilísticas sobre estas matrices. A pesar de ello, matrices estocásticas son definidas en el marco de trabajo MARCH-INSIDE [62, 63] y en los descriptores ToMoCoMD-CARDD 2D [20, 23].

En el caso de los descriptores moleculares MARCH-INSIDE, estos codifican información relativa a la distribución de los electrones en la molécula, basado en un enfoque simple estocástico acerca de la idea de

ecualización de la electronegatividad (principio de Sanderson) [64]. Por su parte, los índices ToMoCoMD-CARDD 2D definen la *matriz estocástica de densidad electrónica grafo-teórica*, la cual describe cambios en la distribución de los electrones en el tiempo a través de un backbone molecular. Para esta matriz es considerado un caso hipotético, en el cual un conjunto de átomos están inicialmente libres en el espacio y distribuidos alrededor de los núcleos atómicos. En este sentido, los electrones en un átomo arbitrario pueden moverse a otros átomos diferentes mediante una red de enlace químico.

Por otra parte, Carbó-Dorca [65] también empleó un escalamiento estocástico aplicado a Matrices de Similitud Cuántica (QSM), bajo el principio de que cualquier fila o columna de una matriz QSM puede fácilmente convertirse en una distribución de probabilidad discreta, realizando así un escalamiento simple estocástico, donde la suma de los elementos de cada fila (o columna) son usados como un factor de escala.

Formalmente, pueden definirse las matrices estocásticas como matrices cuadradas para las cuales la suma de cada fila (matrices estocásticas derechas), o la suma de cada columna (matrices estocásticas izquierdas), es igual a 1, significando esto que los elementos de las filas o columnas consisten en número reales no negativos que pueden ser interpretados como probabilidades [66].

1.3. Formas algebraicas

En matemática, una aplicación $f : V \to F$ se denomina *isomorfismo* entre los espacios vectoriales V y F sobre un mismo campo escalar K, si se cumple que f es biyectiva y además los siguientes axiomas:

$$\begin{aligned} f(\bar{v} + \bar{w}) &= f(\bar{v}) + f(\bar{w}) \\ f(\lambda \bar{v}) &= \lambda f(\bar{v}) \end{aligned} \tag{1.2}$$

para todo $\bar{v}$, $\bar{w} \epsilon V$ y $\lambda \epsilon K$. Esto significa que si se toman dos elementos $\bar{v}$, $\bar{w}$ en V y se determina su suma $\bar{v} + \bar{w}$, esto corresponde en F a buscar las imágenes de los elementos y sumarlas. Lo mismo sucede con el producto por un escalar, debido a que si la imagen $f(\bar{v})$ de un vector $\bar{v}$, se desea multiplicar por un escalar λ, basta multiplicar $\bar{v}$ por λ y buscar la imagen $\lambda \bar{v}$.

Esta aplicación f sólo realiza transformaciones si el espacio de llegada es una copia del espacio de partida. Sin embargo, en ocasiones surgen correspondencias entre espacios que no son idénticos estructuralmente, pero estas correspondencias trasladan en cierta medida, la estructura de uno de los espacios a otro (por ejemplo, $f : R^3 \to R^2$). A estas funciones se les denominan aplicaciones lineales, y las mismas cumplen también con los axiomas de linealidad especificados en la Ecuación 1.2. En particular, si el espacio de llegada

de una aplicación lineal es un número real ($f : E \to \mathbb{R}$), entonces se le conoce como *forma lineal*.

En un espacio vectorial V de dimensión finita n, la forma lineal queda completamente determinada por los valores que se asignan a los elementos de una base. De esta manera, si $E = \{\bar{e}_1, \bar{e}_2, \ldots, \bar{e}_n\}$ es una base de V, y cualquier vector $\bar{x}\,\epsilon\,V$ está determinado por $\bar{x} = \sum_{j=1}^{n} \alpha^j \bar{e}_j$, entonces:

$$f(\bar{x}) = f\left(\sum_{j=1}^{n} \alpha^i \bar{e}_j\right) = \sum_{j=1}^{n} \alpha^i f(\bar{e}_j) \tag{1.3}$$

donde, $\left(\alpha^1, \alpha^2, \ldots, \alpha^n\right)\,\epsilon\,\mathbb{R}^n$ son las coordenadas del vector $\bar{x}$ en la base E, y $f(\bar{e}_j)$ puede expresarse respecto a la base $B = \left\{\bar{b}_1, \bar{b}_2, \ldots, \bar{b}_m\right\}$ del espacio vectorial F de dimensión m, como sigue:

$$f(\bar{e}_j) = \sum_{i=1}^{m} e_{ij} \bar{b}_i \tag{1.4}$$

y sustituyendo 1.4 en 1.3 se obtiene:

$$f(\bar{x}) = \sum_{j=1}^{n} \alpha^i \sum_{i=1}^{m} e_{ij} \bar{b}_i = \sum_{i=1}^{m} \sum_{j=1}^{n} e_{ij} \alpha^i \bar{b}_i \tag{1.5}$$

De manera análoga a la forma algebraica lineal, es definida la *forma bilineal*, como una aplicación $b : V \times V \to \mathbb{R}$, la cual es lineal en todos sus argumentos tomados separadamente. Es decir, que esta función satisface las condiciones de linealidad mostradas en la Ecuación 1.6, para cualquier escalar λ y cualquier vector $\bar{v}$, $\bar{w}$, $\bar{v}_1$, $\bar{v}_2$, $\bar{w}_1$ y $\bar{w}_2$ que pertenecen al espacio vectorial V.

$$\begin{aligned} b(\lambda\bar{v},\, \bar{w}) &= b(\bar{v},\, \lambda\bar{w}) = \lambda b(\bar{v},\, \bar{w}) \\ b(\bar{v}_1 + \bar{v}_2,\, \bar{w}) &= b(\bar{v}_1,\, \bar{w}) + b(\bar{v}_2,\, \bar{w}) \\ b(\bar{v},\, \bar{w}_1 + \bar{w}_2) &= b(\bar{v},\, \bar{w}_1) + b(\bar{v},\, \bar{w}_2) \end{aligned} \tag{1.6}$$

Por lo tanto, si V es un espacio vectorial real en $\mathbb{R}^n$ ($V\,\epsilon\,R^n$) y el conjunto de vectores $E = \{\bar{e}_1, \bar{e}_2, \ldots, \bar{e}_n\}$ es un sistema base de $\mathbb{R}^n$, se pueden definir formas no ambiguas para cualquier vector $\bar{x}$ y $\bar{y}$ de V, siendo sus coordenadas $\left(x^1, x^2, \ldots, x^n\right)\,\epsilon\,R^n$ y $\left(y^1, y^2, \ldots, y^n\right)\,\epsilon\,\mathbb{R}^n$, respectivamente. Esto significa que los vectores $\bar{x}$ y $\bar{y}$ son expresados como combinaciones lineales respecto a la base E del espacio vectorial V, como es mostrado a continuación:

$$\bar{x} = \sum_{i=1}^{n} x^i \bar{e}_i \qquad \bar{y} = \sum_{j=1}^{n} y^j \bar{e}_j \tag{1.7}$$

Por consiguiente,

$$b\left(\bar{x},\,\bar{y}\right)=b\left(x^{i}\bar{e}_{i},\,y^{j}\bar{e}_{j}\right)=x^{i}y^{j}b\left(\bar{e}_{i},\,\bar{e}_{j}\right) \tag{1.8}$$

y si son tomados los coeficientes a_{ij} como $n\times n$ escalares de $b\left(\bar{e}_{i},\,\bar{e}_{j}\right)$, es decir,

$$a_{ij}=b\left(\bar{e}_{i},\,\bar{e}_{j}\right)\qquad\forall\, i=1,\,2,\,\ldots,\,n\,\wedge\,\forall\, j=1,\,2,\,\ldots,\,n \tag{1.9}$$

entonces,

$$b\left(\bar{x},\,\bar{y}\right)=\sum_{i=1}^{n}\sum_{j=1}^{n}a_{ij}x^{i}y^{j}=\left[X\right]^{T}A\left[Y\right]=\left[\begin{array}{ccc} x^{1} & \ldots & x^{n}\end{array}\right]\left[\begin{array}{ccc} a_{11} & \ldots & a_{jn}\\ \ldots & \ldots & \ldots\\ a_{in} & \ldots & a_{nn}\end{array}\right]\left[\begin{array}{c} y^{1}\\ \vdots\\ y^{n}\end{array}\right] \tag{1.10}$$

De esta manera, la ecuación especificada para la forma bilineal, puede ser escrita como una ecuación matricial simple (ver Ecuación 1.10), donde $[Y]$ es un vector columna de las coordenadas de $\bar{y}$ sobre un sistema base canónico de $\mathbb{R}^n$, y $[X]^T$ es la transpuesta del vector columna $[X]$ de las coordenadas de $\bar{x}$ en la misma base de $\mathbb{R}^n$. Finalmente, una forma bilineal b es simétrica si $b\left(\bar{x},\,\bar{y}\right)=b\left(\bar{y},\,\bar{x}\right)\;\;\forall\,\bar{x},\,\bar{y}\,\epsilon\,V$.

Por otro lado y a partir de la definición simétrica de una forma bilineal, es determinada la *forma cuadrática*, como una función $q:\,V\rightarrow\mathbb{R}$, dada por:

$$q\left(\bar{x}\right)=b\left(\bar{x},\,\bar{x}\right) \tag{1.11}$$

donde, $q\left(\lambda\bar{x}\right)=\lambda^{2}q\left(\bar{x}\right)\qquad\forall\,\lambda\,\epsilon\,\mathbb{R},\,\bar{x}\,\epsilon\,V$.

1.3.1. Formas algebraicas n-dimensionales

Sea V un espacio vectorial *n-dimensional* sobre el campo de los número reales $\mathbb{R}$, y $\wedge^{n}V^{*}$ denota el espacio de las *formas n-dimensionales* en V . Se puede definir como *forma n-dimensional* a una función $w:\,V_1\times V_2\times\ldots\times V_n\rightarrow\mathbb{R}$ que cumple las siguientes propiedades:

$$
\begin{aligned}
w\left(\lambda\bar{v}_1,\, \bar{v}_2,\, \ldots,\, \bar{v}_n\right) &= w\left(\bar{v}_1,\, \lambda\bar{v}_2,\, \ldots,\, \bar{v}_n\right) = w\left(\bar{v}_1,\, \bar{v}_2,\, \ldots,\, \lambda\bar{v}_n\right) = \lambda w\left(\bar{v}_1,\, \bar{v}_2,\, \ldots,\, \bar{v}_n\right) \\
w\left(\bar{v}_1 + \bar{v}_3,\, \bar{v}_2,\, \ldots,\, \bar{v}_n\right) &= w\left(\bar{v}_1,\, \bar{v}_2,\, \ldots,\, \bar{v}_n\right) + w\left(\bar{v}_3,\, \bar{v}_2,\, \ldots,\, \bar{v}_n\right) \\
w\left(\bar{v}_1,\, \bar{v}_2 + \bar{v}_3,\, \ldots,\, \bar{v}_n\right) &= w\left(\bar{v}_1,\, \bar{v}_2,\, \ldots,\, \bar{v}_n\right) + w\left(\bar{v}_1,\, \bar{v}_3,\, \ldots,\, \bar{v}_n\right) \\
&\vdots \\
w\left(\bar{v}_1,\, \bar{v}_2,\, \ldots,\, \bar{v}_n + \bar{v}_3\right) &= w\left(\bar{v}_1,\, \bar{v}_2,\, \ldots,\, \bar{v}_n\right) + w\left(\bar{v}_1,\, \bar{v}_2,\, \ldots,\, \bar{v}_3\right)
\end{aligned}
\tag{1.12}
$$

donde, $\bar{v}_1$, $\bar{v}_2$, $\bar{v}_3$, $\bar{v}_n$ son vectores del espacio $\wedge^n V^*$.

Por lo tanto si $\wedge^n V^* \,\epsilon\, \mathbb{R}$ y el sistema de vectores $E = \{\bar{e}_1,\, \bar{e}_2,\, \ldots,\, \bar{e}_n\}$ es una base de $\mathbb{R}^n$, entonces pueden definirse vectores $\bar{w}_1,\, \bar{w}_2,\, \ldots,\, \bar{w}_n$ como combinaciones lineales respecto a la base E. De esta forma se tiene:

$$
\bar{w}_1 = \sum_{i=1}^{n} w_1^i \bar{e}_i \qquad \bar{w}_2 = \sum_{j=1}^{n} w_2^j \bar{e}_j \quad \ldots \quad \bar{w}_n = \sum_{k=1}^{n} w_n^k \bar{e}_k \tag{1.13}
$$

Por consiguiente,

$$
w\left(\bar{w}_1,\, \bar{w}_2,\, \ldots,\, \bar{w}_n\right) = w\left(w_1^i \bar{e}_i,\, w_2^j \bar{e}_j,\, \ldots,\, w_n^k \bar{e}_k\right) \tag{1.14}
$$

y si son considerados los elementos $a_{ij\ldots k}$ como escalares $n \times n \times \ldots \times n$ de $w\left(\bar{e}_i,\, \bar{e}_j,\, \ldots,\, \bar{e}_k\right)$, entonces

$$
w\left(\bar{v}_1,\, \bar{v}_2,\, \ldots,\, \bar{v}_n\right) = A_{ij\ldots k} \cdot \bar{w}_1^{ij\ldots k} \cdot \bar{w}_2^{ij\ldots k} \cdot \ldots \cdot \bar{w}_n^{ij\ldots k} \tag{1.15}
$$

1.4. Métodos estadísticos en Informática Química

La informática química es una disciplina que recopila herramientas matemáticas y estadísticas para tratar con datas químicas complejas [67–70]. Estas técnicas son utilizadas para la recopilación, la elaboración, el análisis y la caracterización de conjuntos de datos, de forma tal que se pueda obtener información útil de ellas. Las mismas, hoy en día, se interceptan no solo con varios campos de la Matemática y la Estadística clásica, sino también de la Inteligencia Artificial (IA) y otras ramas de la ciencia de la computación [71, 72]. En esta sección serán presentadas solo aquellas técnicas que son de interés en el presente trabajo.

1.4.1. Análisis de Variabilidad

El método de Análisis de Variabilidad, propuesto por Godden y colaboradores [73, 74], cuantifica el contenido de información y, por lo tanto, la variabilidad de los descriptores moleculares (DMs). Este método no supervisado está basado en el cálculo de la Entropía de Shannon (SE) [75], bajo el principio de que variables deseables para análisis quimio-métricos pudieran poseer elevados valores de entropía como un indicador de su tendencia a cambiar gradualmente con la modificación de la estructura molecular; mientras que variables redundantes pudieran tener valores bajos, siendo cero el límite para aquellas variables que contienen el mismo valor para estructuras diferentes. Esta técnica permite evaluar la calidad de los DMs como entidades independientes y ha sido utilizada en la literatura para comparar el desempeño de conjuntos de DMs implementados en diferentes paquetes computacionales, así como en estudios de diversidad molecular [73, 76].

1.4.2. Análisis de Componentes Principales

El Análisis de Componentes Principales (ACP) es un procedimiento matemático que transforma un conjunto de variables correlacionadas de respuesta en un conjunto menor de variables no correlacionadas, llamadas *componentes principales* [77,78]. Este procedimiento es útil cuando al trabajar con varias variables (posiblemente con un gran número de variables), se cree que existe cierto grado de redundancia en las mismas. En este caso, redundancia significa que algunas variables están correlacionadas con otras, tal vez porque están midiendo un mismo hecho. Debido a esta redundancia, puede reducirse las variables observadas en *componentes principales*, de forma tal, que en su totalidad expliquen la mayor varianza posible de las variables originales.

El primer *componente* extraído generalmente tiene en consideración la mayor varianza de todos los componentes determinados en el estudio. Por su parte el segundo *componente*, tendrá dos características fundamentales: 1) que explicará la mayor varianza posible que el primero no tuvo en cuenta, y 2) que no está correlacionado con el primero, es decir, la variables cargadas en él son ortogonales a las cargadas en el primer *componente*. Estas dos características son presentadas por el resto de los componentes determinados en el análisis.

Por lo tanto, si se tiene en cuenta que en estudios de informática química se trabaja con un gran conjunto de variables (índices moleculares), el uso del ACP resulta ser de gran utilidad, dado que permite explorar a priori la posible existencia de no colinealidad entre las variables, un requisito importante para nuevos

descriptores moleculares y también en la modelación de las propiedades físico-químicas y biológicas.

1.4.3. Regresión Lineal Múltiple

La Regresión Lineal Múltiple (RLM) estudia las relaciones entre una variable dependiente o criterio y un conjunto de variables independientes o explicativas. Este modelo puede ser expresado como:

$$Y = a + b_1X_1 + b_2X_2 + \ldots + b_nX_n \tag{1.16}$$

donde, Y es la variable dependiente o explicada, a es la intersección o término constante, X_1, X_2, ..., X_n son la variables independientes o explicativas, y b_1, b_2, ..., b_n son los parámetros que miden la influencia de las variables explicativas tienen sobre el regresando.

1.4.3.1. Análisis de la varianza

El ANOVA (ANalysis Of VAriance) sirve para comprobar la hipótesis de que $R^2 = 0$. La variabilidad total de la variable dependiente se divide entre la parte atribuible a la regresión y la parte residual. La distancia de un punto cualquiera Y_i a la $\bar{Y}$ se sub-divide en dos partes [79]:

$$Y_i - \bar{Y} = \left(Y_i - Y''\right) + \left(Y'' - \bar{Y}\right) \tag{1.17}$$

donde, Y'' es el valor predicho, Y_i el valor observado, y $\bar{Y}$ es la media de la variable dependiente. La diferencia $\left(Y_i - Y''\right)$ se denomina *residuo de la regresión*, y la diferencia $\left(Y'' - \bar{Y}\right)$ corresponde a la distancia explicada por la regresión, y representa el aumento en la estimación del Y_i mediante la recta de regresión.

En el ANOVA, F viene dada por:

$$F = \frac{MC_E}{MC_R} \tag{1.18}$$

donde, F sigue una distribución F de *Fisher-Snedecor* con grados de libertad $gl_E = \upsilon - 1$, $gl_R = n - \upsilon$ (υ el número de variables de la ecuación y n la cantidad de instancias), y la *media cuadrática* (MC) se obtiene dividiendo la suma de cuadrados del ANOVA entre los grados de libertad. La F sirve para comprobar si el modelo de regresión se ajusta a los datos y permite evaluar si se rechaza la hipótesis nula, según la cual $R^2 = 0$. Si el modelo se ajusta a los datos, el coeficiente de determinación $\left(R^2\right)$ puede calcularse a partir

de la *suma de cuadrados* (SC) del ANOVA mediante:

$$R^2 = 1 - \frac{SC_R}{SC_{total}} \tag{1.19}$$

Es importante señalar que, la mayoría de las investigaciones QSAR/QSPR han sido realizadas usando la técnica de RLM [80, 81], fundamentalmente por su carácter lineal, paramétrico y su "simplicidad".

1.4.3.2. Validación interna por las técnicas de validación cruzada, re-muestreo y revuelto. Validación externa

La validación cruzada (cross-validation) [82, 83] es una técnica utilizada para evaluar los resultados de un análisis estadístico y garantizar que son independientes de la partición entre datos de entrenamiento y prueba. Consiste en hacer un número (M) de reducidas modificaciones al conjunto de compuestos de la data original, y entonces calcular la precisión de las predicciones sobre las diferentes particiones [82, 83]. Este procedimiento se repite para cada conjunto de datos modificados. El poder predictivo del modelo puede expresarse como Q^2, denominado como la *"varianza predictiva"* o la *"varianza de la validación cruzada"*, la cual puede ser calculada acorde a la siguiente fórmula:

$$Q^2 = 1 - \frac{\sum_{i=1}^{n} \left(Y_i - Y_i^{''}\right)^2}{\sum_{i=1}^{n} \left(Y_i - \bar{Y}\right)} \tag{1.20}$$

donde, Y_i, $Y^{''}$, $\bar{Y}$ es la respuesta observada, estimada y media del *i-ésimo* caso respectivamente.

Cuando se utiliza un solo compuesto en cada grupo de VC (lo cual da N grupos), el procedimiento se conoce como *dejar "uno" fuera* (LOO, acrónimo de Leave-One-Out). No obstante, Shao ha mostrado que desde el punto de vista teórico y práctico, el procedimiento de *dejar "varios" fuera* (LSO, acrónimo de Leave-Several-Out) es preferible al LOO [84].

En la técnica de re-muestreo (bootstrap) [85, 86], el tamaño original de la data (n) es preservado para el conjunto de entrenamiento, mediante la selección de n objetos con repetición, de manera que, el conjunto de entrenamiento contiene algunos objetos repetidos y el conjunto de evaluación está constituido por lo objetos no seleccionados. Este procedimiento de construir conjuntos de entrenamiento y de predicción aleatoriamente es repetido miles de veces, y en cada iteración se calculan los estadísticos relevantes y el promedio de estos constituye el estimado del re-muestreo.

El método del revuelto (Y-scrambling) es empleado para evaluar la correlación al azar [87, 88]. En esta técnica, se calcula un modelo lineal de regresión para la verdadera variable respuesta (Y), junto con un número de regresiones repetidas (200-300 veces) con las mismas variables pero con la variable dependiente aleatoriamente cambiada $\left(\widetilde{Y}\right)$. Luego se calcula para cada modelo la varianza explicada Q^2_{loo}, y se evalúa la correlación entre la respuesta verdadera y la revuelta de la siguiente manera:

$$Q^2_k = a + b * r_k\left(Y, \widetilde{Y}\right) \tag{1.21}$$

donde, Q^2_k es la varianza explicada para el modelo obtenido con los mismos predictores teniendo el *k-ésimo* vector revuelto, y r_k es la correlación entre los vectores para la respuesta verdadera y la *k-ésima* revuelta. Un valor del intercepto cercano a cero implica que el modelo no es obtenido al azar, mientras que un intercepto grande indica que los modelos aleatorios poseen el mismo desempeño que el modelo verdadero, razón por la cual no tendría buena calidad y se pudiera considerar que las variables están aleatoriamente correlacionadas.

La validación externa permite evaluar si los modelos obtenidos son generalizables a nuevos compuestos químicos, y de esta forma analizar el "verdadero" poder predictivo de los mismos [87]. Para esto se divide la data en 2 conjuntos: la serie de entrenamiento para construir el modelo, y la serie de predicción, que no es utilizada en la selección de variables ni en el desarrollo del modelo, sino usada exclusivamente para evaluar el modelo tras su formación. El estadístico utilizado para comparar la capacidad predictiva de los modelos de regresión es el conocido $\mathrm{SDEP^{ext}}$ (desviación estándar de los errores en la predicción) [89], el cual es calculado como se muestra a continuación:

$$SDEP^{ext} = \sqrt{\frac{\sum_{i=1}^{n_{ext}}\left(Y_i - Y_i^{''}\right)^2}{n_{ext}}} \tag{1.22}$$

donde, n_{ext} es la cantidad de objetos o instancias en el conjunto de prueba o predicción, y Y_i, $Y_i^{''}$ es la respuesta estimada y predicha del *i-ésimo* caso respectivamente.

1.5. Conclusiones parciales

En el presente capítulo fueron expuestas las principales características de los índices geométricos definidos hasta la fecha, así como las formas de representación de la geometría molecular. Por último, se mencionaron

las técnicas estadísticas a utilizar para validar los descriptores QuBiLs MIDAS que se proponen en el presente trabajo. Como conclusión se puede decir 1) que los índices geométricos que consideran aspectos de la geometría del receptor y/o están basados en mallas, no son siempre adecuados para estudios quimioinformáticos, 2) que la matriz geométrica, como punto de partida de varios índices 3D, es construida únicamente con la distancia Euclidiana, lo cual puede ser generalizable a la utilización de otras métricas comúnmente usadas en estudios de similitud/disimilitud y en aprendizaje automatizado, y 3) que el empleo de conceptos relacionados con el álgebra lineal y multilineal pudieran contribuir a obtener descriptores moleculares capaces de captar información distinta, con alta variabilidad y con mejor poder discriminatorio que los actuales.

Capítulo 2

Teoría de los índices moleculares QuBiLs MIDAS

En este capítulo es presentada la teoría de los índices moleculares QuBiLs MIDAS Duplas, Ternas y Cuaternas. Son expuestas las definiciones matemáticas de los índices totales y locales, las métricas y medidas empleadas en la codificación de la estructura química para las relaciones entre dos, tres y cuatro átomos, los enfoques matriciales, y los operadores de agregación utilizados como generalización de la combinación lineal de las contribuciones atómicas.

2.1. Vector molecular

El uso de vectores moleculares basados en átomos, como representación de estructuras químicas orgánicas de pequeño y mediano tamaño, ha sido explicado en detalle en varios reportes [5, 21, 22]. Los componentes de un vector molecular son valores numéricos, los cuales representan cierta propiedad atómica. Por lo tanto, una molécula teniendo 5, 10, 15, $\ldots, n$ núcleos atómicos, puede ser representada por medio de vectores con 5, 10, 15, $\ldots, n$ componentes, perteneciendo a los espacios $\mathbb{R}^5$, $\mathbb{R}^{10}$, $\mathbb{R}^{15}$,..., $\mathbb{R}^n$, respectivamente. Es decir, que un vector molecular, no es más que un vector de propiedad n-dimensional de los núcleos atómicos en una molécula.

En el presente trabajo, las propiedades químicas utilizadas como esquemas de ponderación son: 1) masa atómica (M), volumen de van der Waals (V), polarizabilidad (P), electronegatividad en la escala de Pauling (E), Ghose-Crippen LogP (A) [90, 91], carga atómica de Gasteiger-Marsili (C), superficie de área polar

(PSA), refractividad (R), dureza (H - *hardness*) y suavidad (S - *softness*).

2.2. Enfoques matriciales para el cálculo de los índices QuBiLs MIDAS

Con el objetivo de tomar en consideración la cercanía y/o distancia de las interacciones no covalentes dentro de una estructura molecular, fue generalizado el concepto de matriz de distancia geométrica [3, 92], utilizando para ello un conjunto de métricas diferentes a la Euclidiana para evaluar la distancia existente entre pares de átomos. Esta matriz generalizada es denominada como la k^{th} *matriz espacial bidimensional de similitud-disimilitud (B-SDSM)*, GB^k, donde k indica la potencia a la cual GB es elevada. En este sentido, para $k = 0$, la matriz GB^0 tiene cada entrada igual a 1; para $k = 1$, los elementos de la matriz GB^1 representan la distancia entre el átomo i y el átomo j (ver Tabla 2.1), y son definidos como sigue:

$$\begin{aligned} gb^1_{ij} &= D_{ij} \qquad i \neq j \ \wedge \ i, j \ son \ átomos \ de \ la \ molécula \\ &= L_{ij} \qquad i = j \\ &= 0 \qquad en \ cualquier \ otro \ caso \end{aligned} \tag{2.1}$$

donde, D_{ij} es la distancia entre el núcleo atómico i y j (ver Tabla 2.1), y L_{ij} puede ser: 1) el número de pares de electrones no pareados *(lone-pairs)* en el átomo i, o 2) la distancia del átomo i al centro de la molécula (D_{io}). De esta manera, los valores de la diagonal principal no son siempre ceros, logrando así una mayor discriminación de las estructuras moleculares.

Como puede apreciarse, la matriz GB^k, sólo permite representar información de las relaciones entre pares de átomos mediante métricas de distancias. Sin embargo, esta representación no resulta válida cuando se establecen relaciones n-dimensionales ($n > 2$) entre los átomos de una estructura molecular. Para considerar las interacciones no covalentes entre n núcleos atómicos, son construidas matrices tridimensionales y cuatridimensionales, cuando se tienen en cuenta relaciones entre tres y cuatro átomos respectivamente.

Para representar las relaciones ternarias y cuaternarias entre núcleos atómicos, las matrices que se proponen son la k^{th} *matriz espacial tridimensional de similitud-disimilitud (*GT^k*, T-SDSM)*, y la k^{th} *matriz espacial cuatridimensional de similitud-disimilitud (*GQ^k*, Q-SDSM)*, respectivamente. En ambos casos, y al igual que *B-SDSM*, k es la potencia a la cual GT y GQ son elevadas. Así, para $k = 0$, todos los coeficientes gt^0_{ijl} y gq^0_{ijlh} correspondientes a las matrices GT^0 y GQ^0 tienen valor 1; y para $k = 1$, los coeficientes gt^1_{ijl} de la matriz GT^1 y los coeficientes gq^1_{ijlh} de la matriz GQ^1, representan las relaciones existentes entre tres y cuatro átomos de una molécula respectivamente.

Métrica	Fórmula	Métrica	Fórmula
Minkowski	$d_{XY} = \left(\sum_{j=1}^{h} \lvert x_j - y_j \rvert^p \right)^{\frac{1}{p}}$	Soergel	$d_{XY} = \frac{1}{n} \sum_{j=1}^{h} \left(\frac{\lvert x_j - y_j \rvert}{max\{x_j, y_j\}} \right)$
Chebyshev	$d_{XY} = max\{\lvert x_j - y_j \rvert\}$	Bhattacharyya	$d_{XY} = \sqrt{\sum_{j=1}^{h} \left(\sqrt{x_j} - \sqrt{y_j} \right)^2}$
Canberra	$d_{xy} = \sum_{j=1}^{h} \frac{\lvert x_j - y_j \rvert}{\lvert x_j \rvert + \lvert y_j \rvert}$	Wave - Edges	$d_{XY} = \sum_{j=1}^{h} \left(1 - \frac{min\{x_j, y_j\}}{max\{x_j, y_j\}} \right)$
Lance - Williams	$d_{XY} = \frac{\sum_{j=1}^{h} \lvert x_j - y_j \rvert}{\sum_{j=1}^{h} (\lvert x_j \rvert + \lvert y_j \rvert)}$	Angular Separation	$d_{XY} = 1 - \frac{\sum_{j=1}^{h} (x_j * y_j)}{\sqrt{\sum_{j=1}^{h} (x_j)^2 * \sum_{j=1}^{h} (y_j)^2}}$
Clark	$d_{XY} = \sqrt{\sum_{j=1}^{h} \left(\frac{x_j - y_j}{\lvert x_j \rvert + \lvert y_j \rvert} \right)}$		

Tabla 2.1: Métricas utilizadas para el cálculo de distancias inter-atómicas

A continuación se muestran las definiciones formales de los elementos gt^1_{ijl} de la matriz GT^1 (Ecuación 2.2) y de los elementos gq^1_{ijlh} de la matriz GQ^1 (Ecuación 2.3):

$$\begin{aligned} gt^1_{ijl} &= TT_{ijl} && i \neq j \neq l \wedge i, j, l \text{ son átomos de la molécula} \\ &= L_{ijl} && i = j = l \\ &= 0 && \text{en cualquier otro caso} \end{aligned} \tag{2.2}$$

$$\begin{aligned} gq^1_{ijlh} &= QQ_{ijlh} && i \neq j \neq l \neq h \wedge i, j, l, w \text{ son átomos de la molécula} \\ &= L_{ijlh} && i = j = l = h \\ &= 0 && \text{en cualquier otro caso} \end{aligned} \tag{2.3}$$

donde, TT_{ijl} es la medida de asociación para ternas de átomos, QQ_{ijlh} es la medida de asociación para cuaternas de átomos, y L_{ijl} y L_{ijlh} tienen el mismo significado que en la Ecuación 2.1.

Para las medidas utilizadas en el cálculo de las relaciones ternarias o cuaternarias (TT_{ijl}, QQ_{ijlw}) entre los átomos de una molécula, no existen restricciones en los valores de i, j, l y h, siempre que $1 \leq i, j, l\, h \leq n$, donde n es la cantidad de núcleos atómicos de la estructura química. Sin embargo, cuando los átomos en el cálculo de una medida no son todos diferentes, entonces esta puede ser reducida[1]. De esta forma se tienen

[1] *p.ej.* si la métrica es entre tres parámetros y no todos diferentes, entonces puede utilizarse una que considere dos parámetros.

las siguientes opciones:

- *Relaciones ternarias:*

$$3\text{: } TT_{ijl} = \begin{cases} T_{ijl} & tres\,átomos\,diferentes \\ 0 & en\,cualquier\,otro\,caso \end{cases}$$

$$3T\ (total)\text{: } TT_{ijl} = \begin{cases} T_{ijl} & tres\,átomos\,diferentes \\ D_{ij} & dos\,átomos\,iguales\,y\,uno\,diferente \\ 0 & en\,cualquier\,otro\,caso \end{cases}$$

- *Relaciones cuaternarias:*

$$4\text{: } QQ_{ijlh} = \begin{cases} Q_{ijlh} & cuatro\,átomos\,diferentes \\ 0 & en\,cualquier\,otro\,caso \end{cases}$$

$$4T\ (total)\text{: } QQ_{ijlh} = \begin{cases} Q_{ijlh} & cuatro\,átomos\,diferentes \\ T_{ijl} & dos\,átomos\,iguales\,y\,dos\,diferentes \\ D_{ij} & tres\,átomos\,iguales\,y\,uno\,diferentes \\ 0 & en\,cualquier\,otro\,caso \end{cases}$$

donde, Q_{ijlh} es la medida utilizada para establecer una relación entre cuatro átomos (ver Tabla 2.2), T_{ijl} es la medida utilizada para establecer una relación entre tres átomos (ver Tabla 2.2), y D_{ij} es la distancia entre dos átomos (ver Tabla 2.1). En el caso de las medidas *ternarias y cuaternarias totales*, la reducción de las mismas es como se especifica en la Tabla 2.3.

Matrices de orden superior a uno ($k \geq 2$)

Hasta el momento, las únicas matrices definidas son GB, GT y GQ, para órdenes (potencia) de 0 y 1. Por lo tanto, las matrices GB^k, GT^k y GQ^k para $k \geq 2$, son calculadas multiplicando (producto Hadamard) los elementos de la matriz GB^{k-1}, GT^{k-1} y GQ^{k-1} por los elementos correspondientes en la matriz GB^1, GT^1 y GQ^1 respectivamente, de manera tal que los elementos de la matriz GB^k serán igual a $\left(gb_{ij}^1\right)^k$, los de la matriz GT^k serán igual a $\left(gt_{ijl}^1\right)^k$ y los de la matriz GQ^k serán igual a $\left(gq_{ijlh}^1\right)^k$, para todas las duplas, ternas y cuaternas de átomos en una molécula.

Cuando no son empleados procedimientos de normalización para los elementos de las matrices GB^k, GT^k y GQ^k, las mismas son conocidas como matrices no estocásticas, siendo formalmente denominadas como

Medidas ternarias	
Perímetro	$t_{XYZ} = d_{XY} + d_{YZ} + d_{ZX}$
Área	$t_{XYZ} = \sqrt{s\,(s - d_{XY})\,(s - d_{YZ})\,(s - d_{ZX})}$ $s = \dfrac{d_{XY} + d_{YZ} + d_{ZX}}{2}$
Suma de lados	$t_{XYZ} = d_{XY} + d_{YZ}$
Ángulo entre lados	A_X, A_Y, A_Z coordenadas de tres átomos de una molécula $U = A_X - A_Y$, $V = A_Z - A_Y$ $t_{XYZ} = \alpha = \arccos\left(\dfrac{U * V}{\lvert U\rvert * \lvert V\rvert}\right)$
Medidas cuaternarias	
Perímetro	$q_{XYZW} = d_{XY} + d_{YZ} + d_{ZW} + d_{WX}$
Volumen	A_X, A_Y, A_Z, A_W coordenadas de cuatro átomos de una molécula $q_{XYZW} = \dfrac{1}{6}\begin{pmatrix} A_{Y1} - A_{X1} & A_{Z1} - A_{X1} & A_{W1} - A_{X1} \\ A_{Y2} - A_{X2} & A_{Z2} - A_{X2} & A_{W2} - A_{X2} \\ A_{Y3} - A_{X3} & A_{Z3} - A_{X3} & A_{W3} - A_{X3} \end{pmatrix}$
Suma de lados	$q_{XYZW} = d_{XY} + d_{YZ} + d_{ZW}$
Ángulo diedro	A_X, A_Y, A_Z coordenadas de tres átomos de una molécula en el plano A B_W, B_Y, B_Z coordenadas de tres átomos de una molécula en el plano B $U_A = (A_X - A_Y) \times (A_Z - A_Y)$ $U_B = (B_W - B_Y) \times (B_Z - B_Y)$ $q_{XYZW} = \alpha = \arccos\left(\dfrac{U_A * U_B}{\lvert U_A\rvert * \lvert U_B\rvert}\right)$

d_{AB}: distancia entre el átomo A y el átomo B calculada con alguna de las métricas especificadas en la Tabla 2.1

Tabla 2.2: Medidas utilizadas para el cálculo de las relaciones de ternas y cuaternas de átomos

Medida cuaternaria (Q_{ijlh})		**Medida ternaria** (T_{ijl})		**Métrica de distancia** (D_{ij})
Perímetro	$\Longrightarrow$	Perímetro	$\Longrightarrow$	Distancia entre dos átomos
Volumen	$\Longrightarrow$	Área	$\Longrightarrow$	Distancia entre dos átomos
Suma de lados	$\Longrightarrow$	Suma de lados	$\Longrightarrow$	Distancia entre dos átomos
Ángulo diedro	$\Longrightarrow$	Ángulo entre lados	$\Longrightarrow$	Ángulo entre dos átomos

Tabla 2.3: Reducción de las medidas para el cálculo de las relaciones n-dimensionales entre los núcleos atómicos de una molécula

la k^{th} *matriz espacial bidimensional no estocástica de similitud-disimilitud (*${}_{ns}GB^k$*, NS-B-SDSM)*, la k^{th} *matriz espacial tridimensional no estocástica de similitud-disimilitud (*${}_{ns}GT^k$*, NS-T-SDSM)* y la k^{th} *matriz espacial cuatridimensional no estocástica de similitud-disimilitud (*${}_{ns}GQ^k$*, NS-Q-SDSM)*, respectivamente.

Estos tipos de matrices (NS-B-SDSM, NS-T-SDSM, NS-Q-SDSM) serán clasificadas como **matrices**

generalizadas [3], comúnmente denotadas por M^{λ}. Estas matrices generalizadas son obtenidas mediante el producto Hadamard, también conocido como el producto de Schur [93]. Es importante señalar que en todas las aplicaciones algebraicas utilizadas en este trabajo para el cálculo de los índices QuBiLs MIDAS (ver Sección 2.3), la matriz de las formas algebraicas, pertenecen a la clase de matrices generalizadas ($k = \lambda$), donde λ puede ser tanto positiva como negativa.

Cuando el exponente λ es negativo este tipo de matrices se denominan *matrices recíprocas*. Por ejemplo, para $\lambda = -1$, se conoce la matriz de Harary, la matriz geométrica recíproca, entre otras. Cuando $\lambda = -2$, se conoce la *matriz recíproca de la distancia al cuadrado*, la cual es derivada de la matriz geométrica. En esta clase de matrices a los elementos de la diagonal no se les determina el recíproco por tener los valores iguales a cero. En el presente trabajo, cuando el exponente de la matriz es negativo [$(k = \lambda) < 0$] y se consideran los *lone-pairs* para cada átomo i, el recíproco no es aplicado a los elementos de la diagonal. Sin embargo, cuando el exponente de la matriz es negativo [$(k = \lambda) < 0$] y la distancia entre cada átomo i y el centro de la molécula o es seleccionado (ver Ecuación 2.1), entonces el recíproco es aplicado a los elementos de la diagonal. El máximo valor de k es ± 12, estando este valor relacionado con las interacciones no covalentes asociadas con la forma funcional del potencial de Lennard-Jones.

En el campo de los índices geométricos, para obtener algún descriptor molecular, lo típico es usar solamente la distancia Euclidiana con potencia 1 o 2; los que se denominan como índices gravitacionales 3D. Sin embargo, en matrices grafo-teóricas, algunos descriptores como el *momento de distribución de la distancia* (denotado como D_{λ}), han sido definidos utilizando la potencia de la distancia topológica d_{ij} en un grafo molecular, a partir de matrices de distancias generalizadas [94].

Formalismos de normalización basado en los esquemas simple estocástico, doble estocástico y de probabilidad mutua

Con el propósito de normalizar la k^{th} *matriz espacial bidimensional [tridimensional, cuatridimensional] no estocástica de similitud-disimilitud*, ${}_{ns}GB^{k}$ [${}_{ns}GT^{k}$, ${}_{ns}GQ^{k}$], son aplicados tres esquemas de probabilidad. Estos esquemas son asociados con las interacciones inter-atómicas en la estructura química.

Sobre la base de la *matriz estocástica de densidad electrónica grafo-teórica* de los índices QuBiLs MAS 2D y 2.5D [20,23], es definida la k^{th} *matriz espacial bidimensional simple estocástica de similitud-disimilitud* [${}_{ss}GB^{k}$ *(SS-B-SDSM)*]. Esta matriz, en el enfoque geométrico, puede ser interpretada como el cambio en la probabilidad de los átomos en una molécula a interactuar con cada otro. Consecuentemente, puede considerarse esta probabilidad como una medida del esparcimiento de los átomos en el espacio. La matriz

${}_{ss}GB^k$ es obtenida a partir de la matriz ${}_{ns}GB^k$ como sigue:

$$ {}_{ss}gb_{ij}^k = \frac{{}_{ns}gb_{ij}^k}{S_j} = \frac{{}_{ns}gb_{ij}^k}{\sum_j {}_{ns}gb_{ij}^k} \tag{2.4} $$

donde, ${}_{ns}gb_{ij}^k$ son los elementos de la k^{th} potencia de la matriz ${}_{ns}GB$, y S_j es la sumatoria de la fila i de la matriz ${}_{ns}GB^k$, siendo esta suma denominada como el *grado espacial del vértice de similitud-disimilitud de orden k* para el átomo i [66].

De forma análoga son definidas la k^{th} *matriz espacial tridimensional simple estocástica de similitud-disimilitud* [${}_{ss}GT^k$ *(SS-T-SDSM)*], *y la* k^{th} *matriz espacial cuatridimensional simple estocástica de similitud-disimilitud* [${}_{ss}GQ^k$ *(SS-Q-SDSM)*], mostrándose sus definiciones en la Ecuaciones 2.5 y 2.6 respectivamente:

$$ {}_{ss}gt_{ijl}^k = \frac{{}_{ns}gt_{ijl}^k}{S_{jl}} = \frac{{}_{ns}gt_{ijl}^k}{\sum_{j=1}^{n}\sum_{l=1}^{n} {}_{ns}gt_{ijl}^k} \tag{2.5} $$

$$ {}_{ss}gq_{ijlh}^k = \frac{{}_{ns}gq_{ijlh}^k}{S_{jlh}} = \frac{{}_{ns}gq_{ijlh}^k}{\sum_{j=1}^{n}\sum_{l=1}^{n}\sum_{h=1}^{n} {}_{ns}gq_{ijlh}^k} \tag{2.6} $$

donde, ${}_{ns}gt_{ijl}^k$ son los elementos de la k^{th} potencia de la matriz ${}_{ss}GT$, ${}_{ns}gq_{ijlh}^k$ son los elementos de la k^{th} potencia de la matriz ${}_{ss}GQ$, S_{jl} es la suma de todos los elementos de la matriz bidimensional correspondiente a cada átomo i en una matriz de dimensión tres, y S_{jlh} es la sumatoria de las entradas de la matriz tridimensional perteneciente a cada átomo i en una matriz de dimensión cuatro.

Estas transformaciones simples estocásticas generan matrices no simétricas, y por lo tanto otro enfoque a considerar sería obtener matrices donde la suma de los elementos de todas las componentes sumen 1. Sin embargo, para matrices n-dimensionales ($n > 2$), no existen reportados algoritmos que posibiliten este escalamiento. Por otro lado, para matrices bidimensionales, con el objetivo de equilibrar las probabilidades en ambos sentidos, una matriz doble estocástica es empleada, definida como una matriz con entradas reales no negativas cuyas filas y columnas sumen 1 [3]. Este tipo de matriz será referida como la k^{th} *matriz espacial bidimensional doble estocástica de similitud-disimilitud* [${}_{ds}GB^k$ *(DS-B-SDSM)*]. El procedimiento para calcular la matriz doble estocástica asociada a una no estocástica no es trivial. Sinkhorn postuló que una matriz estrictamente positiva A puede ser escalada a una matriz doble estocástica B por [95]:

$$ B = D_g \times A \times D_g \tag{2.7} $$

donde, D_g es una matriz diagonal. Luego, Sinkhorn y Knopp extendieron este teorema para considerar matrices no negativas y propusieron un algoritmo de iteración para el balanceo de matrices [96]. En este sentido, la matriz ${}_{ds}GB^k$ puede ser calculada a partir de la matriz ${}_{ns}GB^k$ usando la ecuación 2.7 y el algoritmo Sinkhorn-Knopp.

Finalmente se introducen la k^{th} *matriz espacial bidimensional de probabilidad mutua de similitud-disimilitud* $[{}_{mp}GB^k$ *(MP-B-SDSM)*$]$, la k^{th} *matriz espacial tridimensional de probabilidad mutua de similitud-disimilitud* $[{}_{mp}GT^k$ *(MP-T-SDSM)*$]$, y la k^{th} *matriz espacial cuatridimensional de probabilidad mutua de similitud-disimilitud* $[{}_{mp}GQ^k$ *(MP-Q-SDSM)*$]$; constituyendo sus elementos los coeficientes ${}_{mp}gb^k_{ij}$, ${}_{mp}gt^k_{ijl}$ y ${}_{mp}gq^k_{ijlh}$ respectivamente, los cuales se definen por las ecuaciones siguientes:

$$ {}_{mp}gb^k_{ij} = \frac{{}_{ns}gb^k_{ij}}{S} = \frac{{}_{ns}gb^k_{ij}}{\sum_{i=1}^{n}\sum_{j=1}^{n} {}_{ns}gb^k_{ij}} \tag{2.8} $$

$$ {}_{mp}gt^k_{ijl} = \frac{{}_{ns}gt^k_{ijl}}{S} = \frac{{}_{ns}gt^k_{ijl}}{\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n} {}_{ns}gt^k_{ijl}} \tag{2.9} $$

$$ {}_{mp}gq^k_{ijlh} = \frac{{}_{ns}gq^k_{ijlh}}{S} = \frac{{}_{ns}gq^k_{ijlh}}{\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\sum_{h=1}^{n} {}_{ns}gq^k_{ijlh}} \tag{2.10} $$

donde, ${}_{mp}gb^k_{ij}$ denota la probabilidad mutua entre los núcleos atómicos i y j, ${}_{mp}gt^k_{ijl}$ representa la probabilidad mutua entre los átomos i, j y l, y el coeficiente ${}_{mp}gq^k_{ijlh}$ denota la probabilidad mutua entre los átomos i, j, l y h. Por su parte, S significa el espacio muestral, el cual es calculado mediante la sumatoria de todos los elementos de ${}_{ns}GB^k$, ${}_{ns}GT^k$ o ${}_{ns}GQ^k$ según sea el caso. Es importante resaltar que mientras el enfoque simple estocástico ha sido previamente usado en otros trabajos [20, 23], los esquemas doble estocástico y de probabilidad mutua son presentados por primera vez como una alternativa a estrategias de normalización.

2.3. Índices moleculares QuBiLs MIDAS

Si una molécula consiste de n átomos, entonces, para las relaciones entre par de ellos son determinados los k^{th} índices lineales, bilineales y cuadráticos para el átomo "a", los cuales son calculados como aplicaciones

(formas) lineales, bilineales y cuadráticas en $\mathbb{R}^n$, sobre un conjunto base canónico, y son expresados por las ecuaciones 2.11, 2.12 y 2.13, respectivamente. Bajo este mismo principio de calcular los índices para un núcleo atómico específico, son definidos para las relaciones n-dimensionales los índices ternarios ($n = 3$) y cuaternarios ($n = 4$) acorde al átomo "a", como es mostrado en las ecuaciones 2.14 y 2.15 respectivamente.

$$ {}_{ns[ss,ds,mp]f}L_a = {}_{ns[ss,ds,mp]}f^{a,k}(\bar{x}) = \sum_{i=1}^{n}\sum_{j=1}^{n} {}_{ns[ss,ds,mp]}gb_{ij}^{a,k}u^i x^j = [U]^T {}_{ns[ss,ds,mp]}G^{a,k}[X] \tag{2.11} $$

$$ {}_{ns[ss,ds,mp]b}L_a = {}_{ns[ss,ds,mp]}b^{a,k}(\bar{x},\bar{y}) = \sum_{i=1}^{n}\sum_{j=1}^{n} {}_{ns[ss,ds,mp]}gb_{ij}^{a,k}x^i y^j = [X]^T {}_{ns[ss,ds,mp]}G^{a,k}[Y] \tag{2.12} $$

$$ {}_{ns[ss,ds,mp]q}L_a = {}_{ns[ss,ds,mp]}q^{a,k}(\bar{x},\bar{x}) = \sum_{i=1}^{n}\sum_{j=1}^{n} {}_{ns[ss,ds,mp]}gb_{ij}^{a,k}x^i x^j = [X]^T {}_{ns[ss,ds,mp]}G^{a,k}[X] \tag{2.13} $$

$$ {}_{ns[ss,mp]tr}L_a = {}_{ns[ss,mp]}tr^{a,k}(\bar{x},\bar{y},\bar{z}) = \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n} {}_{ns[ss,mp]}gt_{ijl}^{a,k}x^i y^j z^l = {}_{ns[ss,mp]}\, GT^{a,k}\cdot\bar{x}\cdot\bar{y}\cdot\bar{z} \tag{2.14} $$

$$ {}_{ns[ss,mp]qu}L_a = {}_{ns[ss,mp]}qu^{a,k}(\bar{x},\bar{y},\bar{z},\bar{w}) = \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\sum_{h=1}^{n} {}_{nsp[ss,mp]}gq_{ijlh}^{a,k}x^i y^j z^l w^h = {}_{ns[ss,mp]}\, GQ^{a,k}\cdot\bar{x}\cdot\bar{y}\cdot\bar{z}\cdot\bar{w} \tag{2.15} $$

En las ecuaciones anteriores, n es la cantidad de núcleos atómicos de la molécula, $\bar{u}$ es el vector unidad, y $x^1,\ldots,x^n$, $y^1,\ldots y^n$, $z^1,\ldots z^n$ y $w^1,\ldots w^n$ son las coordenadas o componentes de los vectores moleculares $\bar{x}$, $\bar{y}$, $\bar{z}$ y $\bar{w}$ en un sistema de vectores base canónico de $\mathbb{R}^n$.

Además, los coeficientes ${}_{ns[ss,ds,mp]}gb_{ij}^{a,k}$ $\left[{}_{ns[ss,mp]}gt_{ijl}^{a,k},\ {}_{ns[ss,mp]}gq_{ijlh}^{a,k}\right]$ son los elementos de la k^{th} *matriz espacial bidimensional [tridimensional, cuatridimensional] no estocástica de similitud-disimilitud de nivel atómico* $\left({}_{ns}GB^{a,k}\ \left[{}_{ns}GT^{a,k},\ {}_{ns}GQ^{a,k}\right]\right)$ para el átomo "a", o de sus correspondientes transformaciones algebraicas *simple estocástica* $\left({}_{ss}GB^{a,k}\ \left[{}_{ss}GT^{a,k},\ {}_{ss}GQ^{a,k}\right]\right)$, *doble estocástica* $\left({}_{ds}GB^{a,k}\right)$ y de *probabilidad mutua* $\left({}_{mp}GB^{a,k}\ \left[{}_{mp}GT^{a,k},\ {}_{mp}GQ^{a,k}\right]\right)$. Los elementos de estas matrices de nivel atómico son obtenidos como se especifica en la Ecuación 2.16 para las relaciones entre pares de átomos, en la Ecuación 2.17 para las relaciones entre tres átomos, y en la Ecuación 2.18 para las relaciones entre cuatro átomos, a partir de la matriz total correspondiente.

$$
\begin{aligned}
{}_{ns[ss,ds,mp]}gb_{ij}^{a,k} &= {}_{ns[ss,ds,mp]}gb_{ij}^{k} && \textit{si los dos átomos son iguales al átomo ``a''} \\
&= \frac{1}{2}\,{}_{ns[ss,ds,mp]}gb_{ij}^{k} && \textit{si un átomo es igual al átomo ``a''} \\
&= 0 && \textit{en cualquier otro caso}
\end{aligned} \tag{2.16}
$$

$$
\begin{aligned}
{}_{ns[ss,mp]}gt_{ijl}^{a,k} &= {}_{ns[ss,mp]}gt_{ijl}^{k} && \textit{si los tres átomos son iguales al átomo ``a''} \\
&= \frac{2}{3}\,{}_{ns[ss,mp]}gt_{ijl}^{k} && \textit{si dos átomos son iguales al átomo ``a''} \\
&= \frac{1}{3}\,{}_{ns[ss,mp]}gt_{ijl}^{k} && \textit{si un átomo es igual al átomo ``a''} \\
&= 0 && \textit{en cualquier otro caso}
\end{aligned} \tag{2.17}
$$

$$
\begin{aligned}
{}_{ns[ss,mp]}gq_{ijlh}^{a,k} &= {}_{ns[ss,mp]}gq_{ijlh}^{k} && \textit{si los cuatro átomos son iguales al átomo ``a''} \\
&= \frac{3}{4}\,{}_{ns[ss,mp]}gq_{ijlh}^{k} && \textit{si tres átomos son iguales al átomo ``a''} \\
&= \frac{2}{4}\,{}_{ns[ss,mp]}gq_{ijlh}^{k} && \textit{si dos átomos son igual al átomo ``a''} \\
&= \frac{1}{4}\,{}_{ns[ss,mp]}gq_{ijlh}^{k} && \textit{si un átomo es igual al átomo ``a''} \\
&= 0 && \textit{en cualquier otro caso}
\end{aligned} \tag{2.18}
$$

Por lo tanto, si una molécula es particionada en "A" átomos, la matrices ${}_{ns[ss,ds,mp]}GB^{k}$, ${}_{ns[ss,mp]}GT^{k}$ y ${}_{ns[ss,mp]}GQ^{k}$, pueden ser particionadas en "A" matrices de nivel atómico ${}_{ns[ss,ds,mp]}GB^{a,k}$, ${}_{ns[ss,mp]}GT^{a,k}$ y ${}_{ns[ss,mp]}GQ^{a,k}$ respectivamente; siendo exactamente la k^{th} potencia de la matriz ${}_{ns[ss,ds,mp]}GB^{k}$, ${}_{ns[ss,mp]}GT^{k}$ y ${}_{ns[ss,mp]}GQ^{k}$, la suma de las k^{th} potencias de las correspondientes matrices de nivel atómico ${}_{ns[ss,ds,mp]}GB^{a,k}$, ${}_{ns[ss,mp]}GT^{a,k}$ y ${}_{ns[ss,mp]}GQ^{a,k}$. Esto significa que cada matriz ${}_{ns[ss,ds,mp]}GB^{a,k}$, ${}_{ns[ss,mp]}GT^{a,k}$ y ${}_{ns[ss,mp]}GQ^{a,k}$, determina un índice de nivel atómico para el átomo "a", denominado en este trabajo con el acrónimo de LOVI[2] [97,98], y denotado como L_a (ver ecuaciones 2.11, 2.12, 2.13, 2.14 y 2.15). De esta forma, los índices totales (molécula completa) lineal, bilineal, cuadrático, trilineal y cuatrilineal, pueden ser representados como un vector $\bar{L}$ de tamaño n, donde cada entrada L_a corresponde al k^{th} índice de nivel atómico lineal, bilineal, cuadrático, trilineal o cuatrilineal para el átomo "a".

Entonces, a partir de la definición anterior, los índices totales (molécula completa) QuBiLs MIDAS son calculados como una combinación lineal de los índices de nivel atómico (componentes del vector $\bar{L}$). Generalizaciones de este enfoque es discutido en la Sección 2.4 usando varios operadores de agregación. En

[2] **LO**cal **V**ertex **I**nvariant

el caso de los índices lineales, bilineales y cuadráticos, la sumatoria sobre $\bar{L}$ es equivalente al producto entre el vector de propiedad (vector unidad) $[X]^T$ (o $[U]^T$), la matriz ${}_{ns[ss,ds,mp]}GB^k$ y el vector de propiedad $[Y]$, coincidiendo así con el enfoque original de las formas algebraicas lineal, bilineal y cuadrática. Por otro lado, para los índices n-dimensionales, la sumatoria de los componentes del vector $\bar{L}$, también coinciden con la multiplicación de un tensor de orden tres y un tensor de orden cuatro, con sus correspondientes vectores de propiedades. Esto es mostrado en las ecuaciones siguientes (ver también ecuaciones 2.11, 2.12, 2.13, 2.14 y 2.15):

$$ {}_{ns[ss,ds,mp]}f^k(\bar{x}) = \sum_{a=1}^{n} {}_{ns[ss,ds,mp]f}L_a = [U]^T \, {}_{ns[ss,ds,mp]}G^k \, [X] \tag{2.19}$$

$$ {}_{ns[ss,ds,mp]}b^k(\bar{x},\bar{y}) = \sum_{a=1}^{n} {}_{ns[ss,ds,mp]b}L_a = [X]^T \, {}_{ns[ss,ds,mp]}G^k \, [Y] \tag{2.20}$$

$$ {}_{ns[ss,ds,mp]}q^k(\bar{x}) = \sum_{a=1}^{n} {}_{ns[ss,ds,mp]q}L_a = [X]^T \, {}_{ns[ss,ds,mp]}G^k \, [X] \tag{2.21}$$

$$ {}_{ns[ss,mp]}tr^k(\bar{x},\bar{y},\bar{z}) = \sum_{a=1}^{n} {}_{ns[ss,mp]}trL_a = {}_{ns[ss,mp]} \, GT^k \cdot \bar{x} \cdot \bar{y} \cdot \bar{z} \tag{2.22}$$

$$ {}_{ns[ss,mp]}qu^k(\bar{x},\bar{y},\bar{z},\bar{w}) = \sum_{a=1}^{n} {}_{ns[ss,mp]}quL_a = {}_{ns[ss,mp]} \, GQ^k \cdot \bar{x} \cdot \bar{y} \cdot \bar{z} \cdot \bar{w} \tag{2.23}$$

donde, $\bar{x}$, $\bar{y}$, $\bar{z}$ y $\bar{w}$ son vectores de propiedades, $[X]$ y $[Y]$ son los vectores columnas (una matriz de $nx1$) de las coordenadas de $\bar{x}$ y $\bar{y}$ en la base canónica de $\mathbb{R}^n$, $[U]^T$ y $[X]^T$ son las transpuestas del vector unitario $[U]$ y el vector de propiedad $[X]$ respectivamente. Finalmente, ${}_{ns[ss,ds,mp]}GB^k\left[{}_{ns[ss,mp]}GT^k, {}_{ns[ss,mp]}GQ^k\right]$ es la k^{th} *matriz espacial bidimensional [tridimensional, cuatridimensional] de similitud-disimilitud*, en su definición no estocástica, simple estocástica, doble estocástica o de probabilidad mutua, según corresponda en cada caso.

Como puede notarse en las ecuaciones 2.14 y 2.22 para los índices ternarios, y en las ecuaciones 2.15 y 2.23 para los índices cuaternarios, son necesitados tres y cuatro vectores de propiedades respectivamente, los cuales pueden estar ponderados con distintas propiedades químicas o pueden ser el vector unidad ($\bar{u}$). De esta manera fueron definidas las formas *"trilineal*es" que se muestran a continuación, según la ponderación de los vectores $\bar{x}$, $\bar{y}$ y $\bar{z}$ de los índices QuBiLs MIDAS Ternas:

Trilineal (Tr): $\bar{x} \neq \bar{y} \neq \bar{z}$

Trilineal-Cuadrática-Bilineal (TrQB): $(\bar{x} = \bar{y}) \neq \bar{z}$

Trilineal-Bilineal (TrB): $\bar{x} \neq \bar{y}$, $\bar{z} = \bar{u}$

Trilineal-Cúbica (TrC): $\bar{x} = \bar{y} = \bar{z}$

Trilineal-Lineal (TrF): $\bar{x} \neq (\bar{y} = \bar{z} = \bar{u})$

De manera análoga ocurre con las ponderaciones de los vectores de propiedad $\bar{x}$, $\bar{y}$, $\bar{z}$ y $\bar{w}$ de los índices QuBiLs MIDAS Cuaternas, dando origen a las siguientes formas *"cuatrilineales"*:

Cuatrilineal (Qu): $\bar{x} \neq \bar{y} \neq \bar{z} \neq \bar{w}$

Cuatrilineal-Cuadrática-Trilineal (QuQTr): $((\bar{x} = \bar{y}) \neq \bar{z} \neq \bar{w})$

Cuatrilineal-Trilineal (QuTr): $(\bar{x} = \bar{u}) \neq \bar{y} \neq \bar{z} \neq \bar{w}$

Cuatrilineal-Cúbica-Bilineal (QuCB): $(\bar{x} = \bar{y} = \bar{z}) \neq \bar{w}$

Cuatrilineal-Bilineal (QuB): $((\bar{x} = \bar{y} = \bar{u}) \neq \bar{z} \neq \bar{w})$

Cuatrilineal-Cuádruple (QuQd): $\bar{x} = \bar{y} = \bar{z} = \bar{w}$

Cuatrilineal-Lineal (QuF): $(\bar{x} = \bar{y} = \bar{z} = \bar{u}) \neq \bar{w}$

Definición de índices locales QuBiLs MIDAS

Las representaciones matriciales propuestas para las relaciones entre dos, tres y cuatro núcleos atómicos $\left({}_{ns[ss,ds,mp]}GB^k,\ {}_{ns[ss,mp]}GT^k,\ {}_{ns[ss,mp]}GQ^k\right)$, pueden ser usadas para codificar información sobre un fragmento molecular F de una molécula. Por lo tanto, las *matrices de similitud-disimilitud* para un fragmento molecular F, ${}_{ns[ss,ds,mp]}GB^k_F$, ${}_{ns[ss,mp]}GT^k_F$ y ${}_{ns[ss,mp]}GQ^k_F$, son obtenidas a partir de la correspondiente matriz total ${}_{ns[ss,ds,mp]}GB^k$, ${}_{ns[ss,mp]}GT^k$ y ${}_{ns[ss,mp]}GQ^k$, respectivamente.

Los elementos ${}_{ns[ss,ds,mp]}gb^k_{ijF}$, ${}_{ns[ss,mp]}gt^k_{ijlF}$, ${}_{ns[ss,mp]}gq^k_{ijlhF}$ de las matrices por tipo de átomos correspondientes, están definidos como sigue:

$$
\begin{aligned}
{}_{ns[ss,ds,mp]}gb^k_{ijF} &= {}_{ns[ss,ds,mp]}gb^k_{ij} && \textit{si los dos átomos pertenecen a "F"} \\
&= \frac{1}{2}\,{}_{ns[ss,ds,mp]}gb^k_{ij} && \textit{si un átomo pertenece a "F"} \\
&= 0 && \textit{en cualquier otro caso}
\end{aligned}
\tag{2.24}
$$

$$
\begin{aligned}
{}_{ns[ss,mp]}gt^{k}_{ijlF} &= {}_{ns[ss,mp]}gt^{k}_{ijl} && \textit{si los tres átomos pertenecen a "F"} \\
&= \frac{2}{3}\,{}_{ns[ss,mp]}gt^{k}_{ijl} && \textit{si dos átomos pertenecen a "F"} \\
&= \frac{1}{3}\,{}_{ns[ss,mp]}gt^{k}_{ijl} && \textit{si un átomo pertenece a "F"} \\
&= 0 && \textit{en cualquier otro caso}
\end{aligned} \tag{2.25}
$$

$$
\begin{aligned}
{}_{ns[ss,mp]}gq^{k}_{ijlhF} &= {}_{ns[ss,mp]}gq^{k}_{ijlh} && \textit{si los cuatro átomos pertenecen a "F"} \\
&= \frac{3}{4}\,{}_{ns[ss,mp]}gq^{k}_{ijlh} && \textit{si tres átomos pertenecen a "F"} \\
&= \frac{2}{4}\,{}_{ns[ss,mp]}gq^{k}_{ijlh} && \textit{si dos átomos pertenecen a "F"} \\
&= \frac{1}{4}\,{}_{ns[ss,mp]}gq^{k}_{ijlh} && \textit{si un átomo pertenece a "F"} \\
&= 0 && \textit{en cualquier otro caso}
\end{aligned} \tag{2.26}
$$

donde, los coeficientes ${}_{ns[ss,ds,mp]}gb^{k}_{ij}$, ${}_{ns[ss,mp]}gt^{k}_{ijl}$ y ${}_{ns[ss,mp]}gq^{k}_{ijlh}$, son los valores de las *matrices totales de similitud-disimilitud* para las relaciones de duplas, ternas y cuaternas de átomos, respectivamente.

Similar a los índices totales de nivel atómico, los índices locales de nivel atómico son calculados como un vector de LOVIs $\bar{L}$, donde cada entrada L_a corresponde a un valor de un índice local acorde al átomo "a" considerado. La definición de estos descriptores es como sigue:

$$
{}_{ns[ss,ds,mp]}f_F L_a = {}_{ns[ss,ds,mp]}\, f_F^{a,k}(\bar{x}) = \sum_{i=1}^{n}\sum_{j=1}^{n} {}_{ns[ss,ds,mp]}gb^{a,k}_{ijF} u^i x^j = [U]^T\, {}_{ns[ss,ds,mp]}G_F^{a,k}\,[X] \tag{2.27}
$$

$$
{}_{ns[ss,ds,mp]}b_F L_a = {}_{ns[ss,ds,mp]}\, b_F^{a,k}(\bar{x},\bar{y}) = \sum_{i=1}^{n}\sum_{j=1}^{n} {}_{ns[ss,ds,mp]}gb^{a,k}_{ijF} x^i y^j = [X]^T\, {}_{ns[ss,ds,mp]}G_F^{a,k}\,[Y] \tag{2.28}
$$

$$
{}_{ns[ss,ds,mp]}q_F L_a = {}_{ns[ss,ds,mp]}\, q_F^{a,k}(\bar{x},\bar{x}) = \sum_{i=1}^{n}\sum_{j=1}^{n} {}_{ns[ss,ds,mp]}gb^{a,k}_{ijF} x^i x^j = [X]^T\, {}_{ns[ss,ds,mp]}G_F^{a,k}\,[X] \tag{2.29}
$$

$$
{}_{ns[ss,mp]}tr_F L_a = {}_{ns[ss,mp]}tr_F^{a,k}(\bar{x},\bar{y},\bar{z}) = \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n} {}_{ns[ss,mp]}gt^{a,k}_{ijlF} x^i y^j z^l = {}_{ns[ss,mp]}GT_F^{a,k}\cdot\bar{x}\cdot\bar{y}\cdot\bar{z} \tag{2.30}
$$

$$
{}_{ns[ss,mp]}qu_F L_a = {}_{ns[ss,mp]}qu_F^{a,k}(\bar{x},\bar{y},\bar{z},\bar{w}) = \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\sum_{h=1}^{n} {}_{nsp[ss,mp]}gq^{a,k}_{ijlhF} x^i y^j z^l w^h = {}_{ns[ss,mp]}GQ_F^{a,k}\cdot\bar{x}\cdot\bar{y}\cdot\bar{z}\cdot\bar{w} \tag{2.31}
$$

donde, ${}_{ns[ss,ds,mp]}gb^{a,k}_{ijF}$, ${}_{ns[ss,mp]}gt^{a,k}_{ijlF}$ y ${}_{ns[ss,mp]}gq^{a,k}_{ijlhF}$ son los k^{th} elementos de la matriz local de nivel atómico ${}_{ns[ss,ds,mp]}GB^{a,k}_{F}$, ${}_{ns[ss,mp]}GT^{a,k}_{F}$ y ${}_{ns[ss,mp]}GQ^{a,k}_{F}$, respectivamente, acorde al átomo "a" considerado. Estas matrices son extraídas para cada núcleo atómico de la molécula, a partir de las correspondientes matrices de fragmento (local) ${}_{ns[ss,ds,mp]}GB^{k}_{F}$, ${}_{ns[ss,mp]}GT^{k}_{F}$ y ${}_{ns[ss,mp]}GQ^{k}_{F}$, las cuales contienen información referente a las relaciones entre dos, tres y cuatro átomos respectivamente. Con la sumatoria sobre el vector de LOVIs $\bar{L}$ se obtienen los k^{th} índices QuBiLs MIDAS para los tipos o grupos de átomos considerados.

En este trabajo, los descriptores moleculares locales QuBiLs MIDAS pueden ser determinados por siete grupos químicos en la molécula, los cuales son: aceptores de enlaces de hidrógeno (A), átomos de carbono en cadenas alifáticas (C), donadores de enlaces de hidrógeno (D), halógenos (G), grupos metilos terminales (M), átomos de carbono en porciones aromáticas (P) y heteroátomos (O, N, S en todos los estados de valencia, denota como X).

2.4. Operadores de agregación de las contribuciones atómicas

La noción de operadores de agregación como un esquema de generalización a la combinación lineal de las contribuciones atómicas para obtener definiciones totales (locales) de los índices QuBils MIDAS, es originada de la hipótesis que la más adecuada definición total de un sistema natural, puede no ser necesariamente aditiva. De hecho esto fue demostrado en [76, 99], donde otros operadores distintos a la suma obtuvieron mejor correlación con determinadas propiedades químicas. Estos operadores de agregación fueron clasificados en cuatro grandes grupos (ver Anexo A) como se muestra a continuación, y los mismos son aplicados al vector $\bar{L}$ de índices por nivel atómico.

1. *Normas (o Métricas):* normas de Minkowski (N1, N2, N3), y Penrose (PN). Puede notarse que el caso de N1, es el equivalente a la sumatoria de los componentes del vector $\bar{L}$.

2. *Estadísticos de tendencia central:* media geométrica (GM), media aritmética (AM), media cuadrática (P2), media potencial (P3) y media armónica (HM).

3. *Estadísticos de dispersión y forma:* varianza (V), skeness (S), kurtosis (K), desviación estándar (SD), coeficiente de variación (VC), rango (RA), percentil 25 (Q1), percentil 50 (Q2), percentil 75 (Q3), XMax (MX) y XMin (MN).

4. *Algoritmos clásicos:* autocorrelación (AC), gravitacional (GV), contenido de información total (TIC), contenido de información media (MIC), contenido de información estandarizada (SIC), suma total

(TS), Ivanciuc-Balaban (IB), estado electrotopológico (ES) y conectividad de Kier-Hall (KH).

El uso de estos operadores matemáticos sobre el vector de LOVIs, posibilita obtener una serie de índices que global o parcialmente caracterizan una molécula; con la particularidad del operador N1, que representa los índices QuBiLs MIDAS lineal, bilineal, cuadrático, trilineal y cuatrilineal, definidos en las ecuaciones 2.19, 2.20, 2.21, 2.22 y 2.23, respectivamente. De la misma forma, estos operadores pueden ser aplicados a un vector $\bar{L}$ compuesto de una clase particular de fragmento (o tipo de átomo), para obtener diversos índices QuBiLs MIDAS locales que describan una determinada molécula. Puede observarse que para los *"Algoritmos clásicos"*, además del uso de índices de nivel atómico como LOVIs, en lugar del grado del vértice, ellos usualmente realizan la operación de suma, la cual es también generalizada con los operadores especificados en los grupos de *Normas, Estadísticos de tendencia central y Estadísticos de dispersión y forma.*

2.5. Conclusiones parciales

En este capítulo fue presentada la teoría de los índices moleculares QuBiLs MIDAS, tanto totales (molécula completa) como locales (fragmento molecular), para las relaciones entre dos, tres y cuatro núcleos atómicos, incluyendo las representaciones matriciales de las estructuras químicas, y los operadores de agregación propuestos para generalizar la combinación lineal de las contribuciones atómicas. Como conclusión se puede decir 1) que se utilizaron métricas distintas a la Euclidiana para el cálculo de la distancia inter-atómica, 2) que en la diagonal de la matriz se tienen en consideración valores diferentes a cero (cantidad de lone-pairs o distancia de cada átomo al centro de la molécula), 3) que se propusieron los enfoques doble estocástico y de probabilidad mutua como estrategias de normalización, 4) que fueron empleadas y extendidas las formas algebraicas lineal, bilineal y cuadrática para el cálculo de índices moleculares basados en relaciones entre pares de átomos, 5) que fueron diseñadas matrices de tres y cuatro dimensiones para representar la información concerniente a las relaciones de ternas y cuaternas de átomos, 6) que se emplearon los conceptos del algebra multi-lineal para el cálculo de índices moleculares n-dimensionales ($n > 2$), y 7) que fueron propuestos operadores matemáticos distintos a la suma para el cálculo de índices moleculares a partir de un vector de índices atómicos (LOVIs).

Capítulo 3

Desarrollo del software ToMoCoMD-CARDD QuBiLs MIDAS

En el presente capítulo es presentado el diseño del software ToMoCoMD-CARDD QuBiLs MIDAS, que incluye el diagrama de clases para el cálculo de los índices que se proponen, y del módulo "Estructura" (Structure) para la limpieza y curado de datos. Son expuestos también, la complejidad computacional de los principales algoritmos, así como los resultados de las pruebas de rendimiento del procesamiento multi-núcleo.

3.1. Lenguaje de programación

Para la implementación del presente trabajo fue seleccionado Java [100] como lenguaje de programación. La razón principal de esta selección es que una aplicación informática desarrollada en este lenguaje puede ejecutarse en una gran variedad de equipos de cómputo independientemente de la arquitectura y sistema operativo que tengan. Al compilar un programa Java lo que se genera es un código intermedio (llamado bytecode), que en el momento de la ejecución es interpretado por la Máquina Virtual de Java (JVM - Java Virtual Machine), quien lo convierte a código nativo de la computadora donde se va a ejecutar. En los primeros días de Java, gran parte de sus críticas se originaban por su pobre rendimiento respecto a lenguajes nativos como C y Fortran. Mucho ha cambiado desde entonces. Actualmente se han producido enormes mejoras en el rendimiento de la JVM, principalmente atribuida a la introducción del compilador just-in-time [101] y la tecnología hotspot [102]. Estas mejoras han dado lugar a que la ejecución de la JVM

sea comparable a la de otros lenguajes nativos [103], y por lo tanto sea un lenguaje útil para el desarrollo de aplicaciones científicas. Además de las características mencionadas con anterioridad, Java constituye un lenguaje simple, orientado a objetos, distribuido, robusto, seguro, de altas prestaciones, multitarea y dinámico.

3.2. Biblioteca Chemical Development Kit (CDK)

Para el desarrollo del presente trabajo fue utilizada la biblioteca Chemistry Development Kit (CDK) [104]. Esta es una biblioteca de código abierto implementada en Java para problemas relacionados con la Informática Química y la Bioinformática. El CDK provee métodos para varias tareas comunes en informática molecular, entre las que se incluyen el renderizado 2D y 3D de estructuras químicas, flujos de lectura y escritura de los formatos de ficheros químicos más utilizados, generación e interpretación de SMILES, comprobación de isomorfismos, entre otras. El diseño de clases que presenta está relacionado con conceptos de la química y la biología, lo cual la convierte en una herramienta fácil de entender por especialistas en ciencias de la vida, y fácil de utilizar por especialistas en informática con conocimientos básicos en estas ciencias. Es ampliamente utilizada por varios proyectos con el objetivo de poner disponible sus funcionalidades, entre los que se destacan CDK-Taverna [105, 106], Cinfony[1] [107], Bioeclipse[2], entre otros.

3.3. Diseño del software ToMoCoMD-CARDD QuBiLs MIDAS

En esta sección son expuestos los principales componentes del modelo de diseño relacionados con el software desarrollado en este trabajo. Específicamente es presentado el diagrama de los paquetes arquitectónicamente significativos para el desarrollo de la aplicación ToMoCoMD-CARDD QuBiLs MIDAS, y por cada uno de estos paquetes el diagrama de clases del diseño correspondiente.

3.3.1. Diagrama de paquetes del diseño

En esta sección se muestra la vista de composición (ver Figura 3.1) y una breve descripción de los paquetes del diseño arquitectónicamente significativos de la aplicación ToMoCoMD-CARDD QuBiLs MIDAS. En cada uno de estos paquetes se encontrarán definidas las clases necesarias para realizar la configuración

[1] sitio oficial: http://code.google.com/p/cinfony/
[2] sitio oficial: http://bioclipse.net/

y cálculo de los índices moleculares presentados en este trabajo.

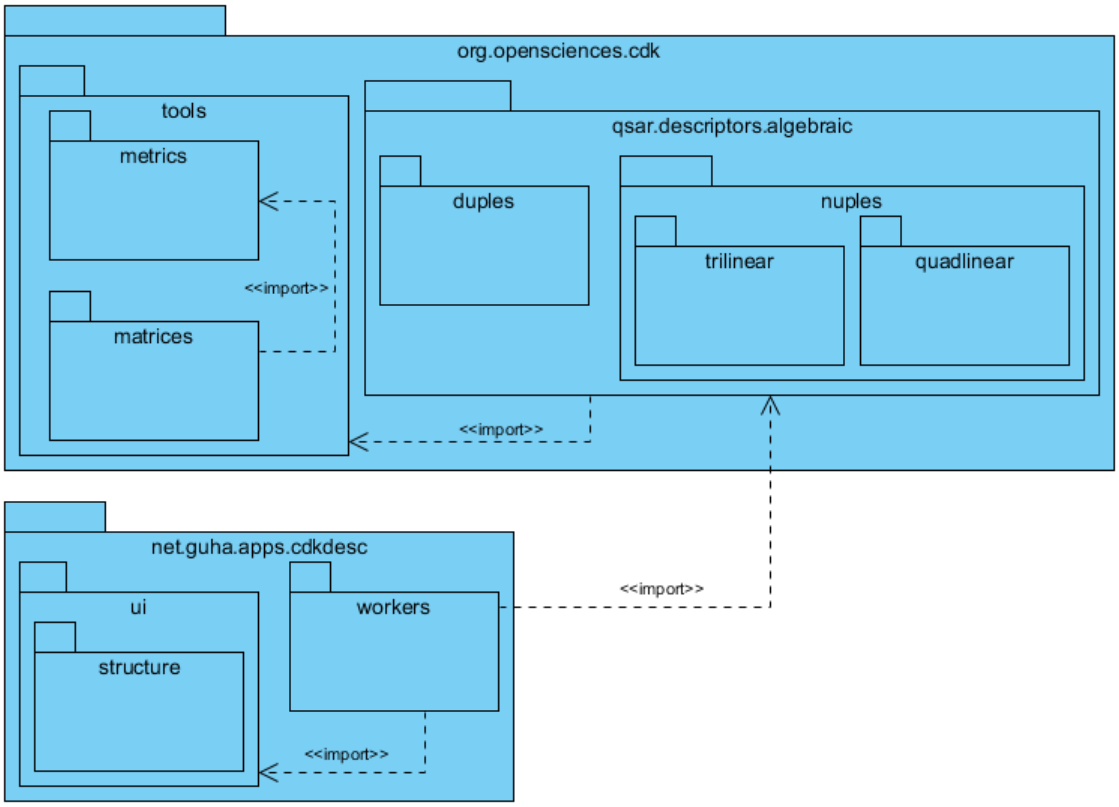

Figura 3.1: Vista de composición de los paquetes arquitectónicamente significativos del software ToMoCoMD-CARDD QuBiLs MIDAS

Descripción de los paquetes

org.opensciences.cdk.tools: contiene las clases que sirven de soporte o colaboran en el cálculo de los índices 3D QuBiLs MIDAS.

org.opensciences.cdk.tools.metrics: contiene las clases que definen las relaciones existentes entre los átomos de una molécula.

org.opensciences.cdk.tools.matrices: contiene las clases que definen los enfoques matriciales no estocásticos, simple estocásticos, doble estocásticos y de probabilidad mutua, a partir de la relaciones de duplas, ternas y cuaternas entre átomos de una molécula.

org.opensciences.cdk.qsar.descriptors.algebraic: contiene las clases que definen a los descriptores moleculares QuBiLs MIDAS.

org.opensciences.cdk.qsar.descriptors.algebraic.duples: contiene las clases que definen los des-

criptores moleculares QuBiLs MIDAS Duplas basados en las formas algebraicas lineal, bilineal y cuadrática.

org.opensciences.cdk.qsar.descriptors.algebraic.nuples.trilinear: contiene las clases que definen a los descriptores moleculares QuBiLs MIDAS Ternas.

org.opensciences.cdk.qsar.descriptors.algebraic.nuples.quadlinear: contiene las clases que definen a los descriptores moleculares QuBiLs MIDAS Cuaternas.

net.guha.apps.cdkdesc: contiene las clases responsables del cálculo y configuración de los descriptores moleculares QuBiLs MIDAS.

net.guha.apps.cdkdesc.ui: contiene las interfaces gráficas para la configuración de los diferentes enfoques de los descriptores moleculares QuBiLs MIDAS.

net.guha.apps.cdkdesc.structure: contiene las clases del módulo "Estructura" responsables de la limpieza de base de datos moleculares, así como de la conversión entre los formatos de ficheros soportados por el software ToMoCoMD-CARDD.

net.guha.apps.cdkdesc.workers: contiene las clases responsables de realizar el cálculo de los descriptores moleculares QuBiLs MIDAS entre los procesadores disponibles en una estación de trabajo.

3.3.2. Diagrama de clases del diseño del paquete *org.openscience.cdk.tools.metrics*

En este paquete se encuentran las clases responsables del cálculo de las distancias entre pares de átomos, así como de las medidas de asociación ternaria y cuaternaria entre núcleos atómicos. En la Figura 3.2 es mostrado el diagrama de clases correspondiente a este paquete, y por cada una de las clases que intervienen en el diseño se realiza una breve descripción.

Descripción de las clases

MetricFactory: el objetivo de esta clase es crear una métrica específica a partir de una configuración dada. Entiéndase por métrica, una función de distancia o una medida que relacione n átomos.

ICalculateMatrixValue: el objetivo de esta interfaz es definir un comportamiento a las clases que establezcan relaciones entre n átomos de una molécula.

Metric: el objetivo de esta clase es modelar una métrica que establezca una relación entre n átomos de una molécula a partir de las coordenadas (x, y, z) correspondientes.

MetricDuples: el objetivo de esta clase es modelar las métricas que establezcan relaciones entre dos átomos (ver Tabla 2.1).

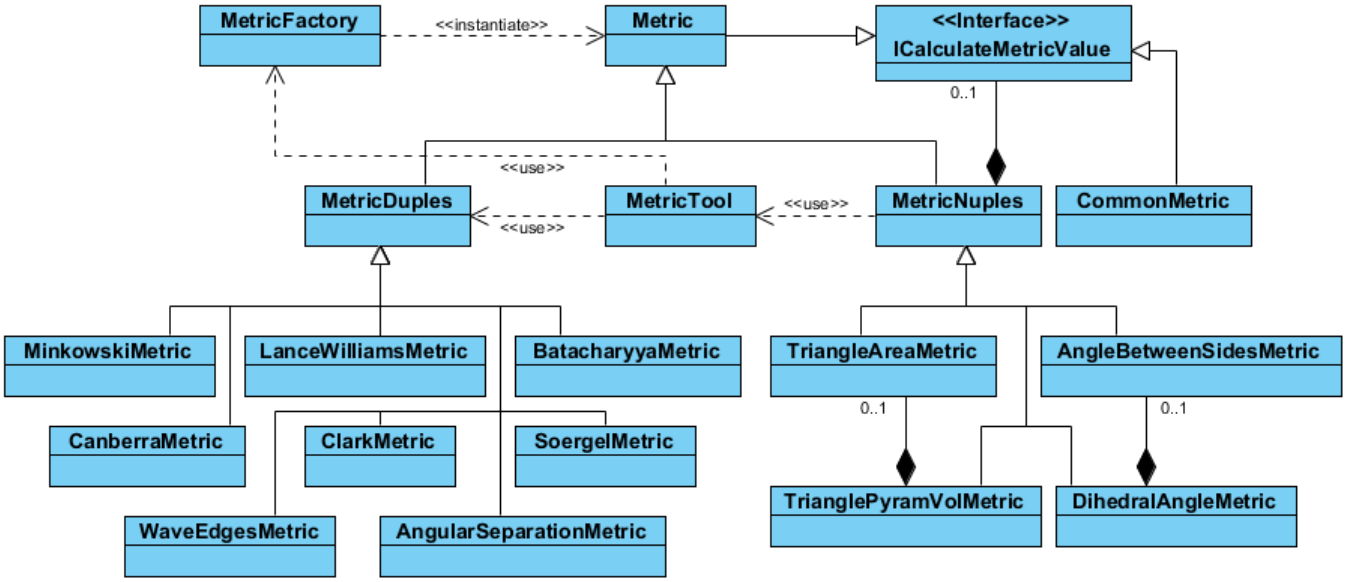

Figura 3.2: Diagrama de clases del diseño del paquete *org.openscience.cdk.tools.metrics*

MinkowskiMetric: el objetivo de esta clase es calcular la distancia entre dos átomos a partir de las coordenadas (x, y, z) utilizando la definición de Minkowski.

CanberraMetric: el objetivo de esta clase es calcular la distancia entre dos átomos a partir de las coordenadas (x, y, z) utilizando la distancia de Canberra.

LanceWilliamsMetric: el objetivo de esta clase es calcular la distancia entre dos átomos a partir de las coordenadas (x, y, z) utilizando la distancia Lance-Williams.

ClarkMetric: el objetivo de esta clase es calcular la distancia entre dos átomos a partir de las coordenadas (x, y, z) utilizando la distancia de Clark.

SoergelMetric: el objetivo de esta clase es calcular la distancia entre dos átomos a partir de las coordenadas (x, y, z) utilizando la distancia Soergel.

BattacharyyaMetric: el objetivo de esta clase es calcular la distancia entre dos átomos a partir de las coordenadas (x, y, z) utilizando la distancia Battacharyya.

WaveEdgesMetric: el objetivo de esta clase es calcular la distancia entre dos átomos a partir de las coordenadas (x, y, z) utilizando la distancia Wave-Edges.

AngularSeparationMetric: el objetivo de esta clase es calcular la distancia entre dos átomos a partir de las coordenadas (x, y, z) utilizando la distancia de Separación Angular.

MetricNuples: el objetivo de esta clase es modelar las métricas que establezcan relaciones entre más de dos átomos (ver Tabla 2.2).

CommonMetric: el objetivo de esta clase es calcular las medidas de perímetro y suma de lados que se establecen entres más de átomos de una molécula a partir de las coordenadas (x, y, z) correspondientes. Esto se debe a que la implementación de estas dos funcionalidades es igual independientemente de la cantidad de núcleos atómicos que relacionen.

TriangleAreaMetric: el objetivo de esta clase es calcular el área del triángulo que existe al relacionar tres átomos de una molécula a partir de las coordenadas (x, y, z) correspondientes.

TriangularPyramVolMetric: el objetivo de esta clase es calcular el volumen de la pirámide triangular que existe al relacionar cuatro átomos de una molécula a partir de las coordenadas (x, y, z) correspondientes.

AngleBetweenSidesMetric: el objetivo de esta clase es calcular el ángulo entre dos lados cuando se establece una relación entre tres átomos de una molécula.

DihedralAngleMetric: el objetivo de esta clase es calcular el ángulo diedro o de intersección entre dos planos al relacionar cuatro átomos de una molécula.

MetricTool: el objetivo de esta clase es colaborar en el cálculo de la relación que se establezca entre n átomos de una molécula, incluyendo la reducción del cálculo en otra métrica especificada cuando no todos los átomos de la relación original son diferentes.

3.3.3. Diagrama de clases del diseño del paquete *org.openscience.cdk.tools.matrices*

En este paquete se encuentran las clases correspondientes a la definición de los enfoques matriciales no estocástico, simple estocástico, doble estocástico y de probabilidad mutua, que son utilizados para el cálculo de los índices moleculares que se proponen en este trabajo. En la Figura 3.2 se muestra el diagrama de clases correspondiente a este paquete, y además en esta sección, aparece una descripción de las clases que lo conforman.

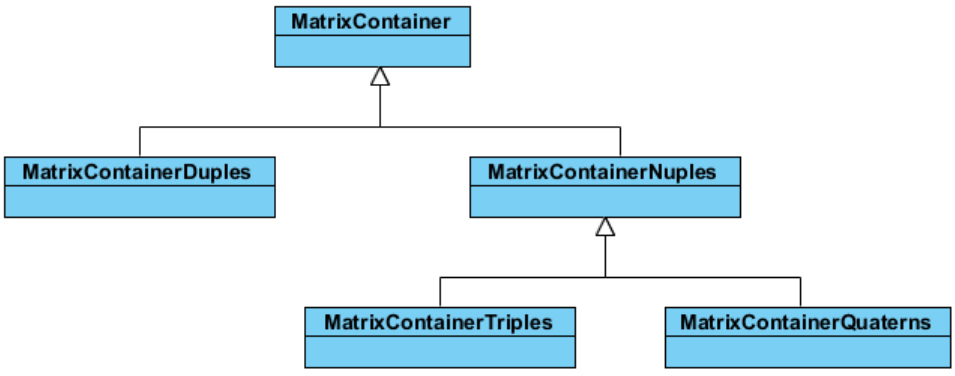

Tabla 3.1: Diagrama de clases del diseño del paquete *org.openscience.cdk.tools.matrices*

Descripción de las clases

MatrixContainer: el objetivo de esta clase es modelar los enfoques matriciales no estocástico, simple estocástico, doble estocástico y de probabilidad mutua para las relaciones que se establecen entre n ($n \geq 2$) átomos de una molécula.

MatrixContainerDuples: el objetivo de esta clase es definir los enfoques matriciales propuestos en este trabajo cuando se establecen relaciones entre dos átomos.

MatrixContainerNuples: el objetivo de esta clase es definir los enfoques matriciales propuestos en este trabajo cuando se establecen relaciones entre n ($n > 2$) átomos de una molécula.

MatrixContainerTriples: el objetivo de esta clase es definir los enfoques matriciales propuestos en este trabajo cuando se establecen relaciones entre tres átomos de una molécula.

MatrixContainerQuaterns: el objetivo de esta clase es definir los enfoques matriciales propuestos en este trabajo cuando se establecen relaciones entre cuatro átomos de una molécula.

3.3.4. Diagrama de clases del diseño del paquete *org.openscience.cdk.qsar.descriptors.algebraic*

En este paquete se encuentran las clases donde se definen los distintos tipos de índices moleculares basados en conceptos del Álgebra Lineal y Multilineal. El diagrama de clases correspondiente a este paquete puede apreciarse en la Figura 3.2. También son mostradas las relaciones existentes entre la clase base *AlgebraicDescriptor* y las otras clases pertenecientes a los paquetes *org.openscience.cdk.tools.metrics* y *org.openscience.cdk.tools.matrices.* En esta sección se realiza también una breve descripción de las clases correspondientes a este diseño.

Descripción de las clases

AlgebraicDescriptorFactory: el objetivo de esta clase es crear todos los posibles descriptores a calcular a partir de un conjunto de propiedades.

AlgebraicDescriptor: el objetivo de esta clase es modelar un descriptor molecular a partir de una relación n-dimensional ($n \geq 2$) entre núcleos atómicos de una molécula.

DuplesDescriptor: el objetivo de esta clase es modelar un descriptor molecular a partir de una relación entre dos átomos de una molécula y basado en conceptos del Álgebra Lineal sobre las formas lineales, bilineales y cuadráticas.

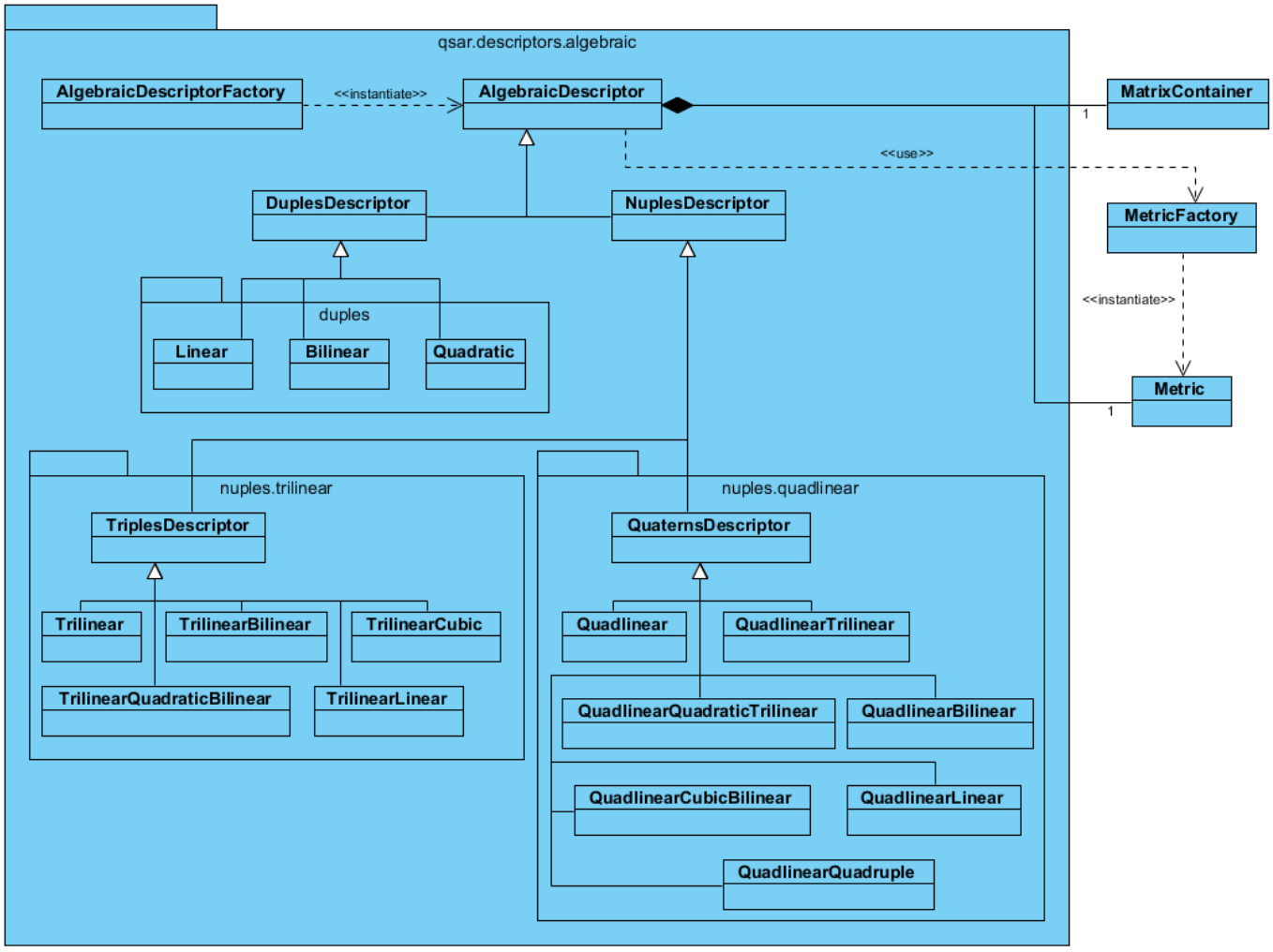

Tabla 3.2: Diagrama de clases del diseño del paquete *org.openscience.cdk.qsar.descriptors.algebraic*

Linear: el objetivo de esta clase es definir un descriptor molecular basado en el concepto algebraico de forma lineal.

Bilinear: el objetivo de esta clase es definir un descriptor molecular basado en el concepto algebraico de forma bilineal.

Quadratic: el objetivo de esta clase es definir un descriptor molecular basado en el concepto algebraico de forma cuadrática.

NuplesDescriptor: el objetivo de esta clase es modelar un descriptor molecular a partir de una relación entre más de dos átomos de una molécula y basado en conceptos del Álgebra Tensorial.

TriplesDescriptor: el objetivo de esta clase es modelar un descriptor molecular a partir de una relación entre tres átomos de una molécula y por lo tanto basado en el cálculo de tensores de orden tres.

Trilinear: el objetivo de esta clase es modelar un descriptor molecular basado en tensores de orden tres, donde cada vector del tensor está ponderado con una propiedad química diferente.

TrilinearQuadraticBilinear: el objetivo de esta clase es modelar un descriptor molecular basado en tensores de orden tres, donde los vectores $\bar{x}$ y $\bar{y}$ están ponderados con la misma propiedad química y ambos son distintos al vector $\bar{z}$ que está codificado con otra propiedad.

TrilinearBilinear: el objetivo de esta clase es modelar un descriptor molecular basado en tensores de orden tres, donde los vectores $\bar{x}$ y $\bar{y}$ están ponderados con propiedades químicas diferentes y el vector $\bar{z}$ es un vector unidad.

TrilinearCubic: el objetivo de esta clase es modelar un descriptor molecular basado en tensores de orden tres, donde cada vector del tensor está ponderado con la misma propiedad química.

TrilinearLinear: el objetivo de esta clase es modelar un descriptor molecular basado en tensores de orden tres, donde el vector $\bar{x}$ está ponderado con una propiedad química determinada y los vectores $\bar{y}$ y $\bar{z}$ son vectores unidad.

QuaternsDescriptor: el objetivo de esta clase es modelar un descriptor molecular a partir de una relación entre cuatro átomos de una molécula y por lo tanto basado en el cálculo de tensores de orden cuatro.

Quadlinear: el objetivo de esta clase es modelar un descriptor molecular basado en tensores de orden cuatro, donde cada vector del tensor está ponderado con una propiedad química diferente.

QuadlinearQuadraticTrilinear: el objetivo de esta clase es modelar un descriptor molecular basado en tensores de orden cuatro, donde los vectores $\bar{x}$ y $\bar{y}$ están ponderados con la misma propiedad química y ambos son distintos a los vectores $\bar{z}$ y $\bar{w}$, los cuales están codificados con propiedades químicas diferentes.

QuadlinearTrilinear: el objetivo de esta clase es modelar un descriptor molecular basado en tensores de orden cuatro, donde el vector $\bar{x}$ es un vector unidad y los vectores $\bar{y}$, $\bar{z}$ y $\bar{w}$ están ponderados con propiedades químicas distintas.

QuadlinearCubicBilinear: el objetivo de esta clase es modelar un descriptor molecular basado en tensores de orden cuatro, donde los vector $\bar{x}$, $\bar{y}$ y $\bar{z}$ están codificados con la misma propiedad química y los tres son distintos al vector $\bar{w}$ que está ponderado con otra propiedad.

QuadlinearBilinear: el objetivo de esta clase es modelar un descriptor molecular basado en tensores de orden cuatro, donde los vectores $\bar{x}$ y $\bar{y}$ son vectores unidad y los vectores $\bar{z}$ y $\bar{w}$ están ponderados con propiedades químicas distintas.

QuadlinearQuadruple: el objetivo de esta clase es modelar un descriptor molecular basado en tensores

de orden cuatro, donde cada vector del tensor está ponderado con la misma propiedad química.

QuadlinearLinear: el objetivo de esta clase es modelar un descriptor molecular basado en tensores de orden cuatro, donde los vectores $\bar{x}$, $\bar{y}$ y $\bar{z}$ son vectores unidad y el vector $\bar{w}$ está codificado con una propiedad química.

MatrixContainer: ver sección 3.3.3.

MetricFactory: ver sección 3.3.2.

Metric: ver sección 3.3.2.

3.3.5. Diagrama de clases del diseño del paquete *net.guha.apps.cdkdesc*

En este paquete están definidas las clases responsables de la configuración y cálculo de los índices moleculares QuBiLs MIDAS. Se puede observar en la Figura 3.3 el diagrama de clases correspondiente, donde se muestra además, las relaciones existentes con otras clases del paquete *org.openscience.cdk.-qsar.descriptors.-algebraic*. Esta sección culmina con una descripción de las clases que intervienen en el diseño.

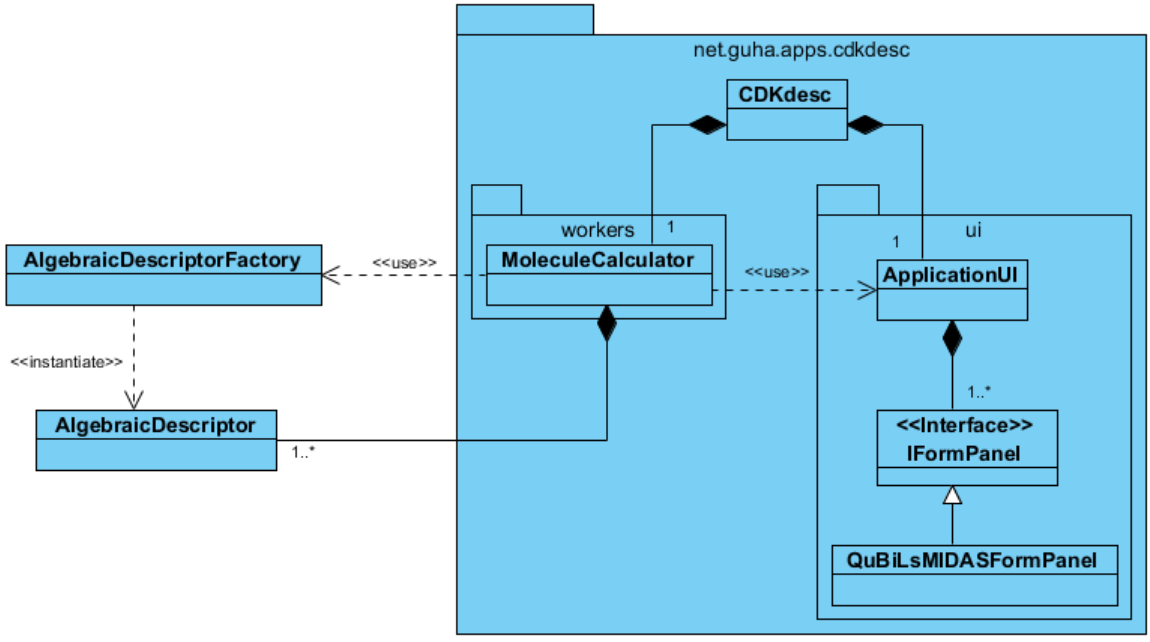

Tabla 3.3: Diagrama de clases del diseño del paquete *net.guha.apps.cdkdesc*

Descripción de las clases

CDKdesc: el objetivo de esta clase es controlar el intercambio de información entre el usuario final y las interfaces gráficas de configuración de los índices moleculares, así como de iniciar el proceso para su

cálculo.

ApplicationUI: esta clase es la encargada de gestionar las interfaces gráficas del software ToMoCoMD-CARDD.

IFormPanel: el objetivo de esta interfaz es definir el comportamiento a seguir por cada uno de los módulos que sean adicionados al software ToMoCoMD-CARDD.

QuBiLsMIDASFormPanel: en esta clase se encuentran definidos todos los enfoques correspondientes a los índices QuBiLs MIDAS.

MoleculeCalculator: el objetivo de esta clase es realizar el procesamiento multi-núcleo de los índices moleculares que fueron especificados por el usuario.

AlgebraicDescriptorFactory: ver sección 3.3.4

AlgebraicDescriptor: ver sección 3.3.4

3.3.6. Diagrama de clases del diseño del paquete *net.guha.apps.cdkdesc.structure*

En este paquete se encuentran declaradas las clases responsables de la limpieza de las base de datos de compuestos orgánicos que pudieran ser utilizadas para el cálculo de los índices moleculares. La correctitud de las datas de entradas es un requisito fundamental para los estudios quimio-informáticos [108]. En la Figura 3.3 es mostrado el diagrama de clases correspondiente a este paquete, y además en esta sección, es realizada la descripción de las clases que intervienen en el diseño.

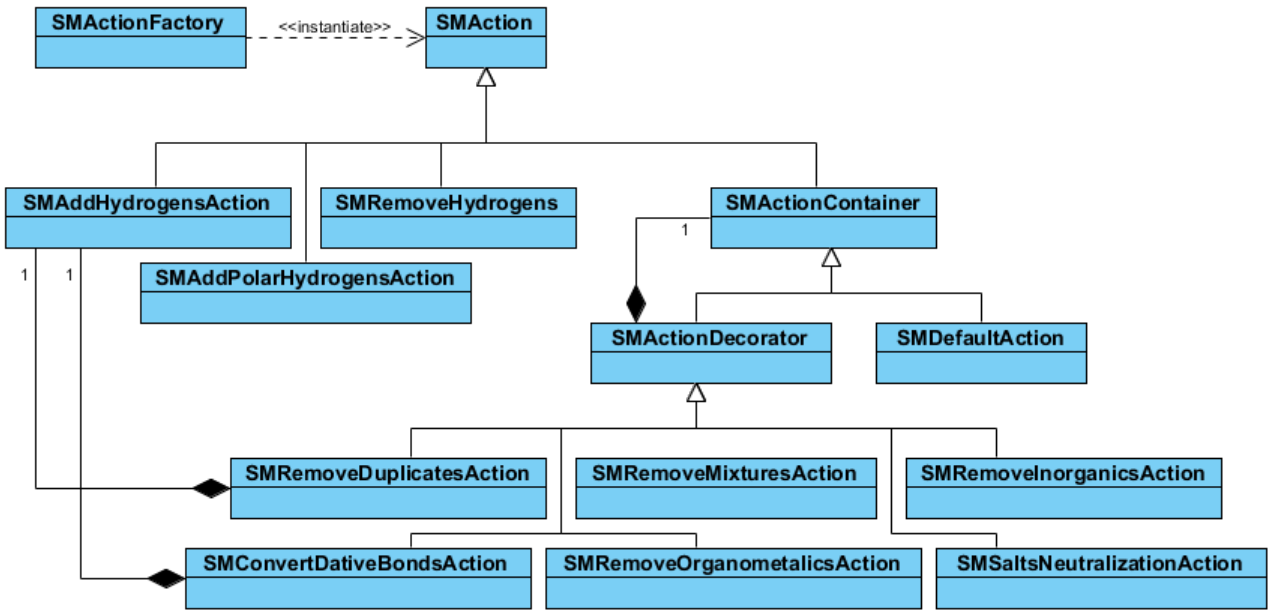

Figura 3.3: Diagrama de clases del diseño del paquete *net.guha.apps.cdkdesc.structure*

Descripción de las clases

SMActionFactory: el objetivo de esta clase es crear una acción que será realizada sobre una base de datos de compuestos químicos.

SMAction: el objetivo de esta clase es modelar una acción genérica sobre una base de datos de compuestos químicos.

SMAddHydrogensAction: el objetivo de esta clase es adicionar explícitamente átomos de hidrógeno a los átomos de una molécula.

SMAddPolarHydrogensAction: el objetivo de esta clase es adicionar explícitamente átomos de hidrógeno a los átomos de una molécula que sean distintos al carbono.

SMRemoveHydrogensAction: el objetivo de esta clase es eliminar los átomos de hidrógeno de una molécula.

SMActionContainer: el objetivo de esta clase es modelar una acción genérica para el curado de datos.

SMDefaultAction: el objetivo de esta clase es definir una acción por defecto sobre una data de compuestos químicos.

SMActionDecorator: el objetivo de esta clase es modelar objetos que encapsulen una acción genérica para el curado de datos, y que a su vez contienen otra acción que será realizada en forma recursiva. Este comportamiento se mantiene hasta que los objetos realicen una acción concreta sobre los datos químicos.

SMRemoveDuplicatesAction: el objetivo de esta clase es eliminar las estructuras químicas repetidas en un conjunto de datos a analizar. Esto es realizado mediante la comparación de los SMILES canónicos de los compuestos [108], de forma tal que si existen dos iguales entonces significa que las estructuras son las mismas.

SMRemoveMixturesAction: el objetivo de esta clase es eliminar en los compuestos de una base de datos todas las mezclas existentes. Por lo general se entiende por mezclas aquellas estructuras que están formadas por una molécula orgánica grande y moléculas inorgánicas pequeñas, tales como, compuestos hidrohalogenados. En este caso solamente se deja la molécula de mayor peso.

SMRemoveInorganicsAction: el objetivo de esta clase es eliminar todas aquellas estructuras que no contienen átomos de carbono y por lo tanto son compuestos inorgánicos.

SMRemoveOrganometalicsAction: el objetivo de esta clase es remover de la base de datos todos aquellos compuestos donde los átomos de carbono se encuentran formando enlaces covalentes con algún metal.

SMConvertDativeBondsAction: el objetivo de esta clase es eliminar de las estructuras químicas los enlaces de coordinación que pudieran tener. Un enlace de coordinación o dative bond, es aquel donde los dos electrones que forman un enlace son donados por el mismo núcleo atómico.

SMSaltsNeutralizationAction: el objetivo de esta clase es neutralizar todas las sales existentes en las estructuras químicas de una base de datos. Con este fin se buscaría por cada estructura, todas las moléculas orgánicas con cargas opuestas, y entonces se dejaría la de mayor peso. También las sales pueden estar constituidas por compuestos hidrohalogenados, como el *HBr*, que aparecen como elementos no conexos. En este último caso la neutralización sería eliminar dichos compuestos.

3.4. Complejidad temporal de los principales algoritmos

En esta sección son presentados los pseudocódigos de los algoritmos fundamentales para el cálculo de los índices QuBiLs MIDAS, entre ellos los relacionados con las transformaciones matriciales simple estocástica y de probabilidad mutua. La transformación doble estocástica es realizada con el procedimiento Sinkorn-Knopp [96]. También están detallados los algoritmos para el cálculo de los índices atómicos QuBiLs MIDAS. Además se determina por cada algoritmo la complejidad computacional correspondiente.

3.4.1. Transformaciones algebraicas de matrices no estocásticas a matrices de probabilidad mutua

A partir de la definición realizada en la Ecuación 2.8, en el Algoritmo 3.1 es mostrado el pseudocódigo para transformar una matriz no estocástica basada en dupla de átomos a una matriz de probabilidad mutua. El funcionamiento básico de este algoritmo es que a partir de una matriz bidimensional simétrica cuadrada, se suman todos sus elementos y posteriormente cada uno de ellos es dividido entre la suma total calculada. La complejidad temporal de este procedimiento es de $O(n^2)$ y está determinado por la ecuación 3.1.

$$T(n) = 3 + 2 * \left[\sum_{i=1}^{n} \left(4 + \sum_{j=i+1}^{n} 4 \right) \right] = O(n^2) \tag{3.1}$$

De forma análoga son definidos los procedimientos para la transformación a probabilidad mutua de matrices (simétricas o no simétricas) no estocásticas pero basadas en relaciones de ternas (Ecuación 2.9) y cuaternas (Ecuación 2.10) de átomos. En los Algoritmos 3.2 y 3.3 se presenta el pseudocódigo y en las ecuaciones 3.2 y 3.3 la función de tiempo correspondiente a cada uno, siendo la complejidad temporal de

$O(n^3)$ y $O(n^4)$ respectivamente.

$$T(n) = 3 + 2 * \left[\sum_{i=1}^{n} \left(3 + \sum_{j=1}^{n} \left(3 + \sum_{k=1}^{n} 3 \right) \right) \right] = O(n^3) \tag{3.2}$$

$$T(n) = 3 + 2 * \left[\sum_{i=1}^{n} \left(3 + \sum_{j=1}^{n} \left(3 + \sum_{k=1}^{n} \left(3 + \sum_{w=1}^{n} 3 \right) \right) \right) \right] = O(n^4) \tag{3.3}$$

Algoritmo 3.1 Algoritmo para transformar una matriz bidimensional simétrica cuadrada no estocástica a una matriz de probabilidad mutua

```
1  real suma = 0
2  para i = 1 hasta n hacer
3      suma = suma + M[i][i]
4      para j = i + 1 hasta n hacer
5          suma = suma + M[i][j]
6          suma = suma + M[j][i]
7      fin para
8  fin para
9  para i = 1 hasta n hacer
10     M[i][i] = M[i][i] / suma
11     para j = i + 1 hasta n hacer
12         M[i][j] = M[i][j] / suma
13         M[j][i] = M[j][i] / suma
14     fin para
15 fin para
```

Algoritmo 3.2 Algoritmo para transformar una matriz tridimensional cuadrada no estocástica a una matriz de probabilidad mutua

```
1  real suma = 0
2  para i = 1 hasta n hacer
3      para j = 1 hasta n hacer
4          para k = 1 hasta n hacer
5              suma = suma + M[i][j][k]
6          fin para
7      fin para
8  fin para
9  para i = 1 hasta n hacer
10     para j = 1 hasta n hacer
11         para k = 1 hasta n hacer
12             M[i][j][k] = M[i][j][k] / suma
13         fin para
14     fin para
15 fin para
```

Algoritmo 3.3 Algoritmo para transformar una matriz cuadrada no estocástica de dimensión cuatro a una matriz de probabilidad mutua

```
1  real suma = 0
2  para i = 1 hasta n hacer
3      para j = 1 hasta n hacer
4          para k = 1 hasta n hacer
5              para w = 1 hasta n hacer
6                  suma = suma + M[i][j][k][w]
7              fin para
8          fin para
9      fin para
10 fin para
11 para i = 1 hasta n hacer
12     para j = 1 hasta n hacer
13         para k = 1 hasta n hacer
14             para w = 1 hasta n hacer
15                 M[i][j][k][w] = M[i][j][k][w] / suma
16             fin para
17         fin para
18     fin para
19 fin para
```

3.4.2. Transformaciones algebraicas de matrices no estocásticas a matrices simple estocásticas

Otra de las transformaciones matriciales planteadas en este trabajo es la relacionada con matrices simple estocásticas a partir de matrices no estocásticas. Para este procedimiento se muestra en el Algoritmo 3.4 el pseudocódigo correspondiente a esta transformación para matrices bidimensionales cuadradas simétricas (Ecuación 2.4), el cual consiste en sumar todas las entradas de una fila determinada y luego dividir cada entrada entre la suma calculada. La complejidad de este procedimiento es de $O(n^2)$ y su función de tiempo está especificada en la ecuación 3.4.

$$T(n) = 2 + \sum_{i=1}^{n} \left(4 + \sum_{j=1}^{n} 3 + \sum_{j=1}^{n} 3 \right) = O(n^2) \tag{3.4}$$

Como los valores de una matriz simple estocástica bidimensional son calculados a partir de la relación existente entre un átomo i y un átomo j dividido entra la suma de las relaciones entre el átomo i y el resto de los átomos de la molécula, entonces las transformaciones simple estocásticas para matrices de más de dos dimensiones deben seguir un comportamiento análogo. Siguiendo este principio, en el Algoritmo y en la Ecuación 3.5 es mostrado el pseudocódigo y la función de tiempo, respectivamente, correspondiente a esta

transformación para matrices tridimensionales (ver Ecuación 2.5). Como puede observarse la complejidad computacional es de $O(n^3)$.

$$T(n) = 1 + \sum_{i=1}^{n} \left(5 + 2 * \left[\sum_{j=1}^{n} \left(3 + \sum_{k=1}^{n} 3 \right) \right] \right) = O(n^3) \tag{3.5}$$

Por otra parte, la transformación algebraica simple estocástica para matrices no estocásticas de dimensión cuatro (Ecuación 2.6) es realizada como se muestra en el Algoritmo 3.6. La complejidad computacional de este procedimiento es de $O(n^4)$ y la función temporal está especificada en la ecuación 3.6.

$$T(n) = 1 + \sum_{i=1}^{n} \left(5 + 2 * \left[\sum_{j=1}^{n} \left(3 + \sum_{k=1}^{n} \left(3 + \sum_{w=1}^{n} 3 \right) \right) \right] \right) = O(n^4) \tag{3.6}$$

Algoritmo 3.4 Algoritmo para transformar una matriz bidimensional simétrica cuadrada no estocástica a una matriz simple estocástica

```
1  real suma = 0
2  para i = 1 hasta n hacer
3      para j = 1 hasta n hacer
4          suma = suma + M[i][j]
5      fin para
6      para j = 1 hasta n hacer
7          M[i][j] = M[i][j] / suma
8      fin para
9  fin para
```

Algoritmo 3.5 Algoritmo para transformar una matriz tridimensional simétrica cuadrada no estocástica a una matriz simple estocástica

```
1  para i = 1 hasta n hacer
2      real suma = 0
3      para j = 1 hasta n hacer
4          para k = 1 hasta n hacer
5              suma = suma + M[i][j][k]
6          fin para
7      fin para
8      para j = 1 hasta n hacer
9          para k = 1 hasta n hacer
10             M[i][j][k] = M[i][j][k] / suma
11         fin para
12     fin para
13 fin para
```

Algoritmo 3.6 Algoritmo para transformar una matriz cuadrada no estocástica de dimensión cuatro a una matriz simple estocástica

```
1  para i = 1 hasta n hacer
2      real suma = 0
3      para j = 1 hasta n hacer
4          para k = 1 hasta n hacer
5              para w = 1 hasta n hacer
6                  suma = suma + M[i][j][k][w]
7              fin para
8          fin para
9      fin para
10     para j = 1 hasta n hacer
11         para k = 1 hasta n hacer
12             para w = 1 hasta n hacer
13                 M[i][j][k][w] = M[i][j][k][w] / suma
14             fin para
15         fin para
16     fin para
17 fin para
```

3.4.3. Cálculo de las contribuciones atómicas de los índices QuBiLs MIDAS Duplas

En esta sección es presentado el procedimiento para determinar las contribuciones atómicas correspondientes a los índices QuBiLs MIDAS Duplas. Este algoritmo es el que se aplica independientemente de la forma algebraica lineal, bilineal o cuadrática que se utilice. Cuando los vectores de propiedades de entrada al algoritmo $\bar{x}$ y $\bar{y}$ son diferentes se calculan los índices atómicos bilineales, cuando son iguales son calculados los índices atómicos cuadráticos, y cuando el vector $\bar{x}$ es el vector unidad entonces se calculan los índices atómicos lineales. La implementación trivial de este procedimiento tiene una cota temporal de $O(n^3)$, y como puede analizarse en el pseudocódigo mostrado en el Algoritmo 3.7, el mismo está mejorado, teniendo de esta forma una complejidad temporal determinada por la ecuación 3.7 de $O(n^2)$.

$$T(n) = 3 + \sum_{i=1}^{n} \left(5 + \sum_{j=1}^{n} 8 \right) = O(n^2) \tag{3.7}$$

3.4.4. Cálculo de las contribuciones atómicas de los índices QuBiLs MIDAS Ternas

En esta sección se presenta el procedimiento para el cálculo de las contribuciones atómicas de los índices ternarios QuBiLs MIDAS. Este algoritmo es genérico para cualquiera de las combinaciones de los vectores de propiedades $\bar{x}$, $\bar{y}$ y $\bar{z}$, siempre y cuando contribuyan al cálculo de alguna de las formas trilineales definidas en

Algoritmo 3.7 Algoritmo para calcular las contribuciones atómicas de los índices QuBiLs MIDAS Duplas

```
1  Entrada: vector de propiedades X, Y; arreglo bidimensional M
2  Salida: el vector de contribuciones atómicas
3
4  arreglo L
5  para iAtom = 1 hasta M.longitud hacer
6      arreglo auxL
7      para col = 1 hasta M.longitud hacer
8          si col == iAtom
9              auxL[iAtom] = auxL[iAtom] + (M[iAtom][col] * Y[col])
10         sino
11             auxL[iAtom] = auxL[iAtom] + ((M[iAtom][col] / 2) * Y[iAtom])
12             auxL[col] = auxL[col] + ((M[col][iAtom] / 2) * Y[col])
13         fin si
14     fin para
15     L[iAtom] = dot product(auxL, X)
16 fin para
17
18 retornar L
```

la Ecuación 2.14. En el Algoritmo 3.8 es mostrado el pseudocódigo correspondiente a este procedimiento, el cual mejora la implementación trivial, que tendría una complejidad de $O(n^4)$. En este caso la cota máxima temporal es de $O(n^3)$ y está determinada por la ecuación 3.8.

$$T(n) = 4 + \sum_{i=1}^{n} \left(7 + \sum_{j=1}^{n} \left(5 + \sum_{k=1}^{n} 21 \right) + \sum_{j=1}^{n} 6 \right) = O(n^3) \tag{3.8}$$

3.4.5. Cálculo de las contribuciones atómicas de los índices QuBiLs MIDAS Cuaternas

El procedimiento para el cálculo de las contribuciones atómicas de los índices cuaternarios QuBiLs MIDAS es presentado en el Algoritmo 3.9. Independientemente de las combinaciones de los vectores de propiedades $\bar{x}$, $\bar{y}$, $\bar{z}$ y $\bar{w}$, este procedimiento es genérico para el cálculo de las formas cuatrilineales especificadas en la Ecuación 2.15. La complejidad temporal de este algoritmo es de $O(n^4)$, siendo mejor que la implementación trivial para este procedimiento, que tendría una cota de $O(n^5)$. Al analizar el pseudocódigo mostrado se obtiene la ecuación 3.9, que es la que determina la complejidad algorítmica del procedimiento propuesto en esta sección.

$$T(n) = 6 + \sum_{i=1}^{n} \left(5 + \sum_{j=1}^{n} \left(10 + \sum_{k=1}^{n} \left(5 + \sum_{l=1}^{n} 23 \right) + \sum_{k=1}^{n} 9 \right) + \sum_{j=1}^{n} 6 \right) = O(n^4) \tag{3.9}$$

Algoritmo 3.8 Algoritmo para calcular las contribuciones atómicas de los índices QuBiLs MIDAS Ternas

```
1  Entrada: vector de propiedades X, Y, Z; arreglo tridimensional M
2  Salida: el vector de contribuciones atómicas
3
4  arreglo real L
5  arreglo real tempL
6  para iAtom = 1 hasta M.longitud hacer
7      arreglo real auxL
8      arreglo real atomL
9      para row = 1 hasta M.longitud hacer
10         para col = 1 hasta M.longitud hacer
11             si row <> iAtom
12                 real prod = 2 / 3
13                 si col <> iAtom
14                     prod = 1 / 3
15                 fin si
16                 auxL[iAtom] = auxL[iAtom] + ((M[row][iAtom][col] * prod) * X[col]))
17                 si col <> iAtom
18                     auxL[col] = auxL[col] + ((M[row][col][iAtom] * prod) + X[iAtom])
19                 fin si
20             fin si
21             si col >= row
22                 real cont = 1
23                 si row == iAtom
24                     cont = cont + 1
25                 fin si
26                 si col == iAtom
27                     cont = cont + 1
28                 fin si
29                 atomL[row] = atomL[row] + ((M[iAtom][row][col] * (cont / 3)) + X[col])
30                 si col > row
31                     atomL[col] = atomL[col] + ((M[iAtom][col][row] * (cont / 3)) + X[row])
32                 fin si
33             fin si
34         fin para
35         si row <> iAtom
36             tempL[row] = dot product(Y, auxL)
37         fin si
38     fin para
39     tempL[iAtom] = dot product(Y, atomL)
40     L[iAtom] = dot product(Z, tempL)
41 fin para
42
43 retornar L
```

Algoritmo 3.9 Algoritmo para calcular las contribuciones atómicas de los índices QuBiLs MIDAS Cuaternas

```
1  Entrada: vector de propiedades X, Y, Z, W; arreglo cuatridimensional M
2  Salida: el vector de contribuciones atómicas
3
4  arreglo real L, temp1L, temp2L, temp3L
5  para iAtom = 1 hasta M.longitud hacer
6      para iMatrix hasta M.longitud hacer
7          arreglo real auxL, atom1L, atom2L
8          para row = 1 hasta M.longitud hacer
9              para col = 1 hasta M.longitud hacer
10                 real cont = (si iMatrix == iAtom entonces 1 sino 0) + (si row == iAtom entonces 1 sino
                       0) + (si col == iAtom entonces 1 sino 0) + 1
11                 si iMatrix <> iAtom
12                     si row <> iAtom
13                         auxL[iAtom]=auxL[iAtom]+((M[iMatrix][row][iAtom][col]*(cont/4))*X[col])
14                         si col <> iAtom
15                             auxL[col]=auxL[col]+((M[iMatrix][row][col][iAtom]*(cont/4))*X[iAtom])
16                         fin si
17                     fin si
18                     si col >= row
19                         atom1L[row]=atom1L[row]+((M[iMatrix][iAtom][row][col])*(cont/4)*X[col])
20                         atom2L[row]=atom2L[row]+((M[iAtom][iMatrix][row][col])*(cont/4)*X[col])
21                         si col > row
22                             atom1L[col]=atom1L[col]+((M[iMatrix][iAtom][col][row]*(cont/4))*X[row])
23                             atom2L[col]=atom2L[col]+((M[iAtom][iMatrix][col][row]*(cont/4))*X[row])
24                         fin si
25                     fin si
26                 sino si col >= row
27                     atom2L[row]=atom2L[row]+((M[iAtom][iMatrix][row][col])*(cont/4)*X[col])
28                     si col > row
29                         atom2L[col]=atom2L[col]+((M[iAtom][iMatrix][col][row]*(cont/4))*X[row])
30                     fin si
31                 fin si
32             fin para
33             si iMatrix <> iAtom && row <> iAtom
34                 temp2L[row] = dot product(auxL, Y)
35             fin si
36         fin para
37         si iMatrix <> iAtom
38             temp2L[iAtom] = dot product(atom1L, Y)
39             temp1L[iMatrix] = dot product(temp2L, Z)
40         fin si
41         temp3L[iMatrix] = dot product(atom2L, Y)
42     fin para
43     temp1L[iAtom] = dot product(temp3L, Z)
44     L[iAtom] = dot product(temp1L, W)
45 fin para
46
47 retornar L
```

3.5. Análisis de escalabilidad en el procesamiento multi-núcleo del software ToMoCoMD-CARDD QuBiLs MIDAS

Para evaluar la capacidad de procesamiento fueron realizadas pruebas de rendimiento a la aplicación ToMoCoMD-CARDD QuBiLs MIDAS, utilizando para ello una estación de trabajo cuyas características son mostradas en la Tabla 3.4. Es válido aclarar que la arquitectura corei7 integra cuatro núcleos de forma nativa *(single die)*, cada uno con tecnología *Simultaneous Multi-Threading (SMT)*. Esto posibilita ejecutar dos instrucciones por cada ciclo de reloj, permitiendo ejecutar ocho hebras de procesamiento.

Motherboard	Intel DH61AGL
Sistema operativo	Windows 7 Ultimate Service Pack 1
Memoria RAM	16 GB
CPU	Intel(R) Core(TM) i7-2600k CPU @ 3.40GHz 3.40GHz

Tabla 3.4: Característica de la estación de trabajo utilizada para realizar las pruebas de procesamiento multi-núcleo

En este experimento, para probar el software desarrollado fue utilizada la base de datos *PrimSreen15*[3] para el cálculo de 4914 índices moleculares QuBiLs MIDAS Duplas, basados en las formas matriciales no estocástica, simple estocástica y de probabilidad mutua. La configuración del proyecto para el cálculo de estos índices es mostrado en el Anexo B. Las pruebas fueron realizadas aumentando la cantidad de procesadores para el mismo conjunto de datos.

En la Tabla 3.5 son mostrados, el tiempo de procesamiento del software desarrollado a medida que la cantidad de procesadores fue aumentándose, y los correspondientes valores de *speed up* alcanzados. En la Figura 3.4 es visualizado gráficamente el *speed up* obtenido cuando el número de procesadores a utilizar en el procesamiento es variado. Puede notarse como el cómputo paralelo de los índices moleculares cuando es utlizado dos procesadores tiene buen comportamiento, sin embargo para cuatro u ocho tiene un menor desempeño.

Procesadores	**Tiempo (s)**	**Speed Up**
1	1416	1
2	782	1.81
4	578	2.45
8	385	3.68

Tabla 3.5: Tiempo de respuesta y speed up alcanzado por el software a medida que aumenta el número de procesadores

[3]URL: http://www.otavachemicals.com/download-compound-libraries/doc_details/10-primscreen-15-db

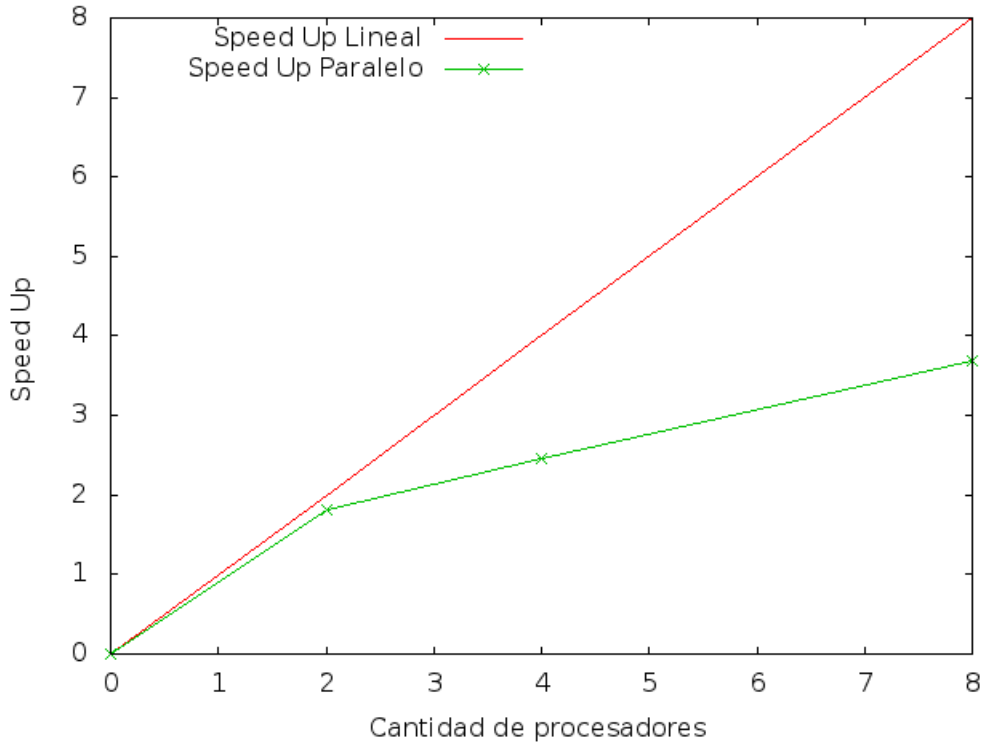

Figura 3.4: Speed Up alcanzado al variar la cantidad de procesadores

3.6. Conclusiones parciales

En este capítulo fueron presentados los diagramas más importantes en el diseño del software ToMoCoMD-CARDD QuBiLs MIDAS, el pseudocódigo y la complejidad algorítmica de los procedimientos críticos en el cálculo de los índices moleculares, y los resultados de las pruebas de escalabilidad multi-núcleo realizadas. Como conclusión se puede decir 1) que la aplicación desarrollada permite la configuración, mediante una interfaz gráfica, de todos los enfoques para el cálculo de los descriptores moleculares; 2) que el diseño realizado brinda flexibilidad para adicionar nuevas características (propiedades químicas, operadores de agregación, entre otras) para el cómputo de los índices propuestos; 3) que se implementaron los algoritmos de la forma más eficiente posible, de manera tal que posibilite calcular los índices moleculares en un tiempo razonable; y 4) que aprovecha las características multi-núcleo de las arquitecturas de computadoras modernas, haciendo uso de todos los procesadores presentes en una estación de trabajo.

Capítulo 4

Validación de los nuevos índices moleculares

En el presente capítulo es presentada la validación de los índices propuestos mediante las técnicas de análisis de variabilidad y análisis de componentes principales. Además, son expuestos los resultados obtenidos en la realización de un estudio QSAR/QSPR, como una medida de la utilidad de los nuevos índices que se introducen.

En todos los análisis son evaluados internamente los índices QuBiLs MIDAS mediante las realizaciones de estudios comparativos respecto a los enfoques matriciales, las métricas de distancias inter-atómicas, las medidas de asociación ternarias y cuaternarias entre átomos, y los operadores de agregación de las contribuciones atómicas. En todos los casos, excepto en la modelación QSAR, fueron utilizadas datas no cogenéricas de estructuras químicas; específicamente, en los estudios 4.1.1, 4.1.2, 4.1.4, 4.2.1, 4.2.2, 4.2.4 y 4.2.5.1 fue utilizada la base de datos Spectrum (1962 moléculas), y en el resto de los estudios fue empleada la data PrimScreen1 [109] (998 moléculas).

4.1. Análisis de variabilidad basado en Entropía de Shannon de los índices QuBiLs MIDAS y comparación con otros enfoques

En esta sección es evaluada la calidad de los índices 3D QuBiLs MIDAS en términos de variabilidad, mostrándose los resultados del comportamiento interno de los enfoques más importantes, y culminando con una comparación entre los índices QuBiLs MIDAS respecto a los descriptores calculados por otras aplicaciones. Para los estudios 4.1.1, 4.1.2 y 4.1.4 que utilizaron la data Spectrum, la máxima entropía está determinada por $log_2 981 = 9{,}93$ bits, mientras que para la data PrimScreen1 la máxima entropía es igual

a $log_2 998 = 9{,}96$ bits.

4.1.1. Análisis comparativo de los índices QuBiLs MIDAS según el enfoque matricial

El objetivo de este estudio es evaluar la contribución, en términos de variabilidad, de los índices 3D QuBiLs MIDAS según el formalismo de matriz no estocástica, simple estocástica, doble estocástica y de probabilidad mutua. Con este fin fueron calculadas 1300 variables para cada uno de los enfoques. Como puede observarse en la Figura 4.1, distribuciones de entropía comparables son observadas para los índices basados en los enfoques de probabilidad mutua, simple estocástico y doble estocástico. Sin embargo, si son comparadas las mejores 200 variables, tomando en consideración sus valores de entropía, puede notarse que el porciento de las variables derivadas del enfoque de probabilidad mutua, presentan mejor comportamiento, lo cual implica que descriptores con alta entropía son obtenidos con este formalismo. Este resultado justifica la contribución teórica del presente trabajo, donde el enfoque de probabilidad mutua es introducido. Por otro lado, puede observarse claramente que los índices basados en el enfoque no estocástico, presentan el peor patrón de distribución. Estos resultados sugieren que un gran porcentaje de descriptores moleculares con alta variabilidad son obtenidos con los enfoques de matriz de probabilidad mutua y matriz simple estocástica.

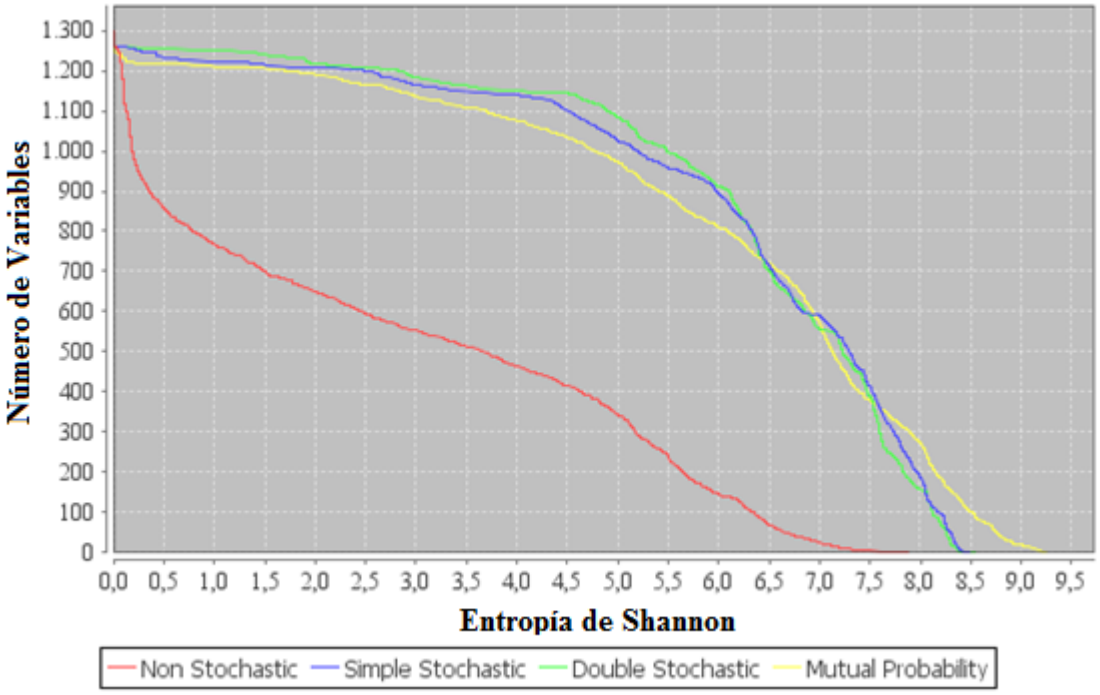

Figura 4.1: Distribución de la entropía de Shannon de los índices 3D según los enfoques matriciales

4.1.2. Análisis comparativo de los índices QuBiLs MIDAS Duplas según las métricas para el cálculo de distancias inter-atómicas

El objetivo de este estudio es analizar la variabilidad de los índices 3D QuBiLs MIDAS Duplas acorde a las diferentes métricas usadas para el cálculo de distancias inter-atómicas. Con este fin fueron calculadas 325 variables para cada una de las métricas. La Figura 4.2 revela que las mejores distribuciones de entropía son obtenidas por los descriptores moleculares calculados acorde a las métricas de Canberra, Lance-Williams y Clark. Este grupo de métricas es seguido por descriptores moleculares basados en Separación Angular y Bhattacharyya respectivamente. Posteriormente aparecen teniendo un comportamiento semejante todos los índices basados en la definición de Minkowski (incluyendo Euclidian $p = 2$), lo cual sugiere que independientemente del valor que tome p la variabilidad de los mismos no cambia significativamente. Finalmente se tiene que las peores distribuciones y también similares en comportamiento, pertenecen a los índices basados en las métricas de Soergel y Wave-Edges. Este estudio demuestra que la incorporación de otras métricas (tradicionalmente utilizadas como medidas de similitud) en el cálculo de la distancia inter-atómica, en lugar de utilizar únicamente la distancia Euclidiana, contribuye a la obtención de índices moleculares con buena variabilidad.

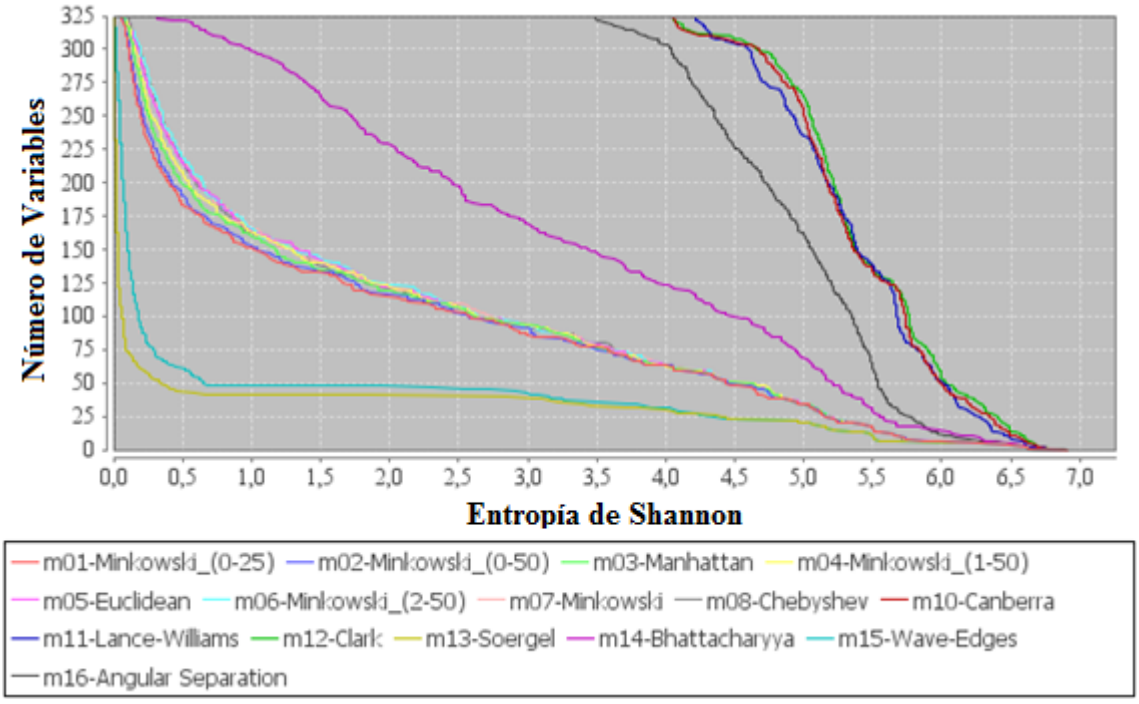

Figura 4.2: Distribución de la entropía de Shannon de los índices 3D según las métricas de distancias inter-atómicas

4.1.3. Análisis comparativo de los índices QuBiLs MIDAS Nuplas según la medida utilizada

El objetivo de este estudio es analizar el comportamiento de los índices ternarios y cuaternarios acorde a la medida utilizada. Con este fin fueron calculadas 130 variables para cada una de las medidas basadas en relaciones de ternas de átomos. La Figura 4.3 muestra que todas las distribuciones superan los 8.0 bits (80 % de la máxima entropía), siendo la más representada la medida Ángulo entre lados (m27) y constituyendo también la de mejor patrón de distribución. Seguidamente aparecen un grupo de índices con comportamiento comparable que son los basados en las medidas de Ángulo entre lados total (m28), Perímetro [m19, m20 (total)] y Suma de lados [m25, m26 (total)]. Finalmente, los índices basados en las medidas de Área del triángulo [m21, m22 (total)] constituyen los de peor comportamiento, tal vez motivado por ser el área una medida isométrica. Por lo tanto se puede concluir, que descriptores moleculares con buena variabilidad pueden ser obtenidos utilizando las medidas ternarias propuestas, exceptuando las relacionadas con el Área del triángulo.

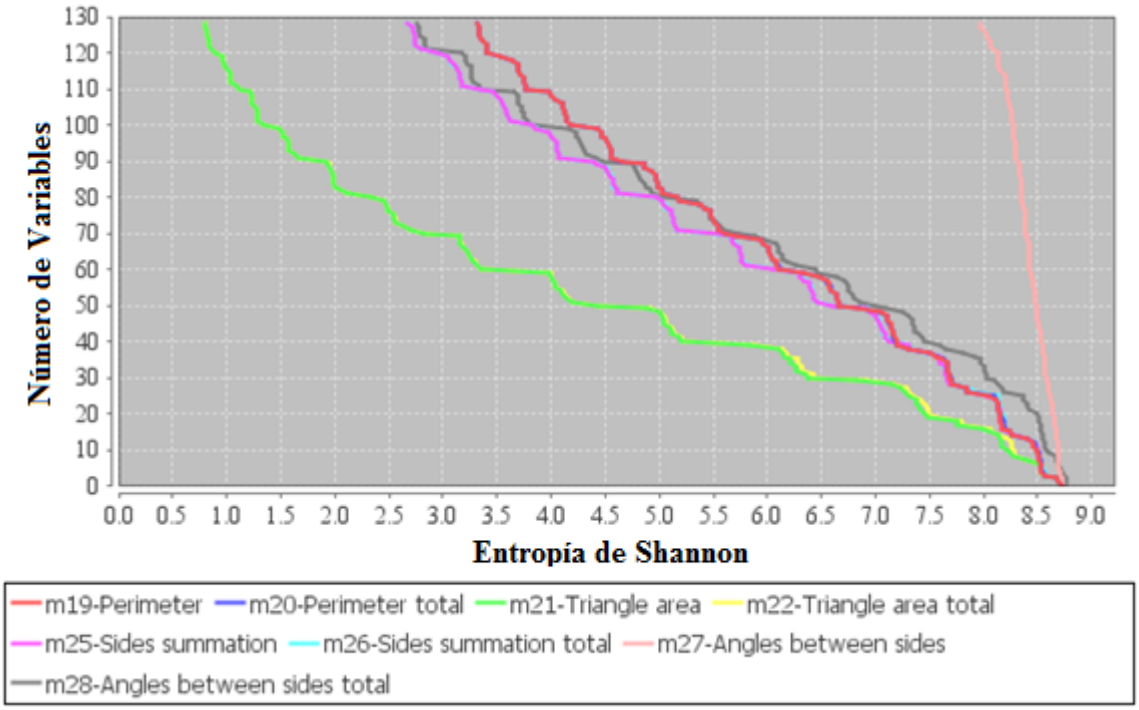

Figura 4.3: Distribución de la entropía de Shannon de los índices 3D ternarios acorde a la medida utilizada

Por otro lado, para el análisis de las medidas basadas en relaciones de cuaternas de átomos fueron seleccionadas las mejores 100 variables de cada una. En la Figura 4.4 puede apreciarse que todas las

distribuciones superan los 7.5 bits (75 % de la máxima entropía), constituyendo las de mayor variabilidad y mejor comportamiento las basadas en las medidas de Ángulo diedro [m29, m30 (total)]. Posteriormente aparecen un conjunto de índices con patrones de distribución similares y comportamiento comparable, que son los calculados a partir de las medidas relacionadas con Perímetro [m19, m20 (total)] y Suma de lados [m25, m26 (total)]. Finalmente los peores índices en cuanto a variabilidad son los correspondientes a las medidas de Volumen [m23, m24 (total)]. Exceptuando estas dos últimas, se puede concluir que descriptores moleculares con buena variabilidad pueden ser obtenidos a partir de las medidas propuestas basadas en relaciones cuaternarias entre átomos.

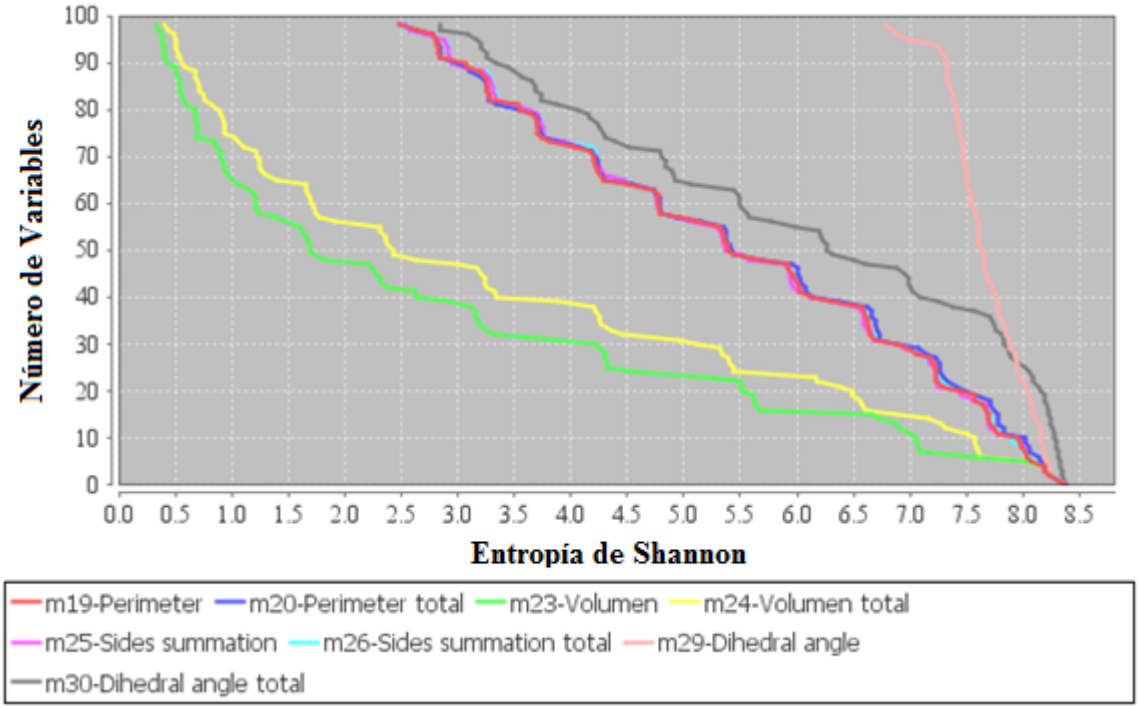

Figura 4.4: Distribución de la entropía de Shannon de los índices 3D cuaternarios acorde a la medida utilizada

4.1.4. Análisis comparativo de los índices QuBiLs MIDAS según el operador de agregación utilizado

El propósito del presente estudio es analizar la variabilidad de los índices 3D QuBiLs MIDAS acorde al operador matemático aplicado al vector de LOVIs determinado para cada una de las moléculas. Este estudio es realizado en dos partes y para ello fueron calculadas 325 variables para cada operador. Primero es analizado el comportamiento de los operadores de agregación denominados como normas, estadísticos

de tendencia central, y estadísticos de dispersión y forma; y en el segundo estudio son comparados los "algoritmos clásicos".

Como puede observarse en la Figura 4.5, los mejores índices desde el punto de vista de variabilidad son los basados en los estadísticos Skewness, Kurtosis y Coeficiente de Variación, presentando por encima de los 6 bits (60.42 % de la entropía máxima) el 96.30 %, el 95.38 % y el 60.61 % del total de variables respectivamente. Los índices de menor variabilidad son los basados en el operador Varianza. El resto de las distribuciones muestran resultados comparables.

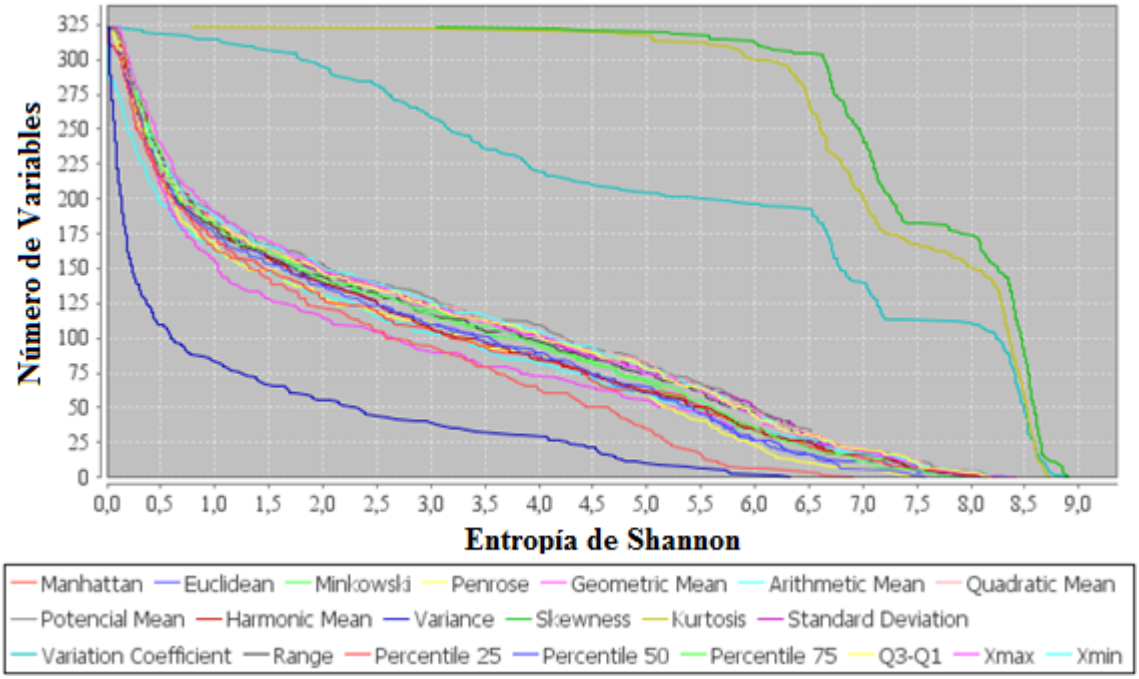

Figura 4.5: Distribución de la entropía de Shannon de los índices QuBiLs MIDAS acorde a los operadores de agregación de norma, estadístico de tendencia central, y estadístico de dispersión y forma aplicados

En el segundo estudio (ver Figura 4.6) puede observarse que los índices basados en Contenido de Información Media (MIC) y Contenido de Información Total (TIC) presentan una distribución similar y contienen muchas variables con un mismo valor de entropía no cercano a la máxima posible, contando solamente con valores superiores a los 5.5 bits (55.38 % de la entropía máxima), el 15 % de las variables analizadas en el caso de MIC y el 7 % en el caso de TIC. Otros índices con valores de entropía mayores a los 5.5 bits son los basados en el Contenido de Información Estandarizado (SIC), con el 16.61 % de sus variables por encima de este valor; y los índices con el mejor patrón de distribución son los basados en Suma Total con el 20 % del total de variables por encima del valor de entropía anteriormente mencionado.

Los índices basados en los algoritmos de Estado Electrotopológico, Gravitacional y Autocorrelación poseen distribuciones comparables, con las últimas dos presentando idéntico patrón de distribución justificado por la similar definición matemática para estos algoritmos. El comportamiento más bajo es presentado por el algoritmo Ivanciuc-Balaban.

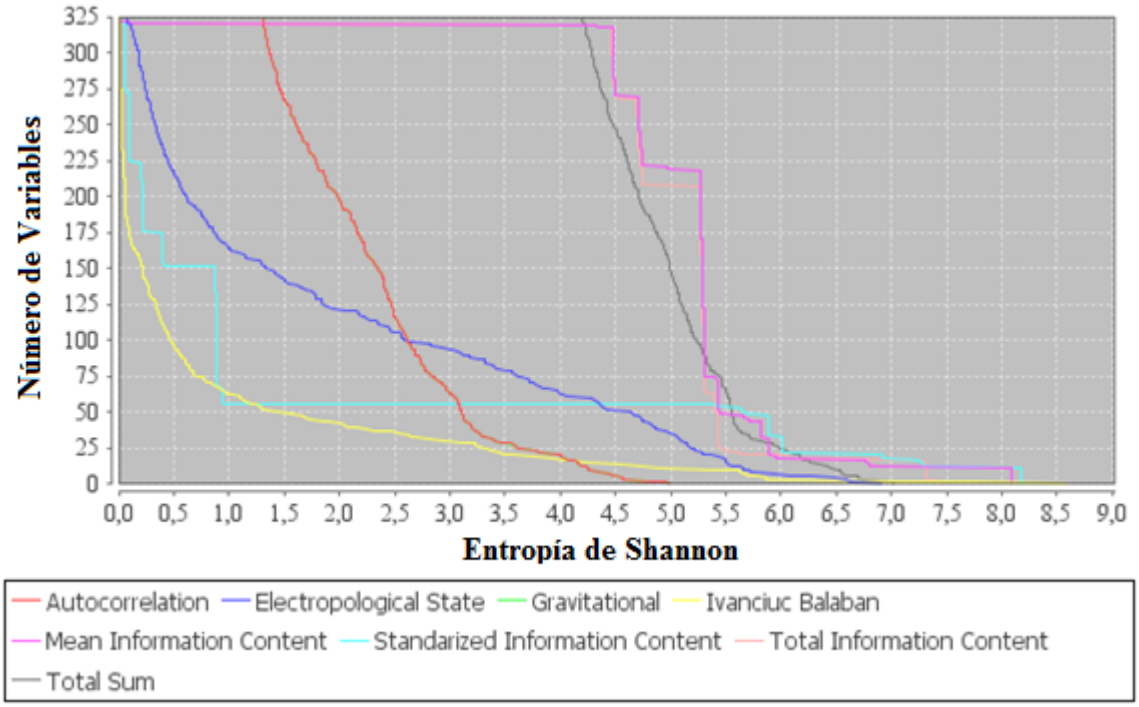

Figura 4.6: Distribución de la entropía de Shannon para los índices QuBiLs MIDAS según los algoritmos clásicos utilizados como operadores de agregación

De estos dos análisis previamente realizados se puede concluir que descriptores moleculares con comportamiento de variabilidad aceptable pueden ser obtenidos con los operadores Skewness, Kurtosis, Coeficiente de Variación, Contenido de Información Media, Contenido de Información Estandarizada y Suma Total; lo cual justifica la utilización de otras definiciones matemáticas diferente a la combinación lineal de las contribuciones atómicas.

4.1.5. Análisis comparativo de los índices QuBiLs MIDAS respecto a los descriptores calculados por otras aplicaciones

El objetivo de este estudio es comparar la variabilidad de los índices determinados por el software QuBiLs MIDAS respecto a los descriptores calculados por algunos programas relevantes usados en estudios quimio-informáticos, tales como: DRAGON [110], Mold2 [111], Padel [112], CDK [104], BlueDesc [113] y

PowerMV [114]. En este estudio los índices del DRAGON son analizados considerándolos a todos, y también teniendo en cuenta solamente los 3D. El número de corte del presente experimento es de 170 variables y fue determinado por el software BlueDesc, al ser el que menor cantidad de índices determina. Se puede apreciar en la Figura 4.7 que los índices del software QuBiLs MIDAS demuestran mejor comportamiento que los descriptores calculados por los otros software analizados. Es válido destacar también que los índices QuBiLs MIDAS Nuplas tienen un ligero mejor comportamiento respecto a los índices QuBiLs MIDAS Duplas, presentando los primeros por encima de los 9 bits el 100 % variables estudiadas respecto a un 33.5 % (57 variables) de los segundos.

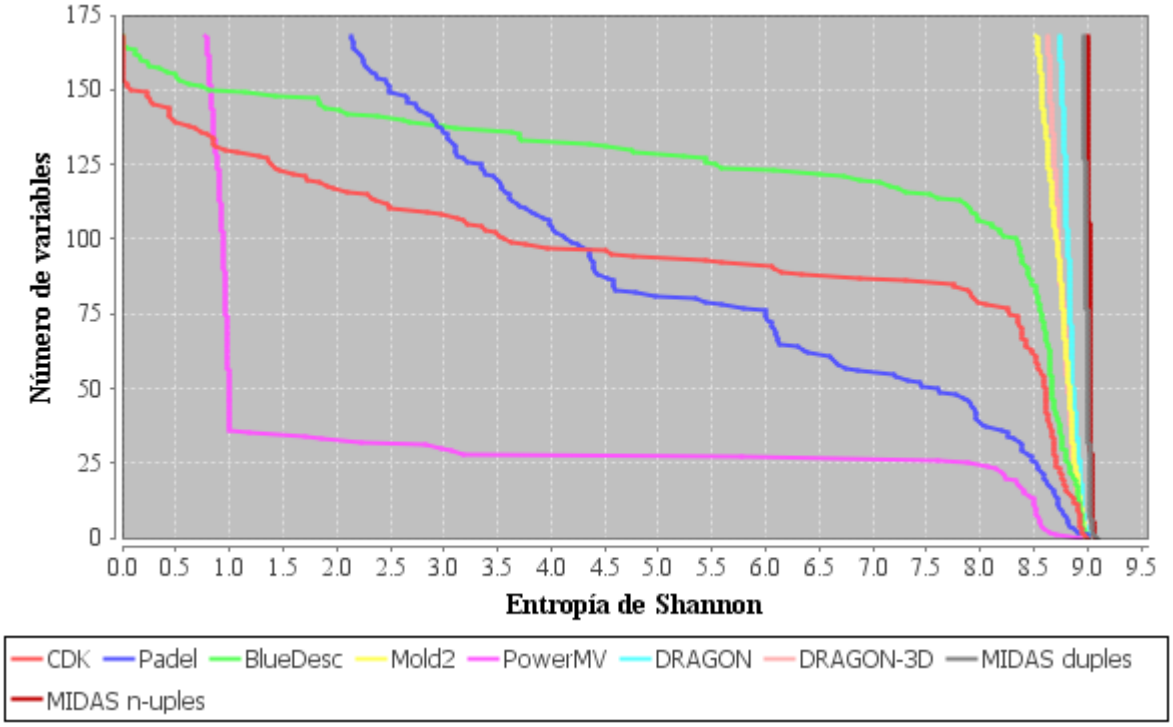

Figura 4.7: Distribución de la entropía de Shannon para los índices QuBiLs MIDAS Duplas y Nuplas, y los descriptores calculados por otras aplicaciones

Como conclusión del estudio de variabilidad realizado se puede decir que con los diferentes enfoques matriciales, la introducción de nuevas métricas para el cálculo de distancias entre pares de átomos, la generalización del esquema aditivo de las contribuciones atómicas, y la consideración de relaciones n-dimensionales entre los núcleos atómicos, contribuyen a que los índices previamente definidos alcancen un comportamiento de variabilidad bueno; y por lo tanto sugiere que el enfoque QuBiLs MIDAS sea una herramienta adecuada para estudios quimio-informáticos. Sin embargo, el método basado en entropía de

Shannon no brinda información sobre la existencia de redundancia (correlación) entre las variables, por lo que la sección siguiente se dedicará a la exploración de posible independencia lineal de los parámetros propuestos.

4.2. Análisis de ortogonalidad de los índices QuBiLs MIDAS

En esta sección, se examina la posible ortogonalidad de los índices 3D QuBiLs MIDAS usando el método de componentes principales [77,78]. La existencia de independencia lineal ha sido declarada por Randić como uno de los atributos deseables para nuevos índices moleculares [3, 115]. A continuación son mostrados los resultados de los análisis de los diferentes enfoques utilizados para el cálculo de los índices QuBiLs MIDAS. Se finaliza esta sección con una comparación entre los índices QuBiLs MIDAS Duplas e índices QuBiLs MIDAS Nuplas respecto a los descriptores 3D del DRAGON.

4.2.1. Independencia lineal de los índices QuBiLs MIDAS según el enfoque matricial

Con el fin de evaluarse la información capturada por los índices 3D propuestos según los enfoques de matriz no estocástica (NS), simple estocástica (SS), doble estocástica (DS) y de probabilidad mutua (MP), fueron calculadas 208 variables. La Tabla C.1 muestra los valores propios y los porcientos de la varianza explicada por los 5 componentes del análisis, los cuales explican aproximadamente el 84.26 % de la varianza acumulada. Analizando los componentes principales se obtiene que el enfoque de matriz de MP capta información ortogonal respecto al resto de los formalismos, encontrándose sus valores exclusivamente cargados en el Factor 5 (6.26 %). Por otra parte, el enfoque de matriz NS para órdenes superiores a 4, codifica información diferente respecto a las otras propuestas matriciales, teniendo sus valores únicamente cargados en el Factor 4 (9.591 %). Finalmente, las variables basadas en la matriz SS y DS son linealmente dependientes, presentando sus valores correlacionados en el Factor 1 (30.655 %). En este último Factor se encuentran también, los índices NS para órdenes entre 0 y 3.

4.2.2. Independencia lineal de los índices QuBiLs MIDAS Duplas según las métricas para el cálculo de distancias inter-atómicas

En este estudio se evalúa la información capturada por los índices 3D respecto a la métrica utilizada para el cálculo de distancias inter-atómicas. Con este fin fueron calculadas 600 variables. La Tabla C.2 muestra los valores propios y los porcientos de la varianza explicada por los 7 componentes del análisis, los cuales explican

aproximadamente el 97.38 % de la varianza acumulada. Analizando los componentes principales se obtiene que en el Factor 1 (57.311 %) están cargadas las variables de las métricas basadas en Minkowski para órdenes de 0 - 3, todos los índices calculados con las métricas de Canberra (M10), Lance-Williams (M11) y Clark (M12), así como los valores basados en las métricas Bhattacharyya (M14) y Separación Angular (M16) para órdenes de 0 - 5 y de 0 - 2 respectivamente. Por otra parte, para órdenes superiores o iguales a 5 las métricas basadas en Minkowski M1 ($p = 0{,}25$) y M2 ($p = 0{,}5$) tienen sus valores cargados en el Factor 6 (3.822 %), mientras que las métricas desde M3 - M8 ($p = 1$, $p = 1{,}5$, $p = 2$, $p = 2{,}5$, $p = 3$, $p = \infty\, respectivamente$) se encuentran en el Factor 2 (18.037 %). Los valores de la métrica Bhattacharyya para órdenes de 7 - 9 están cargados en el Factor 7 (1.849 %), y los de la métrica Separación Angular para órdenes mayores o iguales a 3 están correlacionadas en el Factor 5 (4.814 %). Por último, los índices basados en las métricas de Soergel (M13) y Wave-Edges (M15) están exclusivamente cargados en el Factor 3 (6.001 %) y Factor 4 (5.546 %) respectivamente. Como conclusión del estudio se puede decir 1) que existe colinealidad entre las métricas de Minkowski (para órdenes bajos), Canberra, Lance-Williams y Clark; 2) que para órdenes elevados (>=7) las métricas de Minkowski, Bhattacharyya y Separación Angular codifican información diferente; y 3) que los enfoques de Soergel y Wave-Edges captan información ortogonal respecto a los demás.

4.2.3. Independencia lineal de los índices QuBiLs MIDAS Nuplas según la medida utilizada

Para evaluar la información capturada por los índices ternarios y cuaternarios acorde a la medida utilizada, fueron calculadas 1040 variables para cada una de las medidas. Para las relaciones ternarias y cuaternarias, en la Tabla C.3 y C.4 se muestran los valores propios y los porcientos de la varianza explicada por los componentes obtenidos, respectivamente.

Para los índices QuBiLs MIDAS Ternas, los 5 componentes determinados explican aproximadamente el 92.38 % de la varianza acumulada. Realizando un análisis de los mismos se obtiene que los índices basados en las medidas de Perímetro, Suma de lados (desigualdad triangular) y Ángulo presentan sus valores fuertemente cargados en los Factores 1 (62.86 %), 2 (12.18 %) y 4 (4.86 %). Por otra parte los índices basados en la medida de Área del triángulo para órdenes menores o iguales a 3 son cargados en el Factor 1, mientras que para órdenes superiores (>=5) son cargados en los Factores 3 (10.68 %) o 5 (1.78 %) exclusivamente, en dependencia de la ponderación que se utilice. Por lo tanto se puede concluir 1) que los índices basados en todas las medidas, excepto la del área del triángulo, están muy correlacionados; y 2) que para órdenes superiores a 5 y utilizando la medida de área del triángulo se pueden obtener índices que

capten información ortogonal respecto a los otros enfoques.

En el caso de los índices QuBiLs MIDAS Cuaternas, es explicado aproximadamente el 90.49 % de la varianza acumulada por los 4 componentes obtenidos. Después de un análisis de los mismos se determina que los índices basados en las medidas de Perímetro, Suma de lados y Ángulo diedro presentan sus valores fuertemente cargados en los Factores 1 (58.96 %), 2 (16.14 %) y 4 (4.32 %). Por otra parte, los índices basados en la medida de volumen para órdenes menores o iguales a 3 son cargados en el Factor 2, mientras que para órdenes superiores a 4 son cargados en el Factor 3 (11.06 %) exclusivamente. Por lo tanto se puede concluir 1) que los índices basados en todas las medidas, excepto la de volumen, son muy colineales; y 2) que para órdenes superiores a 5 y utilizando la medida de volumen se pueden obtener índices que codifiquen información diferente a los otros enfoques.

4.2.4. Independencia lineal de los índices QuBiLs MIDAS según el operador de agregación utilizado

El presente estudio tiene como propósito analizar la ortogonalidad de los índices 3D propuestos acorde al operador matemático utilizado para el cálculo de los descriptores moleculares. Primeramente es analizada la independencia lineal de los operadores denominados como normas, estadísticos de tendencia central y estadísticos de dispersión y forma; y en segundo lugar son analizados los "algoritmos clásicos". En la Tablas C.5 y C.6 se muestran los valores propios y los porcientos de la varianza explicada por los componentes determinados, para ambos estudios respectivamente.

Para la realización del primer estudio fueron calculadas 271 variables. Los 4 componentes obtenidos explican aproximadamente el 94.49 % de la varianza acumulada. Analizando estos componentes se obtiene que todos los operadores, excepto Skewness (S), Kurtosis (K) y Coeficiente de Variación (VC), son colineales entre sí, presentando sus valores cargados en el Factor 1 (64.327 %) para órdenes superiores a 4 y en el Factor 2 (18.977 %) para órdenes de 0 - 3. Por otra parte, los índices determinados con S y K están correlacionados en el Factor 3 (8.943 %), y en este mismo Factor también están cargados los relacionados con el operador VC pero para órdenes entre 8 y 12. Finalmente, las variables basadas en el último operador anteriormente mencionado, para órdenes menores o iguales a 7, se encuentran cargadas en el Factor 4 (2.248 %).

Para el segundo estudio fueron calculadas 228 variables. El 95.17 % de la varianza acumulada es explicada por los 6 componentes determinados. Después de analizados dichos componentes se obtiene que las variables calculadas con los operadores Autocorrelación, Gravitacional, Suma Total y Estado Electrotopológico están cargados en el Factor 1 (48.942 %) para órdenes de 0 - 4, y en el Factor 2 (18.787 %) para órdenes mayores

o iguales a 5. En el Factor 1 también se encuentran los índices determinados con el operador Contenido de Información Total. Por otra parte, los valores basados en Ivanciuc-Balaban y Kier-Hall se encuentran correlacionados en el Factor 3 (17.428 %), aunque en el caso de Kier-Hall para orden 0 están cargados en el Factor 4 (5.672 %). Por último, las variables relacionadas con los operadores Contenido de Información Estandarizada y Contenido de Información Media, se encuentran exclusivamente cargadas en los Factores 5 (2.508 %) y 6 (1.834 %) respectivamente.

Como conclusión se puede decir que información ortogonal puede ser codificada haciendo uso de los operadores de agregación Manhattan, Coeficiente de Variación, Suma Total, Ivanciuc-Balaban, Contenido de Información Estandarizada y Contenido de Información Media; o algún operador linealmente dependiente con los anteriormente mencionados.

4.2.5. Independencia lineal de los índices QuBiLs MIDAS respecto a los descriptores 3D del DRAGON

Para la realización de los siguientes estudios fueron calculados con el software DRAGON 721 descriptores 3D. Las familias del DRAGON consideradas son GETAWAY con 197 variables, 3D-MoRSE con 160, RDF con 150, WHIM con 99, y Randic Geometrical Profiles y Geometrical Descriptors con 109 índices.

4.2.5.1. Independencia lineal de los índices QuBiLs MIDAS Duplas respecto a los descriptores 3D del DRAGON

Para este estudio los valores propios y los porcientos de la varianza explicada por los 15 componentes obtenidos se muestran en la Tabla C.7, los cuales explican aproximadamente el 73.24 % de la varianza acumulada. Con un análisis de los mismos se puede observar que entre los índices de ambos software existe colinealidad, presentando valores cargados en los Factores 1 (26.527 %), 2 (13.232 %) y 4 (4.505 %), y explicando aproximadamente en su conjunto el 44.26 % de la varianza acumulada. No obstante, los resultados también indican que los índices 3D del software QuBiLs MIDAS Duplas están exclusivamente cargados en los Factores 3 (5.879 %), 7 (2.784 %) y del 9 al 15 (10.596 %), por lo que explican un 19.26 % de la varianza que los índices 3D del DRAGON no hacen. Por otro lado, considerando los Factores 5 (3.979 %), 6 (3.323 %) y 8 (2.424 %) que es donde existen cargas exclusivas de los índices 3D del DRAGON, se puede plantear que los mismos explican un 9.73 % de la varianza total no explicada por los descriptores propuestos. De forma general se puede decir que los índices 3D QuBiLs MIDAS Duplas explican en su totalidad el 63.52 % de la varianza acumulada respecto al 54 % de los descriptores del DRAGON. Por lo tanto se puede concluir 1) que

los índices Duplas propuestos codifican un grado de información estructural capturada por los descriptores del DRAGON, y 2) que existe información codificada por los índices QuBiLs MIDAS Duplas linealmente independiente a los índices del DRAGON.

4.2.5.2. Independencia lineal de los índices QuBiLs MIDAS Nuplas respecto a los descriptores 3D del DRAGON

El presente estudio fue realizado en tres partes, siempre estableciendo el análisis de ortogonalidad respecto a los descriptores 3D del software DRAGON. Primero fue analizada la independencia lineal de los índices QuBiLs MIDAS Ternas, posteriormente fueron analizados los índices QuBiLs MIDAS Cuaternas, y por último se realizó un estudio complementando los índices ternarios y cuaternarios (QuBiLs MIDAS Nuplas), siendo mostrados los valores propios y los porcientos de la varianza explicada por los componentes obtenidos en las Tablas C.8, C.9 y C.10, respectivamente.

En el caso de los índices QuBiLs MIDAS Ternas, los 11 componentes determinados explican aproximadamente el 64.79 % de la varianza acumulada. Con un análisis de estos componentes se puede observar que entre los índices 3D ternarios de QuBiLs MIDAS y los índices 3D del DRAGON existe colinealidad, al presentar ambos software valores cargados en los Factores 1 (33.21 %) y 3 (4.94 %), explicando aproximadamente en su conjunto el 38.15 % de la varianza acumulada. No obstante, los resultados también indican que los índices 3D ternarios propuestos presentan cargas únicas en los Factores 2 (9.20 %), 4 (3.57 %), 5 (2.79 %) y del Factor 7 al 10 (7.06 %), explicando un 22.64 % de la varianza que los índices 3D del DRAGON no hacen. Por otro lado, puede notarse que donde existen cargas exclusivas de los descriptores 3D del DRAGON son en el Factor 6 (2.67 %) y en el Factor 11 (1.32 %), y explican un 4 % de la varianza total no realizada por los índices ternarios. De forma general se puede decir que los índices 3D basados en ternas de átomos de QuBiLs MIDAS explican en su totalidad el 60.79 % de la varianza acumulada respecto al 42.15 % de los descriptores 3D del DRAGON.

Los 13 componentes obtenidos para los índices QuBiLs MIDAS Cuaternas, explican aproximadamente el 68.07 % de la varianza acumulada. Con un análisis de dichos componentes se puede observar que existen índices QuBiLs MIDAS Cuaternas e índices 3D del DRAGON correlacionados, al presentar ambos software valores cargados en los Factores 1 (29.509 %), 5 (3.42 %) y 7 (2.479 %), explicando aproximadamente en su conjunto el 35.40 % de la varianza acumulada. Los resultados también reflejan que los índices cuaternarios propuestos presentan cargas únicas en los Factores del 2 al 4 (19.19 %), en el Factor 6 (3.15 %) y en los Factores del 8 al 12 (8.96 %), por lo que explican un 31.31 % de la varianza que los índices 3D del DRAGON

no hacen. Además, puede notarse que donde existen cargas exclusivas de los descriptores 3D del DRAGON es solamente en el Factor 13, explicando el 1.34 % de la varianza total no explicada por los índices cuaternarios. A modo de resumen se puede decir que los índices 3D basados en cuaternas de átomos de QuBiLs MIDAS explican en su totalidad el 66.72 % de la varianza acumulada respecto al 36.75 % de los descriptores 3D del software DRAGON.

Finalmente fue realizado un estudio donde se consideran los descriptores n-dimensionales (ternarios y cuaternarios), de manera tal que se pueda analizar la complementariedad que existe entre los mismos para la codificación de información de las estructuras químicas. En este análisis fueron determinados 11 componentes, los cuales explican aproximadamente el 80.28 % de la varianza acumulada. Como resultado se puede decir que en los Factores 1 (41.58 %), 4 (4.32 %), 6 (2.94 %), 8 (2.05 %) y 9 (1.77 %), existe colinealidad entre algunos índices QuBiLs MIDAS Nuplas y algunos descriptores 3D del DRAGON, explicando aproximadamente en su conjunto el 52.68 % de la varianza acumulada. Por otro lado, los resultados indican que los índices n-dimensionales propuestos presentan cargas únicas en el Factor 2 (11.01 %), Factor 3 (7.47 %), Factor 5 (3.60 %), Factor 7 (2.68 %), Factor 10 (1.52 %) y Factor 11 (1.30 %), por lo que explican un 27.60 % de la varianza que los índices 3D del DRAGON no hacen, los cuales no presentan cargas exclusivas en ninguno de los Factores analizados.

Como conclusión del estudio para los índices QuBiLs MIDAS Nuplas, puede decirse que los mismos utilizados de manera independiente (ternas y cuaternas), a pesar de capturar información similar y diferente a los índices 3D del DRAGON, deben ser utilizados con algunos descriptores del último software mencionado para abarcar el espacio geométrico posible de codificación. Sin embargo, cuando estos índices n-dimensionales son empleados de manera conjunta, contribuyen a codificar la misma información que los índices 3D del DRAGON y también información diferente no captada por estos últimos.

Aunque alta variabilidad (ver Sección 4.1) y codificación de información ortogonal respecto a descriptores existentes son cualidades deseables de los índices moleculares, no son suficientes para buenas correlaciones con una propiedad físico-química, química o biológica a ser obtenida. Por lo tanto, la próxima sección será dedicada a valorar el poder de modelación de los índices QuBiLs MIDAS propuestos.

4.3. Modelación QSAR/QSPR de un conjunto de datos de prueba

En los siguientes estudios, es evaluada la capacidad de correlación de los enfoques QuBiLs-MIDAS, como un método libre de alineamiento, siguiendo la conjetura de que los esquemas de generalización aplicados

proporcionan un espectro más amplio y flexible del espacio químico, lo cual pudiera contribuir a alcanzar buena correlación con determinadas actividades biológicas. En esta sección se muestran los resultados del comportamiento interno de los enfoques más importantes, y se culmina con una comparación de los modelos obtenidos respecto a otros métodos reportados en la literatura. Para los estudios de comparación interna, cinco variables fueron utilizadas como tamaño de los modelos.

Para todos los estudios de la presente sección, es utilizada la base de datos de Esteroides propuesta por Cramer (ver Tabla D.1) en el año 1988, usando la metodología CoMFA (del inglés - Comparative Molecular Field Analysis) [13]. Para la realización de los análisis fue empleado el software MobyDigs [116], el cual permite desarrollar modelos a través de la técnica estadística de Regresión Lineal Múltiple (RLM) haciendo uso de Algoritmos Genéticos (AG) como estrategia de selección de variables. En todos los casos los modelos fueron optimizados usando como función objetivo el parámetro estadístico Q^2_{loo} ("leave-one-out" cross validation), y todos fueron validados mediante la técnica *"bootstrapping"* $\left(Q^2_{boot}\right)$.

4.3.1. Evaluación QSAR de los índices QuBiLs MIDAS según el enfoque matricial

En esta sección es evaluado el rendimiento de los formalismos matriciales no- (NS), simple- (SS), y doble estocástico (DS) y de probabilidad mutua (MP), en la modelación QSAR. Con este fin, 260 variables para cada tipo de matriz fueron empleadas en la búsqueda del mejor modelo. Como puede ser observado en la Figura E.1, parámetros estadísticos superiores son obtenidos con el enfoque de MP $\left(Q^2_{loo} = 86{,}93\,\%,\ Q^2_{boot} = 85{,}11\,\%\right)$, seguido por los modelos basados en los índices de matriz SS ($Q^2_{loo} = 76{,}54\,\%$, $Q^2_{boot} = 73{,}09\,\%$) y NS $\left(Q^2_{loo} = 79{,}91\,\%,\ Q^2_{boot} = 71{,}69\,\%\right)$, respectivamente. Por último, aparece como el modelo de rendimiento más bajo el basado en el enfoque matricial DS $\left(Q^2_{loo} = 71{,}34\,\%,\ Q^2_{boot} = 67{,}63\,\%\right)$.

4.3.2. Evaluación QSAR de los índices QuBiLs MIDAS Duplas según las métricas para el cálculo de distancias inter-atómicas

En este estudio fueron calculadas 260 variables para cada métrica, y por cada una fue determinado el mejor modelo QSAR correspondiente. Como se evidencia en la Figura E.2, todas las métricas, como generalización de la distancia Euclidiana clásica, producen modelos con buena capacidad predictiva. Rendimientos superiores se obtienen con los modelos calculados con los índices basados en Separación Angular (m16), seguido por los correspondientes índices de Minkowski con bajo valor de p (0,25, 0,5 1, 1,5) y la métrica Lance-Williams (m11). Posteriormente aparecen, con comportamiento comparable al modelo determinado con los índices basados en la distancia Euclidiana, los relacionados con la misma definición

de Minkowski pero con valores de p mayores que dos [p = 2,5, 3, ∞ ($Chebyshev$)], y las métricas Clark (m12), Soergel (m13) y Bhattacharyya (m14) respectivamente. Finalmente, el menor poder predictivo es presentado por el modelo basado en la métrica Wave-Edges (m15).

4.3.3. Evaluación QSAR de los índices QuBiLs MIDAS Nuplas según la medida utilizada

El presente estudio fue realizado en dos partes. Primero fueron evaluados los índices QuBiLs MIDAS Ternas y luego los índices QuBiLs Cuaternas, en ambos casos, acorde a la medida de asociación ternaria y cuaternaria entre los átomos de la estructura química respectivamente. Con este fin fueron calculadas 520 variables para cada una de las medidas.

Para los índices QuBiLs MIDAS Ternas, puede ser apreciado en la Figura E.3, que los mejores resultados son alcanzados por los modelos obtenidos con las medidas basadas en Ángulo [m27, m28 (total)] y Perímetro [m19, m20 (total)], al presentar los valores más altos de Q^2_{loo} y Q^2_{boot}; seguido por el modelo basado en la medida de Área del triángulo (m21). Finalmente puede notarse, que con las medidas Área del triángulo total (m22) y las basadas en Suma de lados [m25, m26 (total)], son obtenidos modelos con pobre rendimiento, a juzgar por la diferencia existente entre los parámetros estadísticos considerados en el estudio.

En la Figura E.4, puede observarse que en el caso de los índices QuBiLs MIDAS Cuaternas, el mayor poder predictivo y robustez es alcanzado por los modelos que utilizan las medidas basadas en Ángulo diedro [m29, m30 (total)] y Suma de lados [m25, m26 (total)], al presentar los mayores valores de los parámetros estadísticos Q^2_{loo} y Q^2_{boot}; en ambos casos superiores al 79 % y 71 % de la varianza total respectivamente. Seguidamente, aparece como el mejor modelo, el basado en la medida de Volumen total (m24). Por último, los rendimientos más bajos son alcanzados por los modelos correspondientes a las medidas de Volumen (m23) y Perímetro [m19, m20 (total)].

4.3.4. Evaluación QSAR de los índices QuBiLs MIDAS según el operador de agregación utilizado

El presente análisis fue realizado en dos partes: primeramente son interpretados los resultados de los operadores de agregación denominados como normas, estadísticos de tendencia central, y estadísticos descriptivos de dispersión y forma; y en segundo lugar los resultados de los "algoritmos clásicos".

Para la realización del primer estudio fueron calculadas 260 variables por cada uno de los operadores. Como puede observarse en la Figura E.5, el modelo con mejor poder predictivo es el basado en el operador

Rango (RA), con valores de Q^2_{loo} y Q^2_{boot} superiores al 80 % de la varianza total y con poca diferencia entre ellos, lo cual indica la buena calidad de este modelo. Otros modelos con buena capacidad de predicción son los basados en los operadores i50, Desviación Estándar (SD), Media Aritmética (AM), Penrose (PN) y Manhattan (N1). Los peores modelos son los basados en los operadores Media Geométrica (GM), Percentile 25 (Q1) y Máximo (MX), los cuales presentan las mayores diferencias entre los parámetros Q^2_{loo} y Q^2_{boot} correspondientes. El resto de los modelos tienen un comportamiento comparable. De forma general se puede notar, que los operadores denominados como "normas" tienden a ser el mejor grupo, seguido por los operadores "estadísticos de dispersión" y "estadísticos de medias" respectivamente.

En este segundo estudio, según su variabilidad, fueron seleccionadas las mejores 240 variables por cada uno de los operadores. Como puede observarse en la Figura E.6, los mejores modelos fueron obtenidos con los operadores Gravitacional (GV), Estado Electrotopológico (ES), Kier-Hall (KH), Suma Total (TS) y Autocorrelación (AC). Los valores del parámetro Q^2_{loo} de los modelos anteriores se encuentran entre un 78.87 % y un 80.31 % del varianza total, mientras que los valores de la validación realizada con la función Q^2_{boot} están entre un 74.05 % y un 78.08 % de la varianza a explicar. Puede apreciarse también que la diferencia entre ambos parámetros en cada uno de los modelos precedentes no es grande, lo cual da muestra de la calidad del poder predictivo que poseen. Los peores modelos son los basados en los operadores Ivanciuc-Balaban (IB), Contenido de Información Media (MIC), Total (TIC) y Estandarizada (SIC).

A modo de conclusión se puede decir que es válida la utilización de varios enfoques matemáticos como operadores de agregación, al obtener modelos con capacidad predictiva de comparable a superior respecto a la combinación lineal de las contribuciones atómicas de una molécula. En sentido general, los operadores Rango (RA), i50, Desviación Estándar (SD), Media Aritmética (AM), Penrose (PN), Manhattan (N1), Gravitacional (GV), Estado Electrotopológico (ES), Kier-Hall (KH), Suma Total (TS) y Autocorrelación (AC), son buenos en la obtención de modelos con poder predictivo aceptable para describir propiedades inherentes de las estructuras moleculares.

4.3.5. Evaluación QSAR de los índices QuBiLs MIDAS respecto a otros enfoques reportados en la literatura

Uno de los métodos de evaluación de la verdadera contribución y relevancia de nuevos parámetros moleculares o generalización de estos, es valorar su rendimiento en estudios de correlación con determinada propiedad molecular con respecto a los métodos existentes, siguiendo el supuesto de que la novedad de un método pudiera ser medida en términos de correlaciones mejores con propiedades estructurales moleculares,

o al menos proveer mejoras cuando son combinadas con las técnicas reportadas. En este sentido, fue realizada una búsqueda de modelos QSAR para la *afinidad de acoplamiento a la cortiscosteroide-binding globulin (CBG)* para una data de 31 esteroides.

Este estudio consta de dos partes: 1) fueron creados modelos utilizando las 31 estructuras de la base de datos de Esteroides como conjunto de entrenamiento, y 2) fueron creados modelos dividiendo la base de datos en un conjunto de entrenamiento y un conjunto de prueba. En el primer experimento se obtuvieron los mejores modelos para 3, 4, 5 y 6 variables, mientras que en el segundo se obtuvieron los mejores modelos para 2, 3 y 4 variables. En ambos casos fueron reportados los mejores modelos para las relaciones duplas, ternarias y cuaternarias entre átomos; así como para la unión de las mejores variables de las dos últimas relaciones anteriores (enfoque nuplas). La selección de los modelos fue realizada acorde a la calidad de los parámetros estadísticos considerados en el estudio.

Primer experimento

En las Tablas F.1, F.2, F.3 y F.4 se muestran los modelos del primer estudio realizado con sus correspondientes parámetros estadísticos, usando los descriptores 3D QuBiLs MIDAS Duplas, Ternas, Cuaternas y Nuplas (ternarios + cuaternarios) propuestos. Se especifican en la Tabla F.1 los modelos de las relaciones de duplas de átomos, en la Tabla F.2 y en la Tabla F.3 los modelos pertenecientes a las relaciones de ternas y cuaternas de átomos respectivamente, y en la Tabla F.4 los modelos correspondientes a la definición de nuplas.

Puede notarse de manera general, que con el enfoque de duplas se obtienen mejores resultados que con los enfoques de ternas y cuaternas; y que con el enfoque de ternas se alcanzan mejores modelos que con el enfoque cuaternario. Puede observarse además, que los modelos de nuplas presentan mejor poder predictivo que los correspondientes modelos obtenidos con los enfoques QuBiLs MIDAS de manera individual.

Con el fin de comparar el desempeño de los índices 3D QuBiLs MIDAS, se comparó los resultados obtenidos con otros modelos reportados en la literatura. En la Tabla 4.1 se encuentran resumidos los resultados alcanzados, ordenados de forma decreciente acorde al valor del parámetro Q^2_{loo}. Como puede verse, todos los modelos presentados excepto los de dimensión tres, son mejores a los reportados, incluso cuando muchos de ellos utilizan técnicas más complejas. Los modelos de dimensión tres únicamente tienen menor poder predictivo que el modelo de *Descriptores Moleculares basado en Matrices de Similitud (Similarity matrices-based molecular descriptors)*, pero están definidos con una técnica más sencilla y con menor número de variables.

Segundo experimento

El hándicap de cualquier enfoque QSAR consiste en la predicción adecuada de un conjunto de compuestos que no han sido usados para construir el modelo, es el bien llamado conjunto de prueba. Este enfoque en la familia de Esteroides propuesta por Cramer, es usualmente realizado dividiendo la base de datos en dos clases: el conjunto de entrenamiento con los primeros 21 compuestos y el conjunto de prueba con los 10 restantes. Por lo tanto, solamente los 21 esteroides del conjunto de entrenamiento son usados para construir el modelo que después será aplicado al conjunto de prueba. Teniendo en cuenta este enfoque, en las Tablas F.5 (duplas), F.6 (ternas), F.7 (cuaternas) y F.8 (nuplas), se muestran los modelos obtenidos con sus correspondientes parámetros estadísticos usando los descriptores 3D QuBiLs MIDAS propuestos.

En la Tabla 4.2 son recogidos los valores de Q^2_{loo} de diferentes metodologías QSAR existentes, y que utilizan como conjunto de entrenamiento y prueba el explicado al inicio del presente experimento. También son listados en esta tabla los valores del mismo parámetro estadístico pero de los modelos obtenidos en este estudio. Por otro lado, en la Tabla 4.3, son resumidos los valores de la *desviación estándar del error en la predicción externa (SDEP^{ext})* de establecidas metodologías QSAR, como otra medida para evaluar la calidad de los modelos propuestos. En sentido general, el rendimiento de los modelos QuBiLs MIDAS son comparables acorde a los reportados en la literatura.

4.4. Conclusiones parciales

En este capítulo se presentaron los resultados del análisis de variabilidad basado en entropía de Shannon, del análisis de independencia lineal mediante la técnica estadística de componentes principales, y de la modelación QSAR de la afinidad de acoplamiento a la *cortiscosteroide-binding globulin* en una base de datos de 31 esteroides. Como conclusión de esta experimentación, se tiene que los descriptores QuBiLs MIDAS que se proponen en el presente trabajo 1) tienen un comportamiento de variabilidad superior acorde a varios índices calculados por importantes aplicaciones quimio-informáticas, como el software DRAGON, por lo que caracterizan mejor estructuras moleculares distintas estructuralmente; 2) posibilitan codificar información no redundante respecto a otros descriptores geométricos (3D) definidos hasta la fecha; y 3) contribuyen a obtener modelos con poder predictivo de comparable a superior respecto a otras metodologías reportadas, presentando también un desempeño aceptable en la predicción de propiedades de compuestos no utilizados como serie de entrenamiento.

Método QSAR	**N**	**n**	**Método estadístico**	**Q^2**	**Ref**
QuBiLs MIDAS Nuplas	31	6	GA + RLM	0.980	Ec. 16
QuBiLs MIDAS Duplas	31	6	GA + RLM	0.978	Ec. 4
QuBiLs MIDAS Ternas	31	6	GA + RLM	0.970	Ec. 8
QuBiLs MIDAS Nuplas	31	5	GA + RLM	0.969	Ec. 15
QuBiLs MIDAS Cuaternas	31	6	GA + RLM	0.968	Ec. 12
QuBiLs MIDAS Duplas	31	5	GA + RLM	0.964	Ec. 3
QuBiLs MIDAS Ternas	31	5	GA + RLM	0.963	Ec. 7
QuBiLs MIDAS Cuaternas	31	5	GA + RLM	0.956	Ec. 11
QuBiLs MIDAS Duplas	31	4	GA + RLM	0.952	Ec. 2
QuBiLs MIDAS Nuplas	31	4	GA + RLM	0.950	Ec. 14
QuBiLs MIDAS Ternas	31	4	GA + RLM	0.948	Ec. 6
QuBiLs MIDAS Cuaternas	31	4	GA + RLM	0.947	Ec. 10
Similarity matrices-based molecular descriptors	31	6	genetic NN	0.940	[26]
QuBiLs MIDAS Nuplas	31	3	GA + RLM	0.922	Ec. 13
CoSCoSA	30	3[c]	RLM + PCA	0.92	[117]
QuBiLs MIDAS Duplas	31	3	GA + RLM	0.917	Ec. 1
QuBiLs MIDAS Cuaternas	31	3	GA + RLM	0.915	Ec. 9
QuBiLs MIDAS Ternas	31	3	GA + RLM	0.909	Ec. 5
MIDSASA - template	30	2[a]	K-NN	0.88	[118]
TQSAR	31	6	RLM + PCA	0.842	[119]
HE-State/E-State	31	3[c]	-	0.803	[120]
SOM-4D-QSAR	30	4	SOM + NN + PLS	0.80	[121]
CoSA	30	3[b]	-	0.78	[122]
QSAR/E-state	30	3[d]	-	0.78	[123]
TQSI	31	3	RLM	0.775	[124]
MEDV	31	6	GA + RLM	0.765	[125]
CoMSA	31	1	Kohonen-NN + PLS	0.76	[126]
Índices Similitud (Similarity indices)	31	1	PLS	0.734	[127]
CoSASA	30	3[d]	-	0.73	[122]
MQSM	31	4	RLM después PLS	0.727	[128]
E-State and kappa shape index	31	4	RLM	0.720	[129]
MQMS	31	3	RLM + ACP	0.705	[124]
SAMFA-RF	31	-	RF	0.69	[130]
SAMFA-PLS	31	4-5	PLS	0.69	[130]
SOM-4D-QSAR	31	4	SOM + NN + PLS	0.68	[121]
Wagener's	31	-	Kohonen-NN + BP-NN	0.630	[32]
SAMFA-SVM	31	-	SVM	0.60	[130]

N: total de esteroides en el conjunto de entrenamiento
n: número de componentes en el modelo
[a]: compuestos (compounds)
[b]: bins
[c]: componentes principales
[d]: átomos

Tabla 4.1: Comparación de los modelos QuBiLs MIDAS respecto a otros enfoques 3D QSAR, considerando los 31 (o 30) esteroides como conjunto de entrenamiento

Modelo QSAR	**n**	**Q^2**	**Ref**	**Modelo QSAR**	**n**	**Q^2**	**Ref**
QuBiLs MIDAS Duplas	4	0.978	Ec. 19	CoMMA	3	0.828	[15]
QuBiLs MIDAS Ternas	4	0.976	Ec. 22	Tominaga's	3	0.807	[27]
QuBiLs MIDAS Nuplas	4	0.962	Ec. 28	CoMFA-QOVS[b]	3	0.807	[131]
QuBiLs MIDAS Ternas	3	0.957	Ec. 21	PARM	-	0.806	[132]
QuBiLs MIDAS Duplas	3	0.951	Ec. 18	CoMFA-PWPLS[a]	1	0.792	[131]
QuBiLs MIDAS Cuaternas	4	0.948	Ec. 25	Similarity indices	1	0.780	[127]
QuBiLs MIDAS Duplas	2	0.945	Ec. 17	CoMASA	2	0.822	[133]
QuBiLs MIDAS Cuaternas	3	0.924	Ec. 24	CoMFA-PWPLS[b]	3	0.773	[131]
QuBiLs MIDAS Nuplas	3	0.901	Ec. 27	Similarity matrices	2	0.761	[26]
COMPASS	-	0.890	[12]	TDQ	2	0.680	[134]
CoMFA-QOVS[a]	4	0.875	[131]	CoMSIA	4	0.665	[16]
QuBiLs MIDAS Ternas	2	0.885	Ec. 20	CoMFA	2	0.662	[13]
QuBiLs MIDAS Cuaternas	2	0.872	Ec. 23	SHAPE matrix	5	0.633	[135]
QuBiLs MIDAS Nuplas	2	0.853	Ec. 26	MS-WHIM	2	0.631	[30]
TARIS	5	0.84	[136]	ALPHA	2	0.63	[137]
TQSAR	6	0.832	[119]	SHESP matrix	1	0.533	[135]
EVA	2	0.830	[45]	ESP matrix	1	0.501	[135]

n: número de componentes en el modelo
[a]: campo estérico (steric field)
[b]: campo electrostático (electrostatic field)

Tabla 4.2: Comparación de los modelos QuBiLs MIDAS respecto a otros enfoques 3D QSAR, considerando las 21 primeras estructuras como conjunto de entrenamiento y las últimas 10 como conjunto de prueba

Modelo $QSAR^n$	**$SDEP^{ext}$**	**Ref**	**Modelo $QSAR^n$**	**$SDEP^{ext}$**	**Ref**
CoMFA-QOVS[b]	0.258	[131]	*QuBiLs MIDAS Ternas*	0.540	Ec. 21
MFTA	0.30	[138]	CoMFA-PWPLS[b]	0.545	[131]
QuBiLs MIDAS Cuaternas	0.380	Ec. 23	*QuBiLs MIDAS Cuaternas*	0.548	Ec. 24
CoMFA-QOVS[a]	0.404	[131]	EEVA	0.58	[139]
QuBiLs MIDAS Cuaternas	0.433	Ec. 25	*QuBiLs MIDAS Nuplas*	0.585	Ec. 28
QuBiLs MIDAS Nuplas	0.434	Ec. 27	SOMFA	0.585	[140]
QuBiLs MIDAS Nuplas	0.455	Ec. 26	ESP matrix	0.595	[135]
QuBiLs MIDAS Ternas	0.494	Ec. 20	MEDV	0.65	[125]
QuBiLs MIDAS Duplas	0.501	Ec. 18	MS-WHIM	0.662	[30]
QuBiLs MIDAS Duplas	0.512	Ec. 17	SHESP matrix	0.646	[135]
CoMFA-PWPLS[a]	0.522	[131]	COMSA	0.70	[126]
QuBiLs MIDAS Ternas	0.523	Ec. 22	COMPASS	0.705	[12]
EVA	0.53	[45]	PARM	0.709	[132]
QuBiLs MIDAS Duplas	0.533	Ec. 19	ALPHA	0.71	[137]
QS-SM	0.54	[141]	TQSAR	0.762	[119]

[a]: campo estérico (steric field)
[b]: campo electrostático (electrostatic field)

Tabla 4.3: Comparación de los modelos propuestos respecto a otras técnicas QSAR acorde al parámetro SDEP externo

Conclusiones

Atendiendo a los resultados obtenidos se llegó a las siguientes conclusiones:

1. Se propusieron índices moleculares geométricos basados en las formas lineal, bilineal y cuadrática, que utilizan la *matriz espacial de similitud-disimilitud* como matriz de las formas algebraicas anteriores, siendo esta matriz construida a partir de distintas métricas diferentes a la Euclidiana para el cálculo de la distancia inter-atómica.

2. Se definieron índices moleculares geométricos basados en conceptos del álgebra multilineal que consideran relaciones ternarias y cuaternarias entre los átomos de una molécula.

3. Se implementaron los algoritmos correspondientes para el cálculo de los descriptores geométricos propuestos en un software denominado ToMoCoMD-CARDD QuBiLs MIDAS, el cual contiene además las interfaces gráficas necesarias para la configuración de los índices, y tiene como una de las características más relevantes la ejecución en paralelo de los procedimientos necesarios para el cómputo de los descriptores.

4. Los diversos estudios de informática química demuestran que los índices QuBiLs MIDAS propuestos poseen variabilidad superior a la de otros índices reportados en la literatura; codifican información química ortogonal respecto a los índices 3D del software DRAGON, además de captar toda la información codificada por estos índices; y que tienen un desempeño en la modelación QSAR de comparable a superior respecto a metodologías reportadas en la literatura.

Recomendaciones

1. Evaluar el desempeño de los índices QuBiLs MIDAS propuestos con otras propiedades químico-físicas, ADME-Tox y farmacológicas para validar la utilidad y efectividad de los nuevos parámetros moleculares.

2. Extender las definiciones propuestas en el presente trabajo a la codificación de información química en macromoléculas y estructuras inorgánicas, así como en estudios de complejidad (ej. redes complejas).

3. Utilizar otras métricas y/o medidas para establecer relaciones invariantes entre n átomos de una molécula.

Referencias bibliográficas

[1] Helm GF. The Principles of Mathematical Chemistry: The Energetics of Chemical Phenomena. John Wiley and Sons; 1897.

[2] In: Applications. Wiley-VCH Verlag GmbH; 2008. .

[3] Todeschini R, Consonni V. Molecular Descriptors for Chemoinformatics. 2nd ed. WILEY - VCH; 2009.

[4] Yovani MP. Total and local (atom and atom type) molecular quadratic indices: significance interpretation, comparison to other molecular descriptors, and QSPR/QSAR applications. Bioorganic & Medicinal Chemistry. 2004;12(24):6351 – 6369.

[5] Marrero Ponce Y, Torrens F, Garcia Domenech R, Ortega Broche SE, Zaldivar VR. Novel 2D TOMOCOMD-CARDD molecular descriptors: atom-based stochastic and non-stochastic bilinear indices and their QSPR applications. Journal of Mathematical Chemistry. 2008;44(3):650–673.

[6] Marrero-Ponce Y, Martinez-Albelo ER, Casanola-Martin GM, Castillo-Garit JA, Echeveria-Diaz Y, Zaldivar VR, et al. Bond-based linear indices of the non-stochastic and stochastic edge-adjacency matrix. 1. Theory and modeling of ChemPhys properties of organic molecules. Molecular Diversity. 2010;14(4):731–753.

[7] Kier L, Hall L. An Electrotopological-State Index for Atoms in Molecules. Pharmaceutical Research. 1990;7(8):801–807.

[8] Katritzky AR, Kuanar M, Slavov S, Hall CD, Karelson M, Kahn I, et al. Quantitative correlation of physical and chemical properties with chemical structure: utility for prediction. Chemical reviews. 2010;110(10):5714–5789.

[9] Hansch C, Maloney PP, Fujita T, Muir RM. Correlation of biological activity of phenoxyacetic acids with Hammett substituent constants and partition coefficients. 1962;.

[10] Kubinyi H. QSAR and 3D QSAR in drug design Part 1: methodology. Drug Discovery Today. 1997;2(11):457 – 467.

[11] Sutherland JJ, O'Brien LA, Weaver DF. A Comparison of Methods for Modeling Quantitative Structure-Activity Relationships. Journal of Medicinal Chemistry. 2004;47(22):5541–5554.

[12] Sun H. COMPASS: An ab Initio Force-Field Optimized for Condensed-Phase ApplicationsOverview with Details on Alkane and Benzene Compounds. The Journal of Physical Chemistry B. 1998;102(38):7338–7364.

[13] Cramer RD, Patterson DE, Bunce JD. Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. Journal of the American Chemical Society. 1988;110(18):5959–5967.

[14] Chuman H, Karasawa M, Fujita T. A Novel Three-Dimensional QSAR Procedure: Voronoi Field Analysis. Quantitative Structure-Activity Relationships. 1998;17(04):313–326.

[15] Silverman BD, Platt D, Pitman M, Rigoutsos I. Comparative molecular moment analysis (CoMMA). Perspectives in Drug Discovery and Design. 1998;12-14(0):183–196.

[16] Klebe G. Comparative molecular similarity indices analysis: CoMSIA. Perspectives in Drug Discovery and Design. 1998;12-14(0):87–104.

[17] Pastor M, Cruciani G, McLay I, Pickett S, Clementi S. GRid-INdependent Descriptors (GRIND): A Novel Class of Alignment-Independent Three-Dimensional Molecular Descriptors. Journal of Medicinal Chemistry. 2000;43(17):3233–3243.

[18] Marrero-Ponce Y, Torrens F, Alvarado YJ, Rotondo R. Bond-based global and local (bond, group and bond-type) quadratic indices and their applications to computer-aided molecular design. 1. QSPR studies of diverse sets of organic chemicals. Journal of Computer-Aided Molecular Design. 2006;20(10-11):685–701.

[19] Ponce YM. Total and Local Quadratic Indices of the Molecular Pseudograph's Atom Adjacency Matrix: Applications to the Prediction of Physical Properties of Organic Compounds. Molecules. 2003;8(9):687–726.

[20] Marrero-Ponce Y, Huesca-Guillen A, Ibarra-Velarde F. Quadratic indices of the molecular pseudograph atom adjacency matrix and their stochastic forms: a novel approach for virtual screening and in silico discovery of new lead paramphistomicide drugs-like compounds. Journal of Molecular Structure. 2005;717:67–79.

[21] Garit JAC, Ponce YM, Torrens F. Atom-based 3D-chiral quadratic indices. Part 2: Prediction of the corticosteroid-binding globulinbinding affinity of the 31 benchmark steroids data set. Bioorganic & Medicinal Chemistry. 2006;14(7):2398 – 2408.

[22] Castillo JA Garit, Martinez O Santiago, Marrero Y Ponce, Casanola GM Martin, Torrens F. Atom-based non-stochastic and stochastic bilinear indices: Application to QSPR/QSAR studies of organic compounds. Chemical Physics Letters. 2008;464(1):107–112.

[23] Marrero-Ponce Y, Castillo-Garit JA, Castro E, Torrens F, Rotondo R. 3D-chiral (2.5) atom-based TOMOCOMD-CARDD descriptors: theory and QSAR applications to central chirality codification. Journal of Mathematical Chemistry. 2008;44(3):755–786.

[24] Hopfinger A, Tokarski J. Practical applications of computer-aided drug design. Practical Applications of Computer-Aided Design (Charifson PS, ed) New York: Marcel Dekker. 1997;105:164.

[25] Parretti MF, Kroemer RT, Rothman JH, Richards WG. Alignment of molecules by the Monte Carlo optimization of molecular similarity indices. Journal of Computational Chemistry. 1997;18(11):1344–1353.

[26] So SS, Karplus M. Three-Dimensional Quantitative Structure-Activity Relationships from Molecular Similarity Matrices and Genetic Neural Networks. 1. Method and Validations. Journal of Medicinal Chemistry. 1997;40(26):4347–4359.

[27] Tominaga Y, Fujiwara I. Novel 3D Descriptors Using Excluded Volume: Application to 3D Quantitative Structure-Activity Relationships. Journal of Chemical Information and Computer Sciences. 1997;37(6):1158–1161.

[28] Goodford PJ. A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. Journal of Medicinal Chemistry. 1985;28(7):849–857.

[29] Construction of Voronoi polyhedra. Journal of Computational Physics. 1978;29(1):81 – 92.

[30] Bravi G, Gancia E, Mascagni P, Pegna M, Todeschini R, Zaliani A. MS-WHIM, new 3D theoretical descriptors derived from molecular surface properties: A comparative 3D QSAR study in a series of steroids. Journal of Computer-Aided Molecular Design. 1997;11(1):79–92.

[31] Todeschini R, Gramatica P. New 3D Molecular Descriptors: The WHIM theory and QSAR Applications. In: 3D QSAR in Drug Design. vol. 2 of Three-Dimensional Quantitative Structure Activity Relationships. Springer Netherlands; 2002. p. 355–380.

[32] Wagener M, Sadowski J, Gasteiger J. Autocorrelation of Molecular Surface Properties for Modeling Corticosteroid Binding Globulin and Cytosolic Ah Receptor Activity by Neural Networks. Journal of the American Chemical Society. 1995;117(29):7769–7775.

[33] Gasteiger J, Sadowski J, Schuur J, Selzer P, Steinhauer L, Steinhauer V. Chemical Information in 3D Space. Journal of Chemical Information and Computer Sciences. 1996;36(5):1030–1037.

[34] Consonni V, Todeschini R, Pavan M. Structure/Response Correlations and Similarity/Diversity Analysis by GETAWAY Descriptors. 1. Theory of the Novel 3D Molecular Descriptors. Journal of Chemical Information and Computer Sciences. 2002;42(3):682–692.

[35] Bogdanov B, Nikolic S, Trinajstic N. On the three-dimensional wiener number. Journal of Mathematical Chemistry. 1989;3(3):299–309.

[36] Mihalic Z, Nikolic S, Trinajstic N. Comparative study of molecular descriptors derived from the distance matrix. Journal of Chemical Information and Computer Sciences. 1992;32(1):28–37.

[37] Karelson M, Lobanov VS, Katritzky AR. Quantum-Chemical Descriptors in QSAR/QSPR Studies. Chemical Reviews. 1996;96(3):1027–1044.

[38] Bath PA, Poirrette AR, Willett P, Allen FH. The Extent of the Relationship between the Graph-Theoretical and the Geometrical Shape Coefficients of Chemical Compounds. Journal of Chemical Information and Computer Sciences. 1995;35(4):714–716.

[39] Trinajstić N, Nikolić S, Lucić B, Amić D, Mihalić Z. The Detour Matrix in Chemistry. Journal of Chemical Information and Computer Sciences. 1997;37(4):631–638.

[40] Randic M. On Characterization of Cyclic Structures. Journal of Chemical Information and Computer Sciences. 1997;37(6):1063–1071.

[41] Ivanciuc O. QSAR and QSPR molecular descriptors computed from the resistance distance and electrical conductance matrices. ACH MODELS IN CHEMISTRY. 2000;137(5/6):607–632.

[42] Karelson M, Lobanov VS, Katritzky AR. Quantum-Chemical Descriptors in QSAR/QSPR Studies. Chem Rev. 1996 Jan;96(3):1027–1044.

[43] Cruciani G. Molecular interaction fields: applications in drug discovery and ADME prediction. vol. 1. Vch Verlagsgesellschaft Mbh; 2006.

[44] Cruciani G, Pastor M, Guba W. VolSurf: a new tool for the pharmacokinetic optimization of lead compounds. European Journal of Pharmaceutical Sciences. 2000;11, Supplement 2(0):S29 – S39.

[45] Turner D, Willett P, Ferguson A, Heritage T. Evaluation of a novel molecular vibration-based descriptor (EVA) for QSAR studies: 2. Model validation using a benchmark steroid dataset. Journal of Computer-Aided Molecular Design. 1999;13(3):271–296.

[46] Bajzer Z, Randic M, Plavsic D, Basak SC, et al. Novel map descriptors for characterization of toxic effects in proteomics maps. Journal of molecular graphics & modelling. 2003;22(1):1.

[47] Randic M, Kleiner AF, De Alba LM. Distance/Distance Matrixes. Journal of Chemical Information and Computer Sciences. 1994;34(2):277–286.

[48] Deza MM, Deza E. Dictionary of distances. Elsevier Science; 2006.

[49] Vargas-Quesada B, de Moya Anegón F. Visualizing the structure of science. Springer; 2007.

[50] Willett P. Chemoinformatics-similarity and diversity in chemical libraries. Current opinion in biotechnology. 2000;11(1):85.

[51] Balaban AT, Feroiu V. Correlations between structure and critical data or vapor pressures of alkanes by means of topological indices. Rep Mol Theor. 1990;1:133–139.

[52] Balaban AT, Bertelsen S, Basak SC. New centric topological indexes for acyclic molecules (trees) and substituents (rooted trees), and coding of rooted trees. MATCH Commun Math Comput. 1994;30:55–72.

[53] Lance G, Williams W. Mixed-Data Classificatory Programs I - Agglomerative Systems. Australian Computer Journal. 1967;1(1):15–20.

[54] Jurman G, Merler S, Barla A, Paoli S, Galea A, Furlanello C. Algebraic stability indicators for ranked lists in molecular profiling. Bioinformatics. 2008;24(2):258–264.

[55] Emran SM, Ye N. Robustness of Chi-square and Canberra distance metrics for computer intrusion detection. Quality and Reliability Engineering International. 2002;18(1):19–28.

[56] Derpanis KG. The Bhattacharyya Measure; 2008.

[57] Choi E, Lee C. Feature extraction based on the Bhattacharyya distance. Pattern Recognition. 2003;36(8):1703 – 1709.

[58] Minchin P. An evaluation of the relative robustness of techniques for ecological ordination. In: Prentice IC, Maarel E, editors. Theory and models in vegetation science. vol. 8 of Advances in vegetation science. Springer Netherlands; 1987. p. 89–107.

[59] Clarke KR, Somerfield PJ, Chapman MG. On resemblance measures for ecological studies, including taxonomic dissimilarities and a zero adjusted Bray Curtis coefficient for denuded assemblages. Journal of Experimental Marine Biology and Ecology. 2006;330(1):55 – 80.

[60] Frakes WB, Baeza-Yates R. Information retrieval: data structures & algorithms, 1992. Prentice-Hall; 1992.

[61] Singhal A. Modern information retrieval: A brief overview. IEEE Data Engineering Bulletin. 2001;24(4):35–43.

[62] Gonzalez-Diaz H, Uriarte E. Proteins QSAR with Markov average electrostatic potentials. Bioorganic & Medicinal Chemistry Letters. 2005;15(22):5088 – 5094.

[63] Ramos de Armas R, Gonzalez Diaz H, Molina R, Uriarte E. Markovian Backbone Negentropies: Molecular descriptors for protein research. I. Predicting protein stability in Arc repressor mutants. Proteins: Structure, Function, and Bioinformatics. 2004;56(4):715–723.

[64] Zefirov N, Kirpichenok M, Izmailov F, Trofimov M. Scheme for the Calculation of the Electronegativities of Atoms in a Molecule in the Framework of Sanderson Principle. In: Dokl. Akad. Nauk SSSR. vol. 296; 1987. p. 883–887.

[65] Carbo-Dorca R. Stochastic transformation of quantum similarity matrices and their use in quantum QSAR (QQSAR) models. International Journal of Quantum Chemistry. 2000;79(3):163–177.

[66] Edwards CH, Penney DE. Elementary linear algebra. Prentice Hall: Englewoods Cliffs; 1988.

[67] Devillers J, Karcher W. Applied multivariate analysis in SAR and environmental studies. Kluwer Academic Publishers for the European Communities: Dordrecht; 1991.

[68] Frank IE, Todeschini R. The Data Analysis Handbook. Elsevier; 1994.

[69] Massart DL, Vandeginste BGM, Buydens L, De Jong S, Lewi PJ, Smeyers-Verbeke J. Handbook of Chemometrics and Qualimetrics. Part A. Elsevier; 1997.

[70] Massart DL, Vandeginste BGM, Buydens L, De Jong S, Lewi PJ, Smeyers-Verbeke J. Handbook of Chemometrics and Qualimetrics. Part B. Elsevier; 1998.

[71] Rumelhart DE, Hinton GE, Williams RJ. Learning representations by back-propagating errors. Nature. 1986 Oct;323(6088):533–536.

[72] Vapnik V. The Nature of Statistical Learning Theory. Springer; 1995.

[73] Godden JW, Stahura FL, Bajorath J. Variability of Molecular Descriptors in Compound Databases Revealed by Shannon Entropy Calculations. Journal of Chemical Information and Computer Sciences. 2000;40(3):796–800.

[74] Godden JW, Bajorath J. Chemical Descriptors with Distinct Levels of Information Content and Varying Sensitivity to Differences between Selected Compound Databases Identified by SE-DSE Analysis. Journal of Chemical Information and Computer Sciences. 2002;42(1):87–93.

[75] Shannon CE. A mathematical theory of communication. SIGMOBILE Mob Comput Commun Rev. 2001 Jan;5(1):3–55.

[76] Barigye SJ, Marrero-Ponce Y, Martinez-Lopez Y, Torrens F, Artiles-Martinez LM, Pino-Urias RW, et al. Relations frequency hypermatrices in mutual, conditional and joint entropy-based information indices. Journal of Computational Chemistry. 2013;34(4):259–274.

[77] Massy WF. Principal Components Regression in Exploratory Statistical Research. Journal of the American Statistical Association. 1965;60(309):234–256.

[78] Mardia KV, Kent JT, Bibby JM. Multivariate Analysis. London: Academic Press; 1980.

[79] Alzina RB. Introduccion conceptual al analisis multivariable. Un enfoque informatico con los paquetes SPSS-X, BMDP, LISREL Y SPAD; 1989.

[80] Frank lE, Friedman JH. A Statistical View of Some Chemometrics Regression Tools. Technometrics. 1993;35(2):109–135.

[81] Kubinyi H. Evolutionary variable selection in regression and PLS analyses. Journal of Chemometrics. 1996;10(2):119–133.

[82] Wold S. Cross-Validatory Estimation of the Number of Components in Factor and Principal Components Models. Technometrics. 1978;20(4):397–405.

[83] Stone M. Cross-Validatory Choice and Assessment of Statistical Predictions. Journal of the Royal Statistical Society. 1974;36(2):11–147.

[84] Shao J. Linear Model Selection by Cross-validation. Journal of the American Statistical Association. 1993;88(422):486–494.

[85] Leger C, Politis DN, Romano OP. Bootstrap Technology and Applications. Technometrics. 1992;34(4):378–398.

[86] Shao J. Bootstrap Model Selection. Journal of the American Statistical Association. 1996;91(434):655–665.

[87] Tropsha A, Gramatica P, Gombar VK. The Importance of Being Earnest: Validation is the Absolute Essential for Successful Application and Interpretation of QSPR Models. QSAR & Combinatorial Science. 2003;22(1):69–77.

[88] Wold S, Erikson L. In Chemometric Methods in Molecular Design. VCH Publishers: Weinheim; 1995.

[89] Cruciani G, Baroni M, Clementi S, Costantino G, Riganelli D, Skagerberg B. Predictive ability of regression models. Part I: Standard deviation of prediction errors (SDEP). Journal of Chemometrics. 1992;6(6):335–346.

[90] Balaban AT. From Chemical Topology to Three-Dimensional Geometry. New York: Plenum Press; 1997.

[91] Ghose AK, Viswanadhan VN, Wendoloski JJ. Prediction of Hydrophobic (Lipophilic) Properties of Small Organic Molecules Using Fragmental Methods: An Analysis of ALOGP and CLOGP Methods. The Journal of Physical Chemistry A. 1998;102(21):3762–3772.

[92] Nikolic S, Trinajstic N, Mihalic Z, Carter S. On the geometric-distance matrix and the corresponding structural invariants of molecular systems. Chemical Physics Letters. 1991;179:21–28.

[93] Davis C. The norm of the Schur product operation. Numerische Mathematik. 1962;4(1):343–344.

[94] Balaz S, Sturdik E, Rosenberg M, Augustin J, Skara B. Kinetics of drug activities as influenced by their physico-chemical properties: antibacterial effects of alkylating 2-furylethylenes. Journal of theoretical biology;p. 115–134.

[95] A Relationship Between Arbitrary Positive Matrices and Doubly Stochastic Matrices. The Annals of Mathematical Statistics. 1964;35(2):876–879.

[96] Concerning nonnegative matrices and doubly stochastic matrices. Pacific Journal of Mathematics. 1967;21(2):343–348.

[97] Balaban AT. Local versus Global (i.e. Atomic versus Molecular) Numerical Modeling of Molecular Graphs. Journal of Chemical Information and Computer Sciences. 1994;34(2):398–402.

[98] Todeschini R, V C. New Local Vertex Invariants and Molecular Descriptors Based on Functions of the Vertex Degrees. MATCH Commun Math Comput Chem. 2010;64(2):359–372.

[99] Barigye SJ, Marrero-Ponce Y, Martinez-Santiago O, Martinez-Lopez Y, Torrens F. Shannon's, Mutual, Conditional and Joint Entropy Information Indices. Generalization of Global Indices Defined from Local Vertex Invariants. Current computer-aided drug design. 2013;.

[100] Grand M. Java language reference. Sebastopol, CA, USA: O'Reilly & Associates, Inc.; 1997.

[101] Adl-Tabatabai AR, Cierniak M, Lueh GY, Parikh VM, Stichnoth JM. Fast, effective code generation in a just-in-time Java compiler. SIGPLAN Not. 1998 May;33(5):280–290.

[102] Meloan S. The Java HotSpotTM Performance Engine: An In-Depth Look;.

[103] Bull JM, Smith LA, Pottage L, Freeman R. Benchmarking Java against C and Fortran for scientific applications. In: Proceedings of the 2001 joint ACM-ISCOPE conference on Java Grande. JGI '01; 2001. p. 97–105.

[104] Steinbeck C, Han Y, Kuhn S, Horlacher O, Luttmann E, Willighagen E. The Chemistry Development Kit (CDK): An Open-Source Java Library for Chemo- and Bioinformatics. Journal of Chemical Information and Computer Sciences. 2003;43(2):493–500.

[105] Kuhn T, Willighagen E, Zielesny A, Steinbeck C. CDK-Taverna: an open workflow environment for cheminformatics. BMC Bioinformatics. 2010;11(1):159.

[106] Truszkowski A, Jayaseelan K, Neumann S, Willighagen E, Zielesny A, Steinbeck C. New developments on the cheminformatics open workflow environment CDK-Taverna. Journal of Cheminformatics. 2011;3(1):1–10.

[107] O'Boyle N, Hutchison G. Cinfony: combining Open Source cheminformatics toolkits behind a common interface. Chemistry Central Journal. 2008;.

[108] Fourches D, Muratov E, Tropsha A. Trust, But Verify: On the Importance of Chemical Structure Curation in Cheminformatics and QSAR Modeling Research. Journal of Chemical Information and Modeling. 2010;50(7):1189–1204.

[109] otavachemicals; 1997 [cited 2013 May 30]. Available from: `http://www.otavachemicals.com/download-compound-libraries/cat_view/110-diversity-sets/128-primscreen-1`.

[110] Mauri A, Consonni V, Pavan M, Todeschini R. DRAGON software: an easy approach to molecular descriptor calculations. MATCH Communications in Mathematical and in Computer Chemistry. 2006;p. 237–248.

[111] Hong H, Xie Q, Ge W, Qian F, Fang H, Shi L, et al. Mold2, Molecular Descriptors from 2D Structures for Chemoinformatics and Toxicoinformatics. Journal of Chemical Information and Modeling. 2008;48(7):1337–1344.

[112] Yap CW. PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints. Journal of Computational Chemistry;32(7).

[113] BlueDesc-Molecular Descriptor Calculator; [cited 2013 May 30]. Available from: `http://www.ra.cs.uni-tuebingen.de/software/bluedesc/welcome_e.html`.

[114] Liu K, Feng J, Young SS. PowerMV: A Software Environment for Molecular Viewing, Descriptor Generation, Data Analysis and Hit Evaluation. Journal of Chemical Information and Modeling. 2005;45(2):515–522.

[115] R F. Theoretical Drug Design Methods. Elsevier; 1984.

[116] MobyDigs: software for regression and classification models by genetic algorithms. In: Nature-insprired Methods in Chemometrics: Genetic Algorithms and Artificial Neural Networks. vol. 23 of Data Handling in Science and Technology. Elsevier; 2003. p. 141 – 167.

[117] Beger RD, Buzatu DA, Wilkes JG, Jackson. Comparative Structural Connectivity Spectra Analysis (CoSCoSA) Models of Steroid Binding to the Corticosteroid Binding Globulin. Journal of Chemical Information and Computer Sciences. 2002;42(5):1123–1131.

[118] Beger RD, Harris S, Xie Q. Models of Steroid Binding Based on the Minimum Deviation of Structurally Assigned 13C NMR Spectra Analysis (MiDSASA). Journal of Chemical Information and Computer Sciences. 2004;44(4):1489–1496.

[119] Robert D, Amat L, Carbo-Dorca R. Three-Dimensional Quantitative Structure-Activity Relationships from Tuned Molecular Quantum Similarity Measures: Prediction of the Corticosteroid Binding Globulin Binding Affinity for a Steroid Family. Journal of Chemical Information and Computer Sciences. 1999;39(2):333–344.

[120] Kellogg G, Kier L, Gaillard P, Hall L. E-state fields: Applications to 3D QSAR. Journal of Computer-Aided Molecular Design. 1996;10(6):513–520.

[121] Polanski J, Bak A. Modeling Steric and Electronic Effects in 3D- and 4D-QSAR Schemes: Predicting Benzoic pKa Values and Steroid CBG Binding Affinities. Journal of Chemical Information and Computer Sciences. 2003;43(6):2081–2092.

[122] Beger R, Wilkes J. Developing 13C NMR quantitative spectrometric data-activity relationship (QSDAR) models of steroid binding to the corticosteroid binding globulin. Journal of Computer-Aided Molecular Design. 2001;15(7):659–669.

[123] de Gregorio C, Kier L, Hall L. QSAR modeling with the electrotopological state indices: Corticosteroids. Journal of Computer-Aided Molecular Design. 1998;12(6):557–561.

[124] Lobato M, Amat L, Besalu E, Carbo-Dorca R. Structure-Activity Relationships of a Steroid Family using Quantum Similarity Measures and Topological Quantum Similarity Indices. Quantitative Structure-Activity Relationships. 1997;16(6):465–472.

[125] Liu SS, Yin CS, Wang LS. Combined MEDV-GA-MLR Method for QSAR of Three Panels of Steroids, Dipeptides, and COX-2 Inhibitors. Journal of Chemical Information and Computer Sciences. 2002;42(3):749–756.

[126] Polanski J, Walczak B. The comparative molecular surface analysis (COMSA): a novel tool for molecular design. Computers & Chemistry. 2000;24(5):615 – 625.

[127] Parretti MF, Kroemer RT, Rothman JH, Richards WG. Alignment of molecules by the Monte Carlo optimization of molecular similarity indices. Journal of Computational Chemistry. 1997;18(11):1344–1353.

[128] Besalu E, Girones X, Amat L, Carbo-Dorca R. Molecular Quantum Similarity and the Fundamentals of QSAR. Accounts of Chemical Research. 2002;35(5):289–295.

[129] Maw HH, Hall LH. E-State Modeling of Corticosteroids Binding Affinity Validation of Model for Small Data Set. Journal of Chemical Information and Computer Sciences. 2001;41(5):1248–1254.

[130] Manchester J, Czerminski R. SAMFA: Simplifying Molecular Description for 3D-QSAR. Journal of Chemical Information and Modeling. 2008;48(6):1167–1173.

[131] Tominaga Y, Fujiwara I. Prediction-Weighted Partial Least-Squares Regression Method (PWPLS) 2: Application to CoMFA. Journal of Chemical Information and Computer Sciences. 1997;37(6):1152–1157.

[132] Chen H, Zhou J, Xie G. PARM: A Genetic Evolved Algorithm To Predict Bioactivity. Journal of Chemical Information and Computer Sciences. 1998;38(2):243–250.

[133] Kotani T, Higashiura K. Comparative Molecular Active Site Analysis (CoMASA). 1. An Approach to Rapid Evaluation of 3D QSAR. Journal of Medicinal Chemistry. 2004;47(11):2732–2742.

[134] Norinder U. 3D-QSAR investigation of the tripos benchmark steroids and some protein-tyrosine kinase inhibitors of styrene type using the TDQ approach. Journal of Chemometrics;10(5-6).

[135] Good AC, So SS, Richards WG. Structure-activity relationships from molecular similarity matrices. Journal of Medicinal Chemistry. 1993;36(4):433–438.

[136] Marin RM, Aguirre NF, Daza EE. Graph Theoretical Similarity Approach To Compare Molecular Electrostatic Potentials. Journal of Chemical Information and Modeling. 2008;48(1):109–118.

[137] Tuppurainen K, Viisas M, Perakyla M, Laatikainen R. Ligand intramolecular motions in ligand-protein interaction: ALPHA, a novel dynamic descriptor and a QSAR study with extended steroid benchmark dataset. Journal of Computer-Aided Molecular Design. 2004;18(3):175–187.

[138] Palyulin VA, Radchenko EV, Zefirov NS. Molecular Field Topology Analysis Method in QSAR Studies of Organic Compounds. Journal of Chemical Information and Computer Sciences. 2000;40(3):659–667.

[139] Tuppurainen K, Viisas M, Laatikainen R, Perakyla M. Evaluation of a Novel Electronic Eigenvalue (EEVA) Molecular Descriptor for QSAR/QSPR Studies: Validation Using a Benchmark Steroid Data Set. Journal of Chemical Information and Computer Sciences. 2002;42(3):607–613.

[140] Robinson DD, Winn PJ, Lyne PD, Richards WG. Self-Organizing Molecular Field Analysis: A Tool for Structure Activity Studies. Journal of Medicinal Chemistry. 1999;42(4):573–583.

[141] Amat L, Besalu E, Carbo-Dorca R, Ponec R. Identification of Active Molecular Sites Using Quantum-Self-Similarity Measures. Journal of Chemical Information and Computer Sciences. 2001;41(4):978–991.

Apéndice A

Definición matemática de los operadores de agregación

No.	Grupo	Nombre	ID	Fórmula
1	**Normas**	Norma Minkowski (p = 1)	N1	$N = \sqrt[p]{\sum_{a=1}^{n} (L_a)^p}$
		Norma Manhattan		
2		Norma Minkowski (p = 2)	N2	
		Norma Euclidean		
3		Norma Minkowsk (p = 3)	N3	
4		Penrose	PN	$PN = \sqrt{\frac{1}{n^2}\left[\sum_{a=1}^{n} L_a\right]^2}$
5	**Estadísticos de tendencia central**	Media Geométrica	GM	$G = \sqrt[n]{\prod_{a=1}^{n} L_a}$
6		Media Aritmética ($\beta = 1$)	AM	$M_\beta = \left(\frac{L_1^\beta + L_2^\beta + \ldots + L_n^\beta}{n}\right)^{\frac{1}{\beta}}$
7		Media Cuadrática ($\beta = 2$)	P2	
8		Media Potencial ($\beta = 3$)	P3	
9		Media Armónica ($\beta = -1$)	HM	
10		Varianza	V	$V = \frac{\sum_{a-1}^{n} (L_a - M)^2}{n-1}$

11		Skewness	S	$S = \frac{n * (X_3)}{(n-1)(n-2)(DE)^3}$ $X_3 = \sum_{a=1}^{n} (L_a - M)^3$ $M:$ *media aritmética* $DE:$ *desviación estándar*
12	**Estadísticos de dispersión y forma**	Kurtosis	K	$K = \frac{n(n+1)X_4 - 3(X_2)(X_2)(n-1)}{(n-1)(n-2)(n-3)(DE)^4}$ $X_j = \sum_{a=1}^{n} (L_a - M)^j$ $M:$ *media aritmética* $DE:$ *desviación estándar*
13		Desviación Estándar	DE	$DE = \sqrt{\frac{\left(\sum_{a=1}^{n} L_a - M\right)^2}{n-1}}$
14		Coeficiente de Variación	VC	$VC = \frac{DE}{M}$ $DE:$ *desviación estándar* $M:$ *media aritmética*
15		Rango	RA	$RA = L_{max} - L_{min}$
16		Percentile 25	Q1	$Q1 = \left[\frac{N}{4} + \frac{1}{2}\right]$ $N:$ *cantidad de* L_i
17		Percentile 50	Q2	$Q1 = \left[\frac{N}{2} + \frac{1}{2}\right]$ $N:$ *cantidad de* L_i
18		Percentile 75	Q3	$Q1 = \left[\frac{3N}{4} + \frac{1}{2}\right]$ $N:$ *cantidad de* L_i
19		Rango Inter-Cuartil	I50	$I50 = Q3 - Q1$
20		Máximo Valor	MX	$MX = max(L_i)$
21		Mínimo Valor	MN	$MN = min(L_i)$
22		Autocorrelación	AC^k	$AC^k = \sum_{i=1}^{n}\sum_{j=1}^{n} L_i \times L_j \bullet \delta(d_{ij}, k)$ $k = 1, 2, 3, \ldots, 7$
23		Gravitacional	GV^k	$GV^k = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} \frac{L_i L_j}{{}^k d_{ij}} \bullet \delta(d_{ij}, k)$ $k = 1, 2, 3, \ldots, 7$

24		Suma Total	TS^k	$TS^k = \sum_{i=1}^{n}\sum_{j=1}^{n} L_{ij} \bullet \delta\left(d_{ij}, k\right)$ $k = 1,\ 2,\ 3, \ldots,\ 7$
25	**Algoritmos clásicos**	Conectividad de Kier-Hall	KH^m	${}^mKH_t = \sum_{i=1}^{K}\left(\prod_{j=1}^{n_k} L_i,\ w\right)_k^{\lambda}$ donde, K es el número de sub-grafos, n_k es el número de átomos en un fragmento, $\lambda = \frac{1}{2}$, m es el orden y t es el tipo del subgrafo
26		Contenido de Información Media	MIC	$MIC = -\sum_{i=1}^{a}\frac{N_g}{N_o} * \log_2 \frac{N_g}{N_o}$ donde, N_g es el número de átomos con el mismo valor de $LOVI$. N_o es el número de átomos en la molécula.
27		Contenido de Información Total	TIC	$TIC = N_o * \log_2 N_o - \sum_{g=1}^{G} N_g * \log_2 N_g$
28		Contenido de Información Estandarizado	SIC	$SIC = \frac{IT}{N_o * \log_2 N_o}$
29		Estado Electrotopológico	ES	$S_i = I_i + \triangle I_i = I_i + \sum_{j=1}^{n}\frac{I_i - I_j}{\left(d_{ij}+1\right)^2}$ donde, $k = 2$, I_i es el estado intrínseco del i^{th} átomo, y $\triangle I_i$ es el efecto del campo sobre el i^{th} átomo calculado como perturbación del I_i del i^{th} átomo por los otros átomos de la molécula, d_{ij} es la distancia topológica entre el átomo i y j, y n es el número de átomos.
30		Ivanciuc-Balaban	IB	$J_k = \frac{n^2 * B}{n + C + 1}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} a_{ij}\left[L_i \times L_j\right]^{-\frac{1}{2}}$ donde, la sumatoria es realizada sobre los átomos adyacentes a_{ij}. Los parámetros n, B y C son el número de átomos, enlaces y anillos respectivamente.

Apéndice B

Proyecto de los descriptores utilizados para realizar las pruebas de rendimiento multi-núcleo

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<ToMoCoMD-cardd_project>
    <qubils>
        <MID/>
    </qubils>
    <algebraic_form>
        <bilinear/>
        <linear/>
        <quadratic/>
    </algebraic_form>
    <options>
        <hAtomsOFF/>
        <lonePairOFF/>
    </options>
    <constrains_form>
        <nonChiral/>
        <duples/>
    </constrains_form>
```

```
19      <matrix_form order="12">
20          <non_stochastic/>
21          <simple_stochastic/>
22          <mutual_probability/>
23      </matrix_form>
24      <metrics>
25          <m3/>
26      </metrics>
27      <cut_off>
28          <all/>
29      </cut_off>
30      <groups>
31          <total/>
32      </groups>
33      <properties>
34          <mass/>
35          <electronegativity/>
36          <vdw/>
37          <polarizability/>
38      </properties>
39      <invariants standarized="false">
40          <n1/>
41          <n2/>
42          <n3/>
43          <pn/>
44          <am/>
45          <p3/>
46          <v/>
47          <sd/>
48          <vc/>
49      </invariants>
50  </ToMoCoMD-cardd_project>
```

Apéndice C

Resultados de los análisis de componentes principales

Factor	Autovalor	% Total de la Varianza	Autovalor Acumulativo	Acumulado - %
1	63.762	30.655	63.762	30.655
2	50.794	24.420	114.556	55.075
3	27.735	13.334	142.292	68.409
4	19.948	9.591	162.240	78.000
5	13.020	6.260	175.260	84.260

Tabla C.1: Resultado del análisis de componentes principales para los índices 3D propuestos respecto al enfoque de matrices no-, simple- y doble-estocásticas y de probabilidad mutua

Factor	Autovalor	% Total de la Varianza	Autovalor Acumulativo	Acumulado - %
1	85.966	57.311	85.966	57.311
2	27.056	18.037	113.022	75.348
3	9.001	6.001	122.023	81.349
4	8.320	5.546	130.343	86.896
5	7.221	4.814	137.565	91.710
6	5.733	3.822	143.297	95.532
7	2.774	1.849	146.072	97.381

Tabla C.2: Resultado del análisis de componentes principales para los índices 3D propuestos respecto a la métrica para el cálculo de distancias inter-atómicas

Factor	Autovalor	% Total de la Varianza	Autovalor Acumulativo	Acumulado - %
1	653.777	62.863	653.777	62.863
2	126.679	12.181	780.456	75.044
3	111.146	10.687	891.602	85.731
4	50.572	4.863	942.174	90.594
5	18.606	1.789	960.781	92.383

Tabla C.3: Resultado del análisis de componentes principales para los índices 3D ternarios propuestos respecto a la medida utilizada

Factor	Autovalor	% Total de la Varianza	Autovalor Acumulativo	Acumulado - %
1	613.261	58.967	613.261	58.967
2	167.908	16.145	781.169	75.112
3	115.040	11.062	896.209	86.174
4	44.944	4.322	941.153	90.495

Tabla C.4: Resultado del análisis de componentes principales para los índices 3D cuaternarios propuestos respecto a la medida utilizada

Factor	Autovalor	% Total de la Varianza	Autovalor Acumulativo	Acumulado - %
1	174.326	64.327	174.326	64.327
2	51.427	18.977	225.753	83.304
3	24.234	8.943	249.987	92.246
4	6.093	2.248	256.080	94.494

Tabla C.5: Resultado del análisis de componentes principales para los índices 3D propuestos respecto a los operadores de agregación de norma, estadístico de tendencia central, y estadístico de dispersión y forma aplicados

Factor	Autovalor	% Total de la Varianza	Autovalor Acumulativo	Acumulado - %
1	111.588	48.942	111.588	48.942
2	42.835	18.787	154.423	67.729
3	39.736	17.428	194.159	85.157
4	12.931	5.672	207.090	90.829
5	5.719	2.508	212.808	93.337
6	4.183	1.834	216.991	95.171

Tabla C.6: Resultado del análisis de componentes principales para los índices 3D propuestos respecto a los algoritmos clásicos utilizados como operadores de agregación

Factor	Autovalor	% Total de la Varianza	Autovalor Acumulativo	Acumulado - %
1	409.047	26.527	409.047	26.527
2	204.033	13.232	613.080	39.759
3	90.647	5.879	703.727	45.637
4	69.474	4.505	773.201	50.143
5	61.355	3.979	834.556	54.122
6	51.239	3.323	885.795	57.445
7	42.925	2.784	928.720	60.228
8	37.376	2.424	966.096	62.652
9	34.464	2.235	1000.560	64.887
10	26.680	1.730	1027.240	66.617
11	24.396	1.582	1051.636	68.199
12	22.295	1.446	1073.931	69.645
13	20.669	1.340	1094.600	70.986
14	18.462	1.197	1113.062	72.183
15	16.441	1.066	1129.502	73.249

Tabla C.7: Resultado del análisis de componentes principales para los índices 3D QuBiLs MIDAS Duplas respecto a los índices 3D del DRAGON

Factor	Autovalor	% Total de la Varianza	Autovalor Acumulativo	Acumulado - %
1	982.011	33.210	982.011	33.210
2	272.037	9.200	1254.048	42.409
3	146.285	4.947	1400.333	47.357
4	105.699	3.575	1506.032	50.931
5	82.757	2.799	1588.789	53.730
6	79.171	2.677	1667.960	56.407
7	63.131	2.135	1731.091	58.542
8	54.903	1.857	1785.994	60.399
9	48.099	1.627	1834.094	62.025
10	42.801	1.447	1876.895	63.473
11	39.203	1.326	1916.098	64.799

Tabla C.8: Resultado del análisis de componentes principales para los índices 3D QuBiLs MIDAS Ternas respecto a los índices 3D del DRAGON

Factor	Autovalor	% Total de la Varianza	Autovalor Acumulativo	Acumulado - %
1	872.588	29.509	872.588	29.509
2	244.885	8.282	1117.472	37.791
3	191.740	6.484	1309.212	44.275
4	130.840	4.425	1440.052	48.700
5	101.117	3.420	1541.169	52.119
6	93.413	3.159	1634.583	55.278
7	73.308	2.479	1707.891	57.758
8	65.258	2.207	1773.149	59.964
9	58.629	1.983	1831.778	61.947
10	53.769	1.818	1885.547	63.766
11	46.165	1.561	1931.712	65.327
12	41.308	1.397	1973.020	66.724
13	39.804	1.346	2012.824	68.070

Tabla C.9: Resultado del análisis de componentes principales para los índices 3D QuBiLs MIDAS Cuaternas respecto a los índices 3D del DRAGON

Factor	Autovalor	% Total de la Varianza	Autovalor Acumulativo	Acumulado - %
1	1196.676	41.580	1196.676	41.580
2	316.980	11.014	1513.656	52.594
3	215.148	7.476	1728.804	60.070
4	124.401	4.322	1853.204	64.392
5	103.763	3.605	1956.968	67.997
6	84.744	2.945	2041.711	70.942
7	77.239	2.684	2118.950	73.626
8	59.159	2.056	2178.109	75.681
9	51.186	1.779	2229.295	77.460
10	43.746	1.520	2273.041	78.980
11	37.662	1.309	2310.703	80.288

Tabla C.10: Resultado del análisis de componentes principales para los índices 3D QuBiLs MIDAS Nuplas (ternas + cuaternas) respecto a los índices 3D del DRAGON

Apéndice D

Base de datos de Esteroides para la predicción

ID	Nombre	Y. Exp	ID	Nombre	Y. Exp
1	aldosterone	-6.28	17	pregnenolone	-5.23
2	androstanediol	-5.00	18	17a-hydroxypregnenolone	-5.00
3	5-androstenediol	-5.00	19	progesterone	-7.38
4	4-androstenedione	-5.76	20	17a-hydroxyprogesterone	-7.74
5	androsterone	-5.61	21	testosterone	-6.72
6	corticosterone	-7.88	22	prednisolone	-7.51
7	cortisol	-7.88	23	cortisolacetat	-7.55
8	cortisone	-6.89	24	4-pregnene-3,11,20-trione	-6.78
9	dehydroepiandrosterone	-5.00	25	epicorticosterone	-7.20
10	11-deoxycorticosterone	-7.65	26	19-nortestosterone	-6.14
11	11-deoxycortisol	-7.88	27	16a,17a-dihydroxyprogesterone	-6.25
12	dihydrotestosterone	-5.92	28	16a-methylprogesterone	-7.12
13	estradiol	-5.00	29	19-norprogesterone	-6.82
14	estriol	-5.00	30	2a-methylcortisol	-7.69
15	estrone	-5.00	31	2a-methyl-9a-fluoro-cortisol	-5.80
16	etiocholanolone	-5.23			

Tabla D.1: Nombre de las estructuras de la base de datos de Esteroides propuesta por Cramer y la afinidad de unión correspondiente con la CBG

Apéndice E

Gráficas de la modelación interna de los estudios QSAR

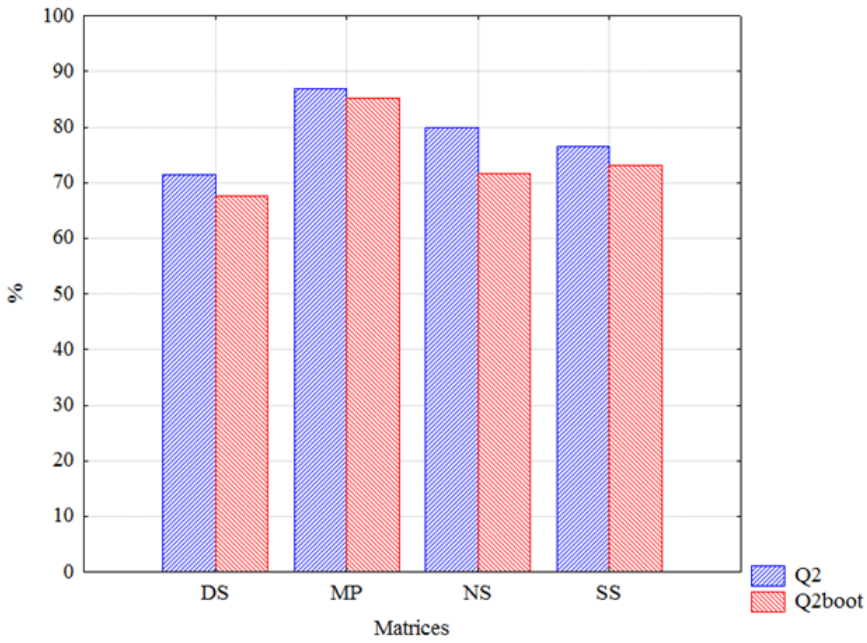

Figura E.1: Comparación en la modelación QSAR de los índices QuBiLs MIDAS propuestos acorde a los enfoques matriciales doble- (DS), simple- (SS) y no estocásticos (NS) y de probabilidad mutua (MP)

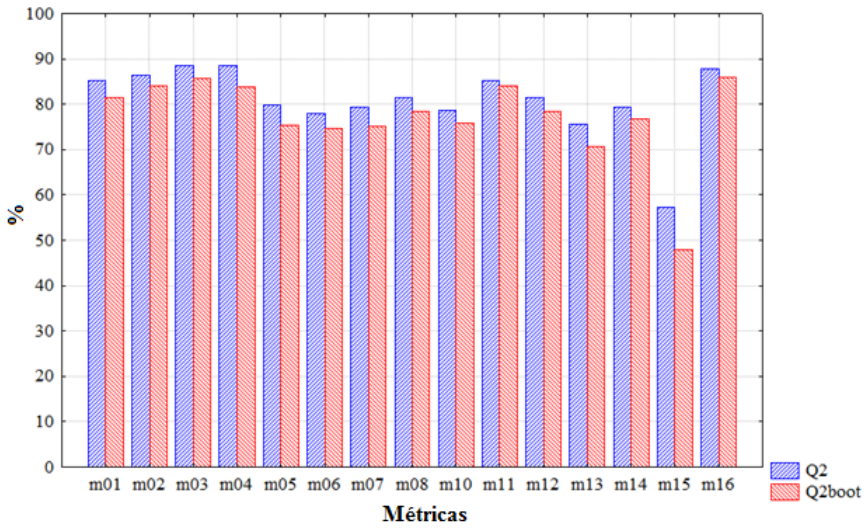

Figura E.2: Comparación en la modelación QSAR de los índices QuBiLs MIDAS propuestos acorde a las métricas de Minkowski [m01 (p=0.25); m02 (p=0.5); m03 (p=1, Manhattan); m04 (p=1.5); m05 (p=2, Euclidian); m06 (p=2.5); m07 (p=3); m08 (Chebyshev)], Canberra (m10), Lance-Williams (m11), Clark (m12), Soergel (m13), Bhattacharyya (m14), Wave-Edges (m15) y Separación Angular (m16)

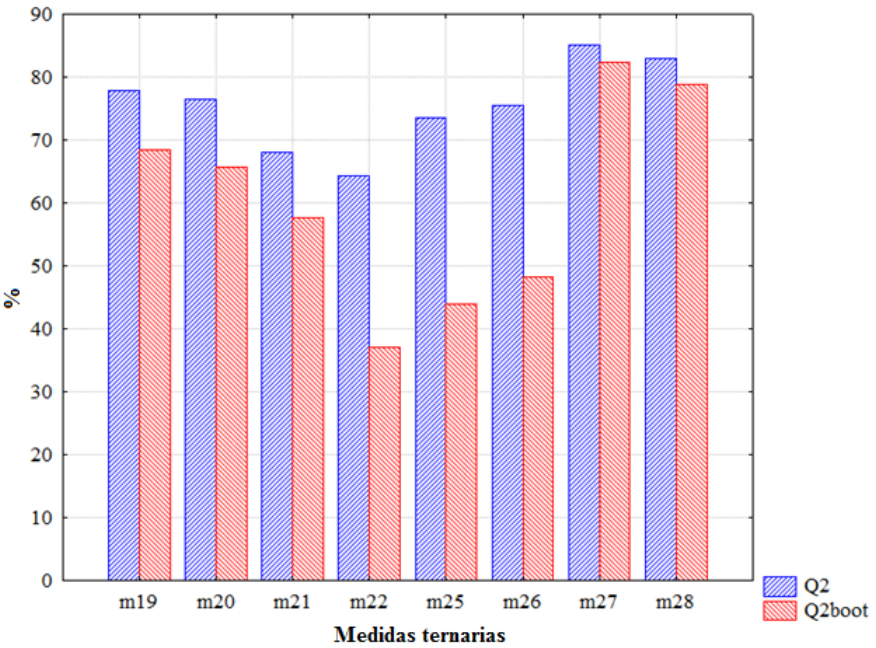

Figura E.3: Comparación en la modelación QSAR de los índices QuBiLs MIDAS propuestos acorde a las medidas basadas en relaciones ternarias entre átomos: Perímetro (m19), Perímetro total (m20), Área (m21), Área total (m22), Suma de lados (m25), Suma de lados total (m26), Ángulo entre lados (m27) y Ángulo entre lados total (m28)

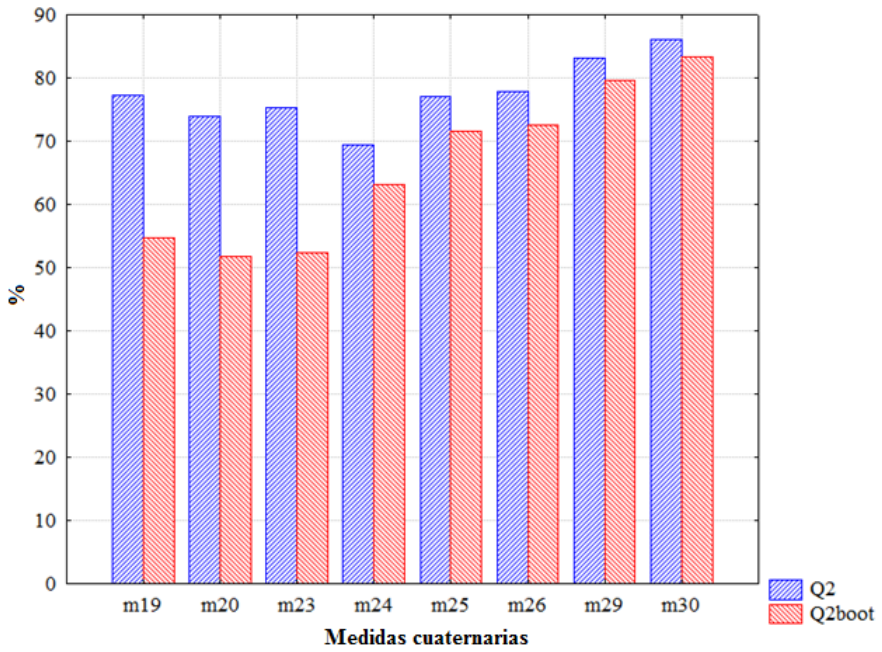

Figura E.4: Comparación en la modelación QSAR de los índices QuBiLs MIDAS propuestos acorde a las medidas basadas en relaciones cuaternarias entre átomos: Perímetro (m19), Perímetro total (m20), Volumen (m23), Volumen total (m22), Sumatoria de lados (m25), Sumatoria de lados total (m26), Ángulo diedro (m29) y Ángulo diedro total (m30)

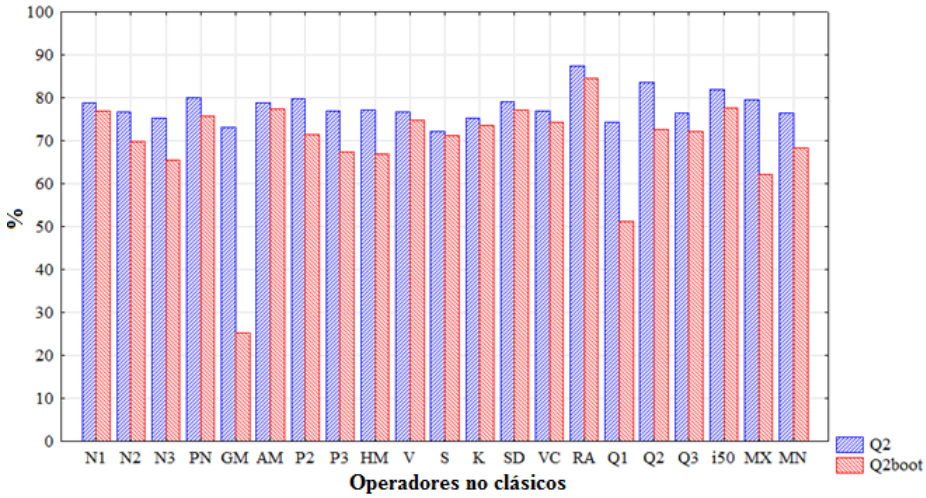

Figura E.5: Comparación en la modelación QSAR de los índices QuBiLs MIDAS propuestos acorde a los operadores Manhattan (N1), Euclidian (N2), Minkowski (N3), Penrose (PN), Media Geométrica (GM), Media Aritmética (AM), Media Cuadrática (P2), Media Potencial (P3), Media Armónica (HM), Varianza (V), Skewness (S), Kurtosis (K), Desviación Estándar (SD), Coeficiente de Variación (VC), Rango (RA), Percentile 25 (Q1), Percentile 50 (Q2), Percentile 75 (Q3), i50, Máximo (MX) y Mínimo (MN)

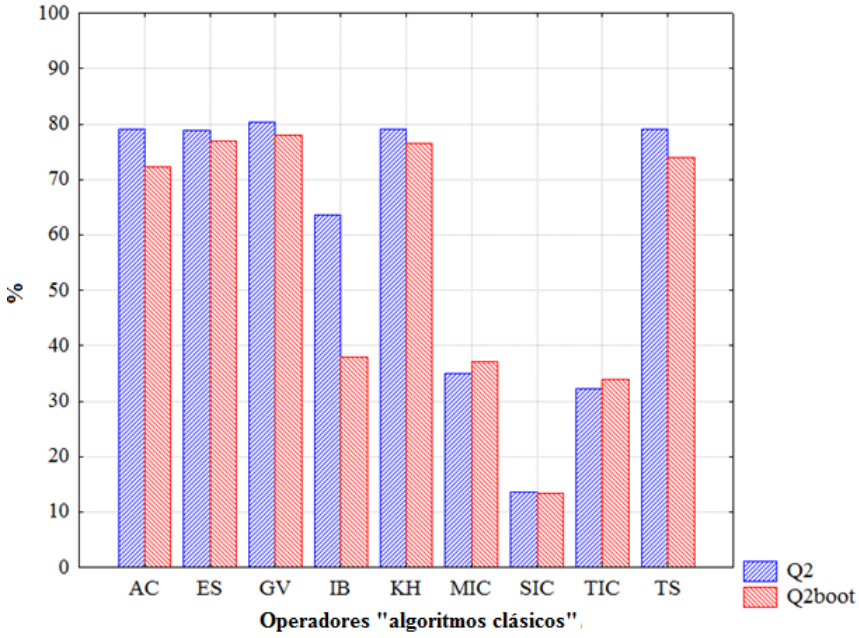

Figura E.6: Comparación en la modelación QSAR de los índices QuBiLs MIDAS propuestos acorde a los operadores clásicos Autocorrelación (AC), Estado Electrotopológico (ES), Gravitacional (GV), Ivanciuc-Balaban (IB), Kier-Hall (KH), Contenido de Información Media (MIC), Contenido de Información Estandarizado (SIC), Contenido de Información Total (TIC) y Suma Total (TS)

Apéndice F

Modelos QSAR obtenidos con los índices QuBiLs MIDAS

	R^2	Q^2loo	Q^2boot	a(Q^2)	SDEC	F	Modelos	Ec.
3	0.939	0.917	0.916	-0.284	0.262	139.7	**logK** = –4.673 (±0.714) – 5.028 (±0.893) ${}^{SD}_{SS4}B^{M2}_{h-e}$ + 66.882 (±11.766) ${}^{RA}_{MP0}B^{M3}_{h-e}$ – 0.090 (±0.011) ${}^{AC[7]-K}_{NS7}B^{M2}_{h-c}$	1
4	0.966	0.952	0.949	-0.312	0.194	189.6	**logK** = –15.840 (±1.082) + 0.305 (±0.055) ${}^{K}_{SS7}B^{M2}_{a-v}$ – 0.051 (±0.012) ${}^{AC[7]-K}_{MP7}B^{M2}_{h-c}$ + 23.339 (±2.666) ${}^{KH[3]-i50}_{SS1}F^{M8}_{h}$ + 13.845 (±1.357) ${}^{GV[6]-AM}_{SS4}F^{M16}_{h}$	2
5	0.977	0.964	0.960	-0.434	0.160	216.2	**logK** = –29.015 (±3.970) + 0.433 (±0.059) ${}^{K}_{SS7}B^{M2}_{a-v}$ – 0.047 (±0.010) ${}^{GV[7]-K}_{NS7}B^{M2}_{h-c}$ + 1.816 (±0.533) ${}^{PN}_{SS7}Q^{M8}_{e}$ + 20.788 (±2.366) ${}^{KH[3]-i50}_{SS1}Q^{M8}_{h}$ + 2.821 (±0.243) ${}^{AC[6]-PN}_{SS4}F^{M16}_{h}$	3
6	0.986	0.978	0.974	-0.423	0.125	286.3	**logK** = –16.228 (±0.729) + 11.664 (±1.531) ${}^{i50}_{NS6}B^{M4}_{a-e}$ – 0.193 (±0.036) ${}^{SD}_{NS1}B^{M2}_{a-v}$ + 0.518 (±0.042) ${}^{K}_{SS7}B^{M2}_{a-v}$ + 28.826 (±1.681) ${}^{KH[3]-i50}_{SS1}F^{M8}_{h}$ – 0.011 (±0.003) ${}^{GV[2]-K}_{NS6}B^{M12}_{c-e}$ + 16.421 (±0.622) ${}^{GV[6]-PN}_{SS4}F^{M16}_{h}$	4

Tabla F.1: Parámetros estadísticos de los mejores modelos de dimensiones desde 3 hasta 6 variables utilizando los índices QuBiLs MIDAS Duplas, considerando las 31 estructuras como conjunto de entrenamiento

	R^2	Q^2loo	Q^2boot	a(Q^2)	SDEC	F	Modelos	Ec.
3	0.932	0.909	0.908	-0.238	0.276	124.4	**logK** = –4.815 (±0.221) + 0.026 (±0.005) ${}^{i50}_{NS1}TrC^{M20(M2)}_{e}$ – 0.041 (±0.005) ${}^{AC[2]-K}_{MP5}TrQB^{M19(M8)}_{a-c}$ – 86.681 (±9.837) ${}^{TS[1]-i50}_{SS1}TrQB^{M21(M2)}_{a-c}$	5
4	0.965	0.948	0.945	-0.374	0.198	180.9	**logK** = 1.239 (±0.598) – 0.046 (±0.004) ${}^{AC[2]-K}_{MP5}TrQB^{M19(M8)}_{a-c}$ – 3.642 (±0.457) ${}^{ES-PN}_{MP1}Tr^{M21(M1)}_{a-v-h}$ – 112.034 (±8.660) ${}^{TS[1]-i50}_{SS1}TrQB^{M21(M2)}_{a-c}$ – 32.933 (±4.115) ${}^{TS[1]-i50}_{MP5}TrB^{M20(M4)}_{e-h}$	6
5	0.975	0.963	0.957	-0.407	0.167	199.0	**logK** = –4.073 (±0.205) + 52.142 (±9.409) ${}^{PN}_{MP2}Tr^{M28}_{a-v-c}$ – 0.036 (±0.003) ${}^{AC[2]-K}_{MP5}TrQB^{M19(M8)}_{a-c}$ + 6.114 (±1.288) ${}^{ES-SD}_{SS0}Tr^{M19(M12)}_{e-h-c}$ – 100.726 (±10.873) ${}^{TS[1]-i50}_{SS1}TrQB^{M21(M2)}_{a-c}$ – 27.281 (±3.170) ${}^{TS[1]-i50}_{MP5}TrB^{M20(M4)}_{e-h}$	7
6	0.981	0.970	0.965	-0.481	0.144	215.1	**logK** = –13.002 (±1.353) – 0.039 (±0.005) ${}^{AC[2]-K}_{SS7}TrB^{M20(M4)}_{a-c}$ – 0.083 (±0.013) ${}^{AC[7]-K}_{SS5}Tr^{M19(M13)}_{a-h-c}$ + 4.839 (±0.814) ${}^{ES-PN}_{MP5}Tr^{M21(M3)}_{e-v-c}$ + 0.168 (±0.022) ${}^{TS[2]-K}_{MP2}TrB^{M19(M3)}_{e-h}$ – 10.767 (±1.940) ${}^{TS[4]-i50}_{SS1}TrB^{M21(M8)}_{a-c}$ + 2.081 (±0.357) ${}^{MX}_{SS4}TrF^{M27}_{e}$	8

Tabla F.2: Parámetros estadísticos de los mejores modelos de dimensiones desde 3 hasta 6 variables utilizando los índices QuBiLs MIDAS Ternas, considerando las 31 estructuras como conjunto de entrenamiento

	R^2	Q^2loo	Q^2boot	a(Q^2)	SDEC	F	Modelos	Ec.
3	0.930	0.915	0.913	-0.328	0.281	120.1	**logK** = –24.147 (±0.956) + 0.059 (±0.006) $^{SD}_{MP7}QuCB^{M26(M3)}_{v-h}$ + 24.435 (±1.925) $^{ES-SD}_{MP2}QuQTr^{M26(M3)}_{m-e-p}$ + 1.186 (±0.282) $^{TS[3]-K}_{SS6}QuCB^{M25(M13)}_{m-v}$	9
4	0.960	0.947	0.944	-0.328	0.212	157.6	**logK** = –21.795 (±0.887) + 0.053 (±0.005) $^{SD}_{MP7}QuCB^{M26(M3)}_{v-h}$ – 0.043 (±0.010) $^{AC[6]-K}_{SS7}QuQd^{M30}_{p}$ + 55.319 (±4.050) $^{ES-SD}_{MP2}QuCB^{M26(M3)}_{m-p}$ + 1.284 (±0.218) $^{TS[3]-K}_{SS6}QuCB^{M25(M13)}_{m-v}$	10
5	0.970	0.956	0.951	-0.435	0.184	162.4	**logK** = –19.455 (±1.135) + 0.043 (±0.006) $^{SD}_{MP7}QuCB^{M26(M3)}_{v-h}$ – 0.048 (±0.009) $^{AC[6]-K}_{SS7}QuQd^{M29}_{p}$ + 50.233 (±4.005) $^{ES-SD}_{MP2}QuCB^{M26(M3)}_{m-p}$ + 1.208 (±0.195) $^{TS[3]-K}_{SS6}QuCB^{M25(M13)}_{m-v}$ + 0.361 (±0.126) $^{TS[5]-K}_{SS7}QuCB^{M26(M13)}_{m-a}$	11
6	0.979	0.968	0.963	-0.482	0.151	195.2	**logK** = –10.484 (±0.3706) – 0.0035 (±0.0007) $^{N1}_{SS5}QuCB^{M29}_{a-v}$ – 6.4181 (±0.5426) $^{N1}_{SS7}QuCB^{M25(M13)}_{a-c}$ + 6.5751 (±0.5183) $^{SD}_{MP4}QuQd^{M25(M8)}_{e}$ – 0.05 (±0.0077) $^{AC[6]-K}_{SS7}QuQd^{M30}_{p}$ – 0.0008 (±0.0001) $^{AC[6]-PN}_{SS3}Qu^{M26(M13)}_{psa-a-v-h}$ + 0.0013 (±0.0003) $^{TS[4]-i50}_{SS5}Qu^{M26(M13)}_{r-e-v-p}$	12

Tabla F.3: Parámetros estadísticos de los mejores modelos de dimensiones desde 3 hasta 6 variables utilizando los índices QuBiLs MIDAS Cuaternas, considerando las 31 estructuras como conjunto de entrenamiento

	R^2	Q^2loo	Q^2boot	a(Q^2)	SDEC	F	Modelos	Ec.
3	0.939	0.922	0.919	-0.334	0.263	138.8	**logK** = –8.343 (±0.523) – 90.249 (±13.083) $^{TS[1]-i50}_{SS1}TrQB^{M21(M2)}_{a-c}$ + 1.192 (±0.159) $^{TS[1]-i50}_{MP3}TrQB^{M20(M4)}_{e-v}$ + 26.424 (±3.862) $^{TS[2]-i50}_{SS0}Qu^{M29}_{a-r-s-c}$	13
4	0.962	0.950	0.948	-0.372	0.205	168.4	**logK** = –7.169 (±0.506) – 98.764 (±10.624) $^{TS[1]-i50}_{SS1}TrQB^{M21(M2)}_{a-c}$ + 0.985 (±0.137) $^{TS[1]-i50}_{MP3}TrQB^{M20(M4)}_{e-v}$ + 26.140 (±3.075) $^{TS[2]-i50}_{SS0}Qu^{M30}_{a-r-s-c}$ – 1.100 (±0.270) $^{TS[6]-i50}_{SS4}QuCB^{M24(M16)}_{a-s}$	14
5	0.979	0.969	0.964	-0.410	0.153	236.1	**logK** = –4.430 (±0.335) + 7.759 (±1.671) $^{RA}_{MP3}TrQB^{M28}_{e-c}$ – 0.024 (±0.003) $^{AC[2]-K}_{NS5}TrQB^{M19(M8)}_{a-c}$ – 110.977 (±6.769) $^{TS[1]-i50}_{SS1}TrQB^{M21(M2)}_{a-c}$ + 1.044 (±0.187) $^{AC[4]-PN}_{SS6}QuQd^{M30}_{m}$ – 1.477 (±0.288) $^{TS[1]-i50}_{SS3}QuQd^{M25(M3)}_{a}$	15
6	0.987	0.980	0.976	-0.521	0.120	309.8	**logK** = –4.688 (±0.278) + 24.208 (±5.295) $^{SD}_{MP3}TrQB^{M27}_{e-c}$ – 0.053 (±0.014) $^{AC[7]-K}_{SS5}Tr^{M19(M13)}_{a-h-c}$ – 0.024 (±0.003) $^{AC[2]-K}_{MP5}TrQB^{M19(M8)}_{a-c}$ – 91.193 (±7.130) $^{TS[1]-i50}_{SS1}TrQB^{M21(M2)}_{a-c}$ + 1.099 (±0.152) $^{AC[4]-PN}_{SS6}QuQd^{M30}_{m}$ – 1.423 (±0.230) $^{TS[1]-i50}_{SS3}QuQd^{M25(M3)}_{a}$	16

Tabla F.4: Parámetros estadísticos de los mejores modelos de dimensiones desde 3 hasta 6 variables utilizando los índices QuBiLs MIDAS Nuplas, considerando las 31 estructuras como conjunto de entrenamiento

	R^2	Q^2loo	Q^2boot	a(Q^2)	$SDEP^{ext}$	F	Modelos	Ec.
2	0.956	0.945	0.944	-0.358	0.512	196.3	**logK** = –3.837 (±0.219) – 0.109 (±0.015) ${}^{AC[7]-K}_{NS7}B^{M2}_{h-c}$ – 0.012 (±0.003) ${}^{TS[2]-i50}_{NS1}F^{M2}_{a}$	17
3	0.964	0.951	0.978	-0.419	0.501	151.1	**logK** = –4.523 (±0.414) + 1.067 (±0.561) ${}^{PN}_{SS5}F^{M4}_{a}$ – 0.109 (±0.014) ${}^{AC[7]-K}_{NS7}B^{M2}_{h-c}$ – 0.011 (±0.003) ${}^{TS[2]-i50}_{NS1}F^{M2}_{a}$	18
4	0.985	0.978	0.971	-0.645	0.533	264.9	**logK** = –1.2719 (±0.4235) – 3.1833 (±0.5336) ${}^{SD}_{SS4}B^{M2}_{h-e}$ – 0.0956 (±0.0100) ${}^{AC[7]-K}_{NS7}B^{M2}_{h-c}$ – 0.0004 (±0.0001) ${}^{N1}_{NS2}Q^{M11}_{a}$ – 22.7123 (±6.9163) ${}^{i50}_{MP3}B^{M16}_{a-e}$	19

Tabla F.5: Parámetros estadísticos de los mejores modelos de dimensiones desde 2 hasta 4 variables utilizando los índices QuBiLs MIDAS Duplas, considerando las 21 primeras estructuras como conjunto de entrenamiento y las últimas 10 como conjunto de prueba

	R^2	Q^2loo	Q^2boot	a(Q^2)	$SDEP^{ext}$	F	Modelos	Ec.
2	0.911	0.885	0.885	-0.317	0.494	92.9	**logK** = –4.239 (±1.552) – 1.198 (±0.152) ${}^{RA}_{SS1}TrB^{M27}_{v-c}$ – 10.304 (±2.252) ${}^{VC}_{NS3}TrF^{M26(M14)}_{a}$	20
3	0.973	0.957	0.956	-0.418	0.540	209.4	**logK** = –12.370 (±0.516) + 1.408 (±0.097) ${}^{TS[1]-i50}_{MP3}TrQB^{M20(M4)}_{e-v}$ – 0.219 (±0.038) ${}^{TS[1]-K}_{NS1}Tr^{M20(M16)}_{e-v-c}$ – 7.936 (±1.339) ${}^{VC}_{NS3}TrF^{M26(M14)}_{a}$	21
4	0.986	0.976	0.973	-0.535	0.523	290.1	**logK** = –5.490 (±0.630) + 0.000 (±0.000) ${}^{PN}_{NS5}TrC^{M20(M11)}_{a}$ – 142.391 (±14.291) ${}^{SD}_{MP3}Tr^{M20(M12)}_{e-h-c}$ – 0.091 (±0.017) ${}^{AC[7]-K}_{SS5}Tr^{M19(M13)}_{a-h-c}$ – 38.183 (±12.881) ${}^{TS[6]-RA}_{MP1}TrF^{M28}_{e}$	22

Tabla F.6: Parámetros estadísticos de los mejores modelos de dimensiones desde 2 hasta 4 variables utilizando los índices QuBiLs MIDAS Ternas, considerando las 21 primeras estructuras como conjunto de entrenamiento y las últimas 10 como conjunto de prueba

	R^2	Q^2loo	Q^2boot	$a(Q^2)$	$SDEP^{ext}$	F	Modelos	Ec.
2	0.897	0.872	0.867	-0.353	0.380	79.2	**logK** = –23.785 (±1.410) + 0.048 (±0.009) ${}^{SD}_{MP7}QuCB^{M26(M3)}_{v-h}$ + 24.820 (±3.175) ${}^{ES-SD}_{MP2}QuQTr^{M26(M3)}_{m-e-p}$	23
3	0.952	0.924	0.921	-0.448	0.548	114.5	**logK** = –2.579 (±0.530) – 1.798 (±0.247) ${}^{RA}_{SS1}Qu^{M26(M16)}_{a-v-h-c}$ – 0.011 (±0.002) ${}^{TS[7]-i50}_{SS1}QuQTr^{M26(M8)}_{e-v-p}$ + 0.070 (±0.009) ${}^{AC[5]-K}_{NS2}QuCB^{M26(M16)}_{m-psa}$	24
4	0.970	0.948	0.938	-0.588	0.433	130.5	**logK** = –5.457 (±0.237) + 4.160 (±0.775) ${}^{i50}_{NS6}QuCB^{M25(M3)}_{e-v}$ – 0.00004 (±0.00) ${}^{i50}_{NS1}QuCB^{M26(M13)}_{a-v}$ – 0.056 (±0.013) ${}^{AC[6]-K}_{SS7}QuQd^{M29}_{p}$ – 8.143 (±1.353) ${}^{TS[5]-i50}_{MP1}Qu^{M26(M16)}_{m-psa-a-s}$	25

Tabla F.7: Parámetros estadísticos de los mejores modelos de dimensiones desde 2 hasta 4 variables utilizando los índices QuBiLs MIDAS Cuaternas, considerando las 21 primeras estructuras como conjunto de entrenamiento y las últimas 10 como conjunto de prueba

	R^2	Q^2loo	Q^2boot	$a(Q^2)$	$SDEP^{ext}$	F	Modelos	Ec.
2	0.886	0.853	0.851	-0.335	0.455	70.1	**logK** = –8.874 (±0.547) – 1.387 (±0.205) ${}^{TS[1]-i50}_{MP3}TrQB^{M20(M4)}_{e-v}$ – 0.074 (±0.020) ${}^{AC[6]-K}_{SS7}QuQd^{M29}_{p}$	26
3	0.939	0.901	0.898	-0.450	0.434	88.5	**logK** = –8.926 (±0.410) + 1.482 (±0.155) ${}^{TS[1]-i50}_{MP3}TrQB^{M20(M4)}_{e-v}$ – 0.235 (±0.060) ${}^{TS[1]-K}_{NS1}Tr^{M20(M16)}_{e-v-c}$ – 0.042 (±0.017) ${}^{AC[6]-K}_{SS7}QuQd^{M29}_{p}$	27
4	0.975	0.962	0.956	-0.602	0.585	161.2	**logK** = –4.580 (±0.232) – 0.644 (±0.16962) ${}^{TS[3]-K}_{SS2}Tr^{M20(M12)}_{a-v-h}$ – 0.00003 (±0.000) ${}^{i50}_{NS1}QuCB^{M26(M13)}_{a-v}$ – 0.077 (±0.009) ${}^{AC[6]-K}_{SS7}QuQd^{M29}_{p}$ – 0.063 (±0.018) ${}^{TS[6]-i50}_{SS4}QuQTr^{M25(M3)}_{m-r-p}$	28

Tabla F.8: Parámetros estadísticos de los mejores modelos de dimensiones desde 2 hasta 4 variables utilizando los índices QuBiLs MIDAS Nuplas, considerando las 21 primeras estructuras como conjunto de entrenamiento y las últimas 10 como conjunto de prueba