UNIVERSITAT POLITÈCNICA DE VALÈNCIA

INSTITUTO DE TECNOLOGÍA QUÍMICA

Aplicación de la topología molecular a la predicción del factor de bioconcentración de un grupo heterogéneo de compuestos orgánicos

Master Interuniversitario en Química Sostenible

Trabajo de fin de máster presentado por

Raffaele Rea

Directores

Dr.D. JORGE GÁLVEZ ÁLVAREZ

Dr.D. RAMÓN GARCÍA DOMENECH

Tutor

Dr.D. ANTONIO EDUARDO PALOMARES GIMENO

Valencia, Julio 2013

JORGE GÁLVEZ ÁLVAREZ, Catedrático de Química Física del Departamento de Química Física de la Universitat de Valencia y ***RAMÓN GARCÍA DOMENECH***, Catedrático de Química Física del Departamento de Química Física de la Universitat de Valencia

CERTIFICAN:

Que el presente trabajo de fin del Máster de Química Sostenible, que lleva por título "***APLICACIÓN DE LA TOPOLOGÍA MOLECULAR A LA PREDICCIÓN DEL FACTOR DE BIOCONCENTRACIÓN DE UN GRUPO DE COMPUESTOS ORGÁNICOS***" ha sido realizado en el Departamento de Química Física de la Universitat de Valencia por ***RAFFAELE REA***, licenciado en Ingeniería Química por la Universitá di Roma "La Sapienza", y para que así conste a efectos oportunos, expiden y firman la siguiente autorización.

Burjassot, julio 2013

Fdo: Jorge Gálvez Álvarez **Fdo: Ramón García Domenech**

Tanto más fatiga el bien deseado cuanto más cerca está la esperanza de poseerlo.

Miguel de Cervantes (1547-1616)

Un agradecimiento a los profesores Dr. Jorge Gálvez Álvarez y Dr. Ramón García Domenech por la realización de este trabajo de fin de master.

Gracias también a Oksana por su presencia y comprensión.

Un saludo va a todos los amigos que he tenido el placer de conocer en este master “internacional” que han trasformado esta experiencia de estudio en algo más.

Índice

1. INTRODUCCIÓN

Miles de productos químicos se liberan en el medioambiente cada año durante su manufactura, explotación y eliminación. Las vías de contaminación del medioambiente son el aire, el suelo y el agua.

Cuando los productos químicos son liberados en el ambiente acuático, los organismos y las plantas son expuestos a estas sustancias que en la mayoría de los casos no son biocompatibles. Los organismos acuáticos viven gracias a las membranas de respiración que les permiten filtrar el oxígeno presente en el agua. Si además del oxígeno están presentes xenobióticos hay la posibilidad de bioacumulación.

Cuando los productos químicos se bioacumulan en los peces, estas sustancias alcanzan una concentración en el cuerpo del organismo aceptor mucho mayor respecto a la concentración presente en el mismo ambiente contaminado. Una elevada concentración de los productos químicos en la base de la cadena de alimentación puede causar efectos tóxicos en la parte superior de la misma, hasta llegar al ser humano. Por estas razones, a partir de los años 60, los gobiernos de varios países empiezan a catalizar su atención sobre como evaluar el riesgo potencial de bioacumulación de los productos químicos en el pescado [1,2].

Varios reglamentos se han elaborado para medir correctamente el potencial de bioacumulación de una sustancia química. El término "bioacumulación endpoint" no se refiere necesariamente al mismo término científico cuando nos referimos a los reglamentos. A la hora de hablar de bioacumulación en términos de variables a medir ("endpoints"), por la reglamentación europea significa en realidad evaluar el factor de bioconcentración. En la tabla 1.1 se ven los principales criterios de evaluación utilizados por las agencias reguladoras de diferentes países.

Tabla 1.1 Visión general de los "endpoints" de bioacumulación reglamentada y de sus criterios de evaluación [3].

Regulatory agency	Bioaccumulation endpoint	Criteria (log values)	Program
Environment Canada	K_{OW}	≥100 000 (5)	CEPA (1999)*
Environment Canada	BCF	≥5 000 (3.7)	CEPA (1999)
Environment Canada	BAF	≥5 000 (3.7)	CEPA (1999)
European Union 'bioaccumulative'	BCF	≥2 000 (3.3)	REACH†
European Union 'very bioaccumulative'	BCF	≥5 000 (3.7)	REACH
United States 'bioaccumulative'	BCF	1 000 (3)–5 000 (3.7)	TSCA‡, TRI
United States 'very bioaccumulative'	BCF	≥5 000 (3.7)	TSCA, TRI
United Nations Environment Programme	K_{OW}	≥100 000 (5)	Stockholm Convention§
United Nations Environment Programme	BCF	≥5 000 (3.7)	Stockholm Convention

*CEPA, Canadian Environmental Protection Act, 1999 (Government of Canada 1999; Government of Canada 2000).
†Registration, Evaluation and Authorization of Chemicals (REACH) Annex XII (European Commission 2001).
‡Currently being used by the US Environmental Protection Agency in its Toxic Substances Control Act (TSCA) and Toxic Release Inventory (TRI) programs (USEPA 1976).
§Stockholm Convention on Persistent Organic Pollutants (UNEP 2001).

En este trabajo fin de Master seguiremos, como criterio de evaluación de bioacumulación, el programa europeo REACH. Es notorio que el factor tenido en cuenta en la mayoría de las agencias de reglamentación para evaluar la bioacumulación de los productos químicos es el BCF, que corresponde al factor de bioconcentración. La única excepción es la agencia

medioambiental canadiense que tiene en cuenta el BAF (factor de bioacumulación) como parámetro para medir la bioacumulación. Para aclarar la diferencia entre ambos términos es necesario definirlos, lo que haremos a continuación [3].

1.1 Bioconcentración

La Bioconcentración es el proceso con el que una sustancia química es absorbida por un organismo desde el ambiente solo a través de la superficie dérmica y de la membrana de respiración (branquias).
A la vez, la bioconcentración es el resultado neto de las velocidades de absorción de una sustancia química particular absorbida a través de las superficies de respiración y la eliminación de la misma por intercambio respiratorio, defecación, biotransformación metabólica del compuesto original y dilución por crecimiento del organismo involucrado.
La dilución por crecimiento se considera un mecanismo de "pseudo-eliminación" porque el compuesto químico en realidad no se elimina desde el organismo pero su concentración disminuye a causa de un incremento del volumen del tejido.
El grado en que se produce la bioconcentración se expresa como factor de bioconcentración (BCF) y puede ser medida en condiciones controladas de laboratorio sin considerar el aporte del compuesto químico a través de la dieta.
Si consideramos el organismo acuático como una unidad en la cual el compuesto químico esta homogéneamente mezclado podemos escribir este balance de materia:

$$\frac{dC_B}{dt} = (k_1 C_{WD}) - (k_2 + k_E + k_M + k_G) C_B \qquad \text{Ec. (1)}$$

Donde C_B es la concentración de la sustancia química en el organismo (g/Kg), k_1 es la constante de la velocidad de absorción del producto químico desde el agua a la superficie respiratoria (L/Kg·d), C_{WD} es la concentración de la sustancia química en el agua (g/L), y k_2, k_E, k_M, k_G son las constantes de velocidad de la eliminación del producto químico a través del intercambio respiratorio, defecación, biotransformación metabólica del compuesto original, y dilución por crecimiento.
Cuando ambas concentraciones C_B y C_{WD} son constantes durante el tiempo de la exposición, se dice que el sistema ha alcanzado el estado estacionario ($dC_B/dt = 0$) y la ecuación (1) se transforma en:

$$BCF = C_B/C_{WD} = k_1/(k_2 + k_E + k_M + k_G) \qquad \text{Ec. (2)}$$

Esto significa que el factor de bioconcentración puede ser calculado simplemente como la relación entre la concentración de una sustancia química dada en un organismo acuático y la concentración de la misma en el agua en estado estacionario. Normalmente en el cálculo del BCF, en el denominador aparece el termino C_{WT}, que es la concentración total de la sustancia química en el ambiente acuático sea en forma libre o enlazada a partículas o materias orgánicas. Se cree que las sustancias químicas asociadas a materia orgánica no pueden pasar la membrana respiratoria, por esto para tener en cuenta solo la concentración de la sustancia pura en un medio acuoso se define un coeficiente φ = C_{WD}/C_{WT} , que es la fracción del soluto

biodisponible. Además, el peso del organismo contaminado puede ser expresado en peso húmedo (WW), peso seco (DW) o en contenido de lípidos (LW). El BCF que normalmente se calcula teniendo en cuenta el peso húmedo de la muestra analizada y que se expresa en L/Kg, a veces puede ser normalizado en base al contenido de lípidos, en este caso la formula se escribe $BCF_{LW} = BCF_{WW}/\%\ lipidos$ [3].
En esta tesis el BCF utilizado en la base de datos, se expresa en peso húmedo y a la minina concentración de exposición al contaminante [4].

1.2 Bioacumulación

La Bioacumulación es el proceso por el que una sustancia química es absorbida por un organismo desde el ambiente a través de todas las posibles vías de exposición presentes en el entorno natural, (fig.1.1).

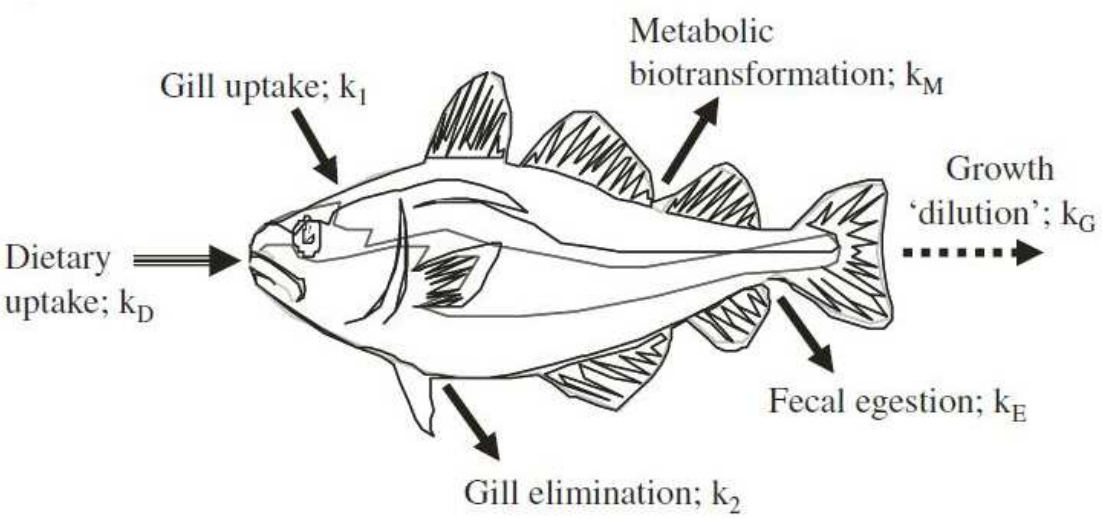


Fig.1.1 Constantes de absorción y de eliminación asociadas a las principales rutas de exposición de una sustancia química en pescado [3].

A diferencia de la bioconcentración tenemos un término adjunto que es la absorción de la sustancia química a través de la dieta.
Entonces la fórmula que describe la bioacumulación es igual a:

$$\frac{dC_B}{dt} = (k_1 C_{WD} + k_D C_D) - (k_2 + k_E + k_M + k_G) C_B \qquad \text{Ec. (3)}$$

En el estado estacionario, la ecuación (3) se simplifica a:

$$BAF = C_B/C_{WD} = [k_1 + k_D(C_D/C_{WD})]/(k_2 + k_E + k_M + k_G) \qquad \text{Ec. (4)}$$

Donde k_D y C_{WD} representan respectivamente la constante de velocidad de absorción de la sustancia química en la dieta (Kg/Kg·d) y la concentración de la misma en la dieta (g /Kg).
Está claro que la diferencia entre la bioconcentración y la bioacumulación está en el numerador que en el segundo caso es mayor respecto al primero.
También la bioacumulación se mide por unidad de masa húmeda de muestra y es expresa como L/Kg [3].

1.3 Test "in silico"

Con el paso de los años ha ido creciendo, en todo el mundo, la atención a las adecuadas medidas de gestión de las sustancia identificadas con la sigla PBT (persistente, bioacumulable y tóxica). Para identificar las sustancias químicas pertenecientes al amplio grupo de los PBT, se han desarrollado criterios y métodos precisos para estudiar los parámetros oportunos que permiten describir la persistencia, la bioacumulación y la toxicidad. Los criterios para evaluar el factor de bioconcentración, parámetro fundamental para determinar la bioacumulación de una sustancia, han sido analizados en el apartado 1.1 y 1.2.
Los criterios vistos hasta ahora, en realidad son solo una pequeña parte de la línea directriz adoptada de la OECD (Organization for Economic Cooperation and Development) a la cual casi todos los gobiernos mundiales hacen referencia.
En realidad todos los tests para evaluar la bioconcentración en pescado tienen que seguir las normas del test 305 de la OECD (Bioconcentration: flow-through fish test).
Los costes de estos tipos de test son muy elevados y necesitan de mucho tiempo. Basta pensar que para calcular el factor de bioconcentración de una sola sustancia química se necesitan alrededor de 125,000 dólares para tener resultados químicamente validos en un plazo de tiempo que supera los 40 días [5].
Además la reproducibilidad de los tests es muy complicada porque se deben tener en cuenta una cantidad de parámetros muy diferentes tales como la magnitud, la edad y el sexo de la especie de pez sometida a la prueba, la estabilidad de la sustancia química en el agua, la presencia de surfactantes, el pH y la capacidad tampón, la dureza del agua, la presencia de materia orgánica en suspensión y los efectos de un co-soluto.
En este contexto complejo e internacional, el 18 diciembre de 2006 el parlamento europeo adopta una legislación para una nueva gestión de las sustancias químicas llamada REACH (Registration, Evaluation, Authorization and Restriction of Chemicals).
El reglamento REACH entra en vigor el 1 de junio de 2007 y uno de los tantos objetivos es la eficiente identificación, evaluación y regulación de PTB de cada sustancia.
A este fin, proporciona criterios (en el anexo XIII del texto legal) para la identificación de PBT y sustancias químicas muy persistentes y muy bioacumulables (vPvB).
Además, el Anexo XI del Reglamento REACH prevé la utilización de tests validos del tipo relaciones cuantitativas estructura-actividad (QSAR), para predecir las propiedades ambientales y toxicológicas de las sustancias químicas, para ahorrar tiempo, mejorar la relación coste-efectividad y proteger el bienestar animal.
Por lo tanto, un mayor uso de los modelos (QSAR) está previsto para la evaluación de peligros y riesgos de los productos químicos en la Unión Europea (UE) [6].

Entre los métodos QSAR se encuentran aquellos que emplean la topología molecular y que han sido los utilizados en el trabajo descrito en esta memoria. Vamos pues a introducir la topología molecular en el contexto QSAR.

1.4 Topología molecular

Los métodos QSAR (Quantitative Structure Activity Relationships), relacionan la estructura de una molécula con la actividad de la misma, gracias a la caracterización numérica de la estructura molecular. Las diferencias entre estructuras se detectan a través de una cuantificación numérica que permite distinguir unas estructuras de otras.
En el caso de la topología molecular, esta caracterización numérica se realiza a través de unos índices o descriptores topológicos, que están basados en la teoría de los grafos. Esta teoría fue desarrollada principalmente por el matemático inglés Arthur Cayley [7].
Por su parte, la topología es aquella parte del algebra que estudia las posiciones e interconexiones entre los elementos de un conjunto [8]. La aplicación de la topología a la química se llama "Topología Molecular" que tiene como objetivo analizar las posiciones y las interconexiones entre los átomos de una molécula dada.
Mediante la topología molecular se puede describir en modo único la estructura molecular de cada compuesto. Ya que la estructura molecular es responsable de la propiedad de la molécula podemos, a partir de una molécula dada, diferenciar entre la forma (estructura) y la función (propiedades) [9].
El papel de la topología molecular es convertir la formula estructural en valores numéricos o índices que por definición llevan consigo la información estructural. Esto nos permite manipular las estructuras moleculares como números, de modo que es posible comparar, clasificar y almacenar las estructuras entre si y además correlacionarlas con sus propiedades. Ya que la forma de una molécula es modelada como expresión matemática o como número, existen distintos modelos que pueden describir la misma molécula [10].
Tal aproximación expresa la estructura topológica de la molécula y puede ser utilizada como la base de una caracterización cuantitativa de la conectividad de la estructura. Mediante esta cuantificación es posible obtener una correlación entre las moléculas y sus propiedades físicas, químicas y biológicas. Este es el objetivo del método de la topología molecular. Se trata de obtener ecuaciones de regresión multilineal entre dichas propiedades y los descriptores topológicos, llamadas funciones topológicas:

$$P(IT) = A_0 + \sum_{i=1}^{m} A_i\,(IT)_i \qquad \text{Ec. (5)}$$

Donde A_0 y el conjunto de términos A_i, son los coeficientes de regresión de la ecuación obtenida e $(IT)_i$ representa cada uno de los descriptores moleculares.

Cada ecuación de regresión se acompaña de una serie de parámetros estadísticos:

N = número de moléculas empleadas en la regresión.

r = coeficiente de correlación.

EEE = error estándar de estimación de la ecuación de regresión.

F = parámetro de Fisher-Snedecor.

p = significación estadística de la regresión.

La utilidad de llegar a una expresión matemática función de los índices topológicos consiste en calcular los valores de los índices $(IT)_i$ para moléculas no utilizadas en la correlación y después sustituir estos valores en la ecuación seleccionada. Así se puede predecir con una cierta precisión el valor teórico de la propiedad "P" para ese compuesto.

Para poder calcular los índices topológicos, primero hemos de representar la molécula mediante un grafo. Los fundamentos de la teoría de grafos aplicada a la química se describen a continuación [11].

1.4.1 Introducción a la teoría de grafos

La teoría de grafos es un área importante de la matemática aplicada que encuentra empleos en prácticamente todas las ramas de la ciencia (redes de comunicación, circuitos eléctricos, optimización de vías de comunicación, etc).
También el análisis topológico de una molécula utiliza la teoría de los grafos para abstraer el concepto de estructura química. Un grafo representa las interconexiones de los elementos dentro de un conjunto, y en el caso de una molécula las conexiones son los enlaces químicos entre sus átomos [12,13].
Aunque el grafo, como entidad química, fue introducido en el siglo XIX, sus principales impulsores fueron Randić [14] y Kier-Hall [15,16] entre los años 70 y 80 del siglo XX, los cuales aplicaron los conceptos matemáticos en el entorno químico. El sentido de esta serie de artículos es que un grafo puede ser asociado con una estructura química ya que los vértices (los puntos de un grafo) son asociables a los átomos y las aristas (los segmentos que unen los puntos) a los enlaces químicos.
Un grafo químico no es pues un simple ordenamiento de puntos y segmentos sino un diagrama de la estructura molecular a través de su topología [9].
Lo que importa en el análisis topológico es el camino para ir de un átomo a otro dentro de la molécula, sin necesidad de estudiar la morfología tridimensional real de la molécula, la naturaleza y la longitud de los enlaces químicos que ligan los átomos o los ángulos entre dichos enlaces.

Como antes señalábamos, el primer paso para obtener el grafo de una molécula es representar a los átomos por puntos (que se conocen como "vértices") y a los enlaces por segmentos (a los que se denomina "aristas"), eliminándose los átomos de hidrógeno. En este caso se habla de "grafo de hidrógenos suprimidos"**,** porque tenemos solo el esqueleto carbonado de una molécula y los átomos de hidrógenos se presuponen en función de la valencia del carbono.

Una vez obtenido el grafo se numeran aleatoriamente los distintos vértices y se construye la "matriz topológica" o "matriz de adyacencia". La primera persona que demostró que es posible representar una molécula por una matriz fue el matemático Sylvester [17] en el siglo XIX (Sylvester, 1874).

Los elementos $a_{ij,}$ de la matriz de adyacencia **A**, toman el valor (1) o (0) dependiendo de que el átomo "i" esté enlazado al "j" o no. En el caso de que la molécula posea enlaces múltiples, el correspondiente termino a_{ij} de la matriz será la multiplicidad del enlace (por ejemplo, en un

doble enlace será 2, en uno triple 3, etc.). Esta matriz cuadrada de *n* filas por *n* columnas, siendo *n* el número de vértices del grafo, es simétrica respecto a su diagonal principal [18].

La valencia de cada vértice, δ_{ij}, es igual a la suma de los valores que hay en la fila o columna correspondientes a dicho vértice o lo que es lo mismo, el número de aristas que llegan a él, es decir:

$$\delta_{ij} = \sum_{j=1}^{m} a_{ij} \qquad \text{Ec. (6)}$$

La letra *m* se emplea para designar el número total de aristas que aparecen en el grafo.

Como se aprecia, es un tipo de expresión matemática que relaciona la estructura química mediante una descripción numérica. Su importancia en los estudios estructura-actividad radica en que a partir de ella se calculan la mayoría de los índices topológicos, que son la base del presente trabajo.
A título de ejemplo la figura 1.2 enseña la construcción de la matriz topológica para el isopentano y el ciclohexano, junto con los valores obtenidos para las valencias de los diferentes átomos [11].

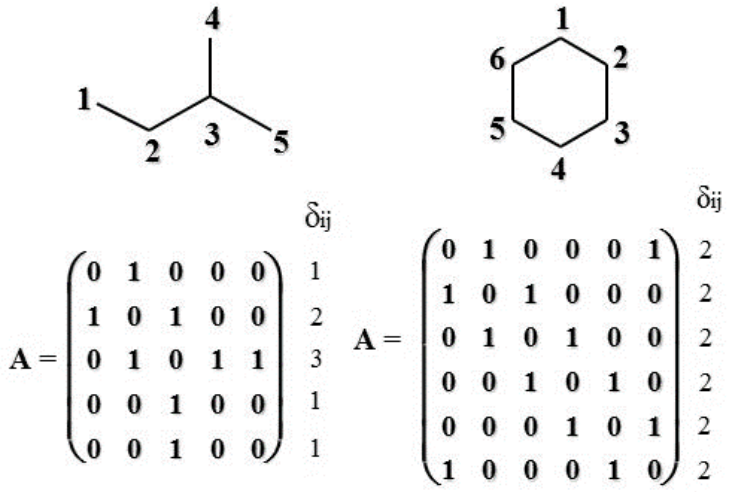
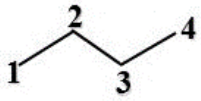


Figura 1. 2 Representación de estructuras como grafos y matrices.

1.4.2 Índices topológicos

Una vez representado el grafo de una molécula y construida su matriz topológica, es posible calcular los índices topológicos por medio de algoritmos adecuados.
Los índices topológicos son descriptores numéricos de la estructura molecular, que describen diferentes aspectos de la molécula como los tipos de átomos, enlaces y ensamblaje topológico de la misma. Un aspecto importante de los descriptores topológicos es que son, por definición,

"invariantes de grafo", es decir que su valor debe ser independiente del orden de numeración de los vértices del grafo.
El primer índice topológico, para la caracterización de las ramificaciones de una molécula, fue propuesto por Wiener en 1947 [19].
El índice de Wiener (W) introduce el concepto topológico de "distancia", es decir el camino más corto entre dos vértices, entendido como el número mínimo de aristas capaces de conectar esos dos vértices. Este índice se determina a partir de la "matriz de distancia" D de Hosoya [11].
Sirva como ejemplo ilustrativo la construcción de la matriz de distancia y el cálculo de W para la molécula de butano.



$$\mathbf{D} = \begin{pmatrix} \mathbf{0} & \mathbf{1} & \mathbf{2} & \mathbf{3} \\ \mathbf{1} & \mathbf{0} & \mathbf{1} & \mathbf{2} \\ \mathbf{2} & \mathbf{1} & \mathbf{0} & \mathbf{1} \\ \mathbf{3} & \mathbf{2} & \mathbf{1} & \mathbf{0} \end{pmatrix} \begin{matrix} \delta_i \\ 6 \\ 4 \\ 4 \\ 6 \end{matrix}$$

$$W = {}^{1}/_{2} \sum_{i}^{m} \delta i = \frac{6+4+4+6}{2} = 10$$

Figura 1.3 Obtención del número de Wiener para el n-butano

Para obtener esta matriz se tiene en cuenta que el término a_{ij} de la matriz es igual al número de aristas existentes entre el vértice "i" y el vértice "j" por el camino más corto. Al igual que la matriz topológica, la matriz de distancia es una matriz cuadrada de "n" filas por "n" columnas, donde "n" es el número de vértices del grafo, [20] siendo también simétrica respecto su diagonal principal. En este caso la valencia δ_i de cada vértice es la suma del número de aristas existentes entre el propio vértice "i" y cada uno "j" del resto de vértices, lo que también es igual a la suma de los términos que hay en la fila o columna correspondiente a dicho vértice. El índice de Wiener será igual a la mitad de la suma de la valencia de cada vértice.

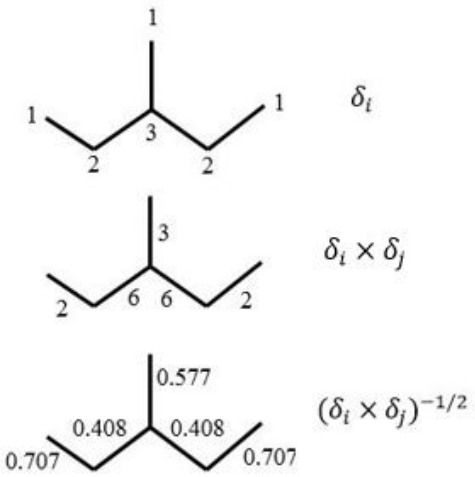
El índice de Wiener muestra muy buenas correlaciones cuando se trata de simular propiedades sencillas como la temperatura de ebullición [19]. Además se comprobó en los años siguientes su eficacia para predecir otras propiedades más complejas, como por ejemplo las energías de enlaces de hidrocarburos policíclicos aromáticos. Esto permitía conocer las frecuencias de aparición de bandas UV-visible de compuestos aun no sintetizados.

Al mencionar la teoría de los grafos, es imprescindible hablar de "índices de conectividad molecular", introducidos por Kier y Hall [15], como generalización del índice de Randić [14]. El índice de Randić se describe por la siguiente ecuación:

$$\chi = \sum_i \sum_j \left(\delta_i \delta_j\right)^{-1/2} \qquad \text{Ec. (7)}$$

O bien es la suma total de los inversos de las raíces cuadradas de los productos de las valencias de los dos vértices adyacentes que limitan cada arista del grafo.

A título de ejemplo se calcula el índice de Randić para una molécula de isopentano.



$$\chi = \sum_i \sum_j \left(\delta_i \delta_j\right)^{-1/2} = 2.807$$

Figura 1.4 Cálculo del índice de ramificación de Randić para el 3-metil-pentano

Si queremos definir los índices de conectividad molecular [11,15], es necesario introducir el concepto de subgrafo, entendiendo por tal, cualquier parte de un grafo constituido por varias de sus aristas interconectadas. Los subgrafos se clasifican según su orden (m) y su tipo (t). El orden de un subgrafo no es más que el número de aristas que contiene, sin distinguir entre enlaces múltiples y simples.
El orden de cada uno de los subgrafos depende del número de enlaces que los formen, así, varían desde orden cero para cada uno de los átomos aislados que constituyen la molécula hasta el orden máximo de la molécula que se atribuye al grafo completo. Los subgrafos se clasifican en cuatro tipos:

- Tipo I o tipo PATH (t = p), que son aquellos subgrafos en los que las valencias de sus vértices son menores o iguales a 2, sin formar un ciclo
- Tipo II o tipo CLUSTER (t = c), constituido por aquellos subgrafos que tienen al menos algún vértice con valencia 3 o 4, pero ninguno con valencia 2, siempre que no formen un ciclo.

- Tipo III o tipo PATH-CLUSTER (t = pc). Son los subgrafos que incluyen vértices con valencias 2, además de algunos con valores 3 o 4, sin formar un ciclo.
- Tipo IV o tipo CHAIN (t = ch), formado por secuencias de enlaces conteniendo al menos un ciclo.

La figura 1.4 muestra los distintos subgrafos del isopentano.

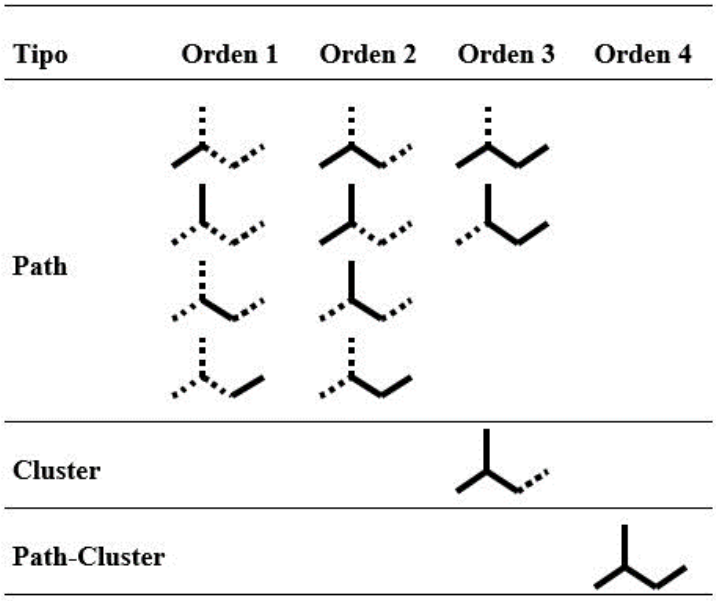


Figura 1. 4 Clasificación de los subgrafos del isopentano (trazos continuos)

Los índices de conectividad simples o de grafo, ${}^{m}\chi_{t}$, corresponden en cada caso a la suma, para todos los subgrafos de tipo *t* y orden *m*, a la siguiente expresión:

$$ {}^{m}\chi_{t} = \sum_{j=1}^{n_m} {}^{m}S_{j} \qquad \text{Ec. (8)} $$

donde n_m es el número de subgrafo de tipo t y orden m. Los términos ${}^{m}S_{j}$ se definen como el inverso de la raíz cuadrada del producto de las valencias de cada uno de los vértices del subgrafo y vienen representados según la expresión:

$$ {}^{m}S_{j} = \left[\prod_{i=1}^{m+1} \delta_{i} \right]^{-1/2} \qquad \text{Ec. (9)} $$

donde j define cada uno de los subgrafos. El número de valencias a multiplicar δ_i depende del tipo de subgrafo. Por ejemplo, los subgrafos tipo "chain" están definidos por *m* vértices como máximo, mientras que los restantes tipos están definidos por *m*+1 vértices, siendo *m* el orden del subgrafo. Es decir, los índices de conectividad se obtienen como suma, para todos los subgrafos de un mismo tipo, de los inversos de las raíces cuadradas de los productos de las valencias de los vértices adyacentes que forman parte de cada subgrafo.

Si sustituimos el término ${}^{m}S_j$ de la ecuación (9) en la ecuación (8), para subgrafos de "m" enlaces, es decir de orden m, obtenemos:

$$ {}^{m}\chi = \sum_{s=1}^{n} \left\{ \left[\prod_{i=1}^{m+1} (\delta_i) \right]^{-1/2} \right\}_s \qquad \text{Ec. (10)} $$

donde "n" es el subgrafo de orden "m" con m+1 vértices, δi es la valencia de cada vértice y "s" representa un subgrafo en particular. Desde esta ecuación se deduce que la serie completa de índices es una caracterización única de la estructura química, es decir que no se repiten en su conjunto. La especificidad para casa estructura química, hace de la conectividad molecular uno de los mejores métodos para caracterizar una molécula en particular.

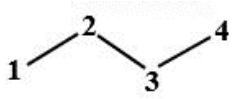
Hasta ahora, hemos considerado a las moléculas sin tener en cuenta los enlaces múltiples y los heteroátomos. Para tener esto en consideración hay que introducir algunas modificaciones en la matriz de adyacencia. Así, los *índices de conectividad de valencia* ${}^{m}\chi_t^{v}$ definen las moléculas de modo idéntico a los anteriores pero sustituyen el grado topológico δ por un grado químico o de valencia definido como:

$$ \delta^{v} = (Z^{v} - H) / Z - Z^{v} - 1 \qquad \text{Ec. (11)} $$

Donde Z es el número atómico, Z^v el número de electrones de valencia del átomo considerado y H el número de átomos de hidrógeno a los que se une dicho átomo. Como se comprueba fácilmente, para el caso del carbono ambos grados coinciden siempre.

Otros tipos de índices son los que poseen información electrónica como las diferencias entre los índices de valencia y no valencia descritos por Kier y Hall y los índices topológicos de carga introducidos por la Unidad de Investigación de Conectividad Molecular y Diseño de Fármacos de la Universidad de Valencia [20].

Estos índices de carga, G_k y J_k , describen la distribución global de la carga molecular mediante la evaluación de la transferencia de carga entre pares de átomos. Para definir estos índices se introduce la matriz de términos de carga **C**. En la siguiente figura se ve como se obtiene dicha matriz para una molécula de n-butano:



$$\mathbf{A}=\begin{pmatrix}0&1&0&0\\1&0&1&0\\0&1&0&1\\0&0&1&0\end{pmatrix} \qquad \mathbf{D}=\begin{pmatrix}0&1&2&3\\1&0&1&2\\2&1&0&1\\3&2&1&0\end{pmatrix} \qquad \mathbf{D}^{*}=\begin{pmatrix}0&1&1/4&1/9\\1&0&1&1/4\\1/4&1&0&1\\1/9&1/4&1&0\end{pmatrix}$$

$$\mathbf{M}=\mathbf{A}\times\mathbf{D}^{*}=\begin{pmatrix}1&0&1&1/4\\1/4&2&1/4&10/9\\10/9&1/4&2&1/4\\1/4&1&0&1\end{pmatrix} \qquad \mathbf{C}=\begin{pmatrix}0&-1/4&-1/9&0\\1/4&2&0&1/9\\1/9&0&2&1/4\\0&-1/9&-1/4&1\end{pmatrix}$$

Figura 1.6 Calculo de la matriz de carga "C" para el n-butano

Donde **A** es la matriz de adyacencia, **D** la matriz de distancia, **D*** es la matriz inversa del cuadrado de la distancia, puesto que la influencia de la carga decrece con el cuadrado de la distancia, también conocida matriz coulombiana [11].

Esta matriz ya es el primer descriptor de carga ya que tiene como elementos q_{ij} el valor de la inversa del cuadrado de la distancia topológica (número de aristas) entre los vértices *i* y *j*.

Se define la matriz **M** como el producto de la matriz de adyacencia **A** por la matriz coulombiana **D***, (**M** = **A** X **D***). La matriz de términos de carga **C** se obtiene a partir de la matriz **M**, quedando definidos los elementos c_{ij} de la misma como:

$$c_{ij} = m_{ij} - m_{ji} \text{ (para } i \neq j)$$

$$c_{ij} = \text{valencia del vértice (para } i = j)$$

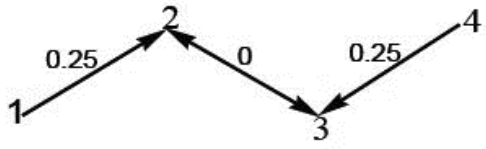
Donde m_{ij} representa el término de posición *ij* en la matriz **M**. Para i=j el término c_{ij} representa la valencia topológica del vértice, que coincide con el valor de los elementos de la diagonal principal de la matriz **M**. Para i≠j el término c_{ij} representa una medida de la carga neta transferida desde el átomo *j* al átomo *i*, por ello cuando este término toma valor negativo será el átomo *i* el que transfiere carga al átomo *j*.

Por fin podemos calcular los índices topológicos de carga G_k y J_k de orden k, para un grafo dado, según las ecuaciones:

$$G_k = \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \left| tc_{ij} \right| \delta(k, d_{ij}) \qquad \text{Ec. (12)}$$

$$J_k = \frac{G_k}{(N-1)} \qquad \text{Ec. (13)}$$

donde N es el número de vértices en el grafo, N - 1 el número de aristas para compuestos no cíclicos, c_{ij} es el término de carga entre los vértices i y j y d(k,d_{ij}) es la función de Kronecker, igual a 1 si k=d_{ij} y 0 en el resto de los casos. Estos nuevos descriptores evalúan el total de la carga transferida entre los átomos situados a una distancia topológica *m.* A título de ejemplo calculamos el valor de los dichos índices para la molécula de n-butano ilustrado según topología como un dígrafo dirigido:



En el caso del n-butano los pares de vértices situados a una distancia *m* = 1 son tres: (*v1*,*v2*), (*v2*,*v3*) y (*v3*,*v4*); los pares situados a distancia *m* = 2 son dos: (*v1*,*v3*) y (*v2*,*v4*). Por último, a distancia *m* = 3 sólo hay un par de vértices: (*v1*,*v4*), por tanto:

G_1= │c12 + c23+ c34│= 0,25 + 0 + 0,25 = 0,5

G_2= │c13 + c24│= 0,11 + 0,11 = 0,22

G_3= │c14 │= 0

También es posible definir los índices de carga ponderados **(**J_k) que representan el valor de la carga transferida por cada enlace o arista [20].

Se recuerda al lector que en este apartado introductorio solamente se han definido de forma breve aquellos índices topológicos de importancia histórica o especial significación. En cambio, en los apartados posteriores se tendrán en cuenta otros índices que se han utilizado también para el desarrollo del modelo QSAR.

2. OBJETIVOS DEL TRABAJO

Las herramientas adquiridas durante el curso de Aplicación de la Topología Molecular a la Química Sostenible y Medioambiental, durante los estudios del Máster en Química Sostenible, pueden ser utilizadas en el nuevo reglamento europeo de las sustancias químicas (REACH).

La política ambiental adoptada del reglamento REACH (articulo [13(1), 25(1), Anexo XI]) prevé la recogida de datos sobre el efecto toxico de una sustancia química, que pertenece a una cierta categoría, mediante técnicas informáticas de extrapolación, análisis de tendencias y metodologías (Q)SAR, dejando la experimentación animal como última opción [21].

En este contexto de innovación cultural y tecnológica surge este trabajo, cuyo objetivo principal es el desarrollo de un modelo "*in silico*" para predecir correctamente el factor de bioconcentración (BCF), parámetro de gran interés para evaluar la sostenibilidad ambiental de un compuesto químico.

La primera etapa del trabajo se centra en la obtención de una ecuación de predicción, construida mediante una serie de datos experimentales de un conjunto de moléculas muy heterogéneas. Posteriormente se divide el conjunto de moléculas en dos grandes categorías, clorados y no clorados, para afinar el modelo matemático.
Estas ecuaciones deben ser capaces de predecir si una molécula se bioacumula o no en los organismos de ensayos, que en el nuestro caso es una tipología de pescado, precisamente la carpa común. Una vez obtenidas todas las ecuaciones multilineales se construye por cada una de ellas, una red neuronal, empleando las mismas variables topológicas para una mejor comparación de los resultados.
El objetivo es comprobar si existen correlaciones más complejas y adecuadas para la predicción de un factor complejo como lo es la bioacumulación. Se recuerda al lector que en este contexto de la tesis, como ya se ha explicado en el apartado de introdución, el termino bioacumulación es un sinónimo de bioconcentración.

3. MATERIALES Y MÉTODOS

3.1 Base de datos e información experimental

La base de datos utilizadas en el estudio de predicción desarrollado en el presente trabajo ha sido extraído del artículo publicado por el profesor Chunyan Zhao y colaboradores [2] los cuales, hasta hoy en día, participan en el proyecto europeo CAESAR (Computer Assisted Evaluation of Industrial Chemical Substance According to Regulations).
La calidad de los datos experimentales está confirmada por el hecho de que la misma "*database*" fue empleada ya en los estudios de Dimitrov [4], que fue el primero que la construyó uniendo dos bases de datos japoneses, el MITI (Ministry of International Trade and Industry) y del NITE (National Institute of Techonology and Evaluation) a la hora de construir un modelo QSAR y comprobarlo con un test externo.
Los datos experimentales de logBCF se refieren a la misma especie de pescado, Cyprinos Carpio, vulgarmente conocido como carpa común, y son calculados a la mínima concentración de exposición al contaminante [4].
Las tablas con todas las moléculas completa del training y del test externo se encuentran en el material suplementario tablas S, en el CD adjunto al presente trabajo.
La base de datos se compone de 466 moléculas que tienen un intervalo de variación de logBCF que va de -1 a 4,85 y un rango de peso molecular de 63 a 943 g/mol.

3.2 Generación y selección de los índices topológicos

Para el cálculo de los índices topológicos se utilizó el programa Dragon [22] desarrollado en la Universidad de Milán por el profesor Roberto Todeschini y colaboradores.

El programa Dragon calcula los índices topológicos partiendo da los grafos con hidrógenos suprimido, por lo tanto, el primer paso es tener moléculas sin hidrógenos y sucesivamente guardarlas con el formato específico ".mol". Estos archivos se pueden obtener directamente desde la red (Search for Species Data by CAS Registry Number) o bien a través de alguno de los programas de dibujo de estructuras químicas, como el ChemDraw.
Una vez que se introduce el archivo comentado anteriormente, el software se encarga automáticamente de determinar las matrices topológicas correspondientes que permiten calcular los 449 descriptores topológicos.
Una vez obtenida la matriz de datos es necesario encontrar entre las 449 variables las que mejor describen la propiedad de bioconcentración que se quiere predecir.
En este trabajo la metodología de elaboración de datos, para encontrar los mejores índices, ha sido la de dividir el total de las 449 variables en nueve grupos de cincuenta y "filtrar" cada vez los índices que mejor describen la propiedad de bioconcentración. Después de la primera selección de los mejores se continúa según la misma metodología hasta llegar a una serie óptima de variables seleccionadas.

3.3 Construcción de los modelos predictivos: Regresiones multilineales y Redes neuronales artificiales

Para llevar a cabo las ecuaciones de regresiones multilineales (RML) se hizo uso del paquete informático STATISTICA, versión 8.0., usando como variable dependiente el factor de bioconcentración "LogBCF" de los compuestos y como variables independientes el resto de descriptores topológicos obtenidos a partir de DRAGON.
Los modelos realizados se refieren a tres escenarios diferentes:

- RML para todas las moléculas,
- RML para las moléculas con cloro,
- RML para las moléculas sin cloro.

El conjunto de datos, de cada modelo, ha sido dividido en un training y en un test set utilizando una función del software STATISTICA que permite crear un subconjunto de datos aleatorios. El 80% del conjunto de datos, de cada escenario, ha sido utilizado para la construcción del modelo (*training*) y el 20% para su validación (*test externo*). A continuación resumimos la partición inherente a cada modelo:

- Todas las moléculas training (N=381); test externo (N=85); total (N=466)
- Moléculas con cloro training (N=126); test externo (N=32); total (N=158)
- Moléculas sin cloro training (N=246); test externo (N=62); total (N=308)

Como se aprecia, todas las moléculas de la base de datos (N=466), han sido divididas en dos grupos en base a la presencia o no de cloro.
La robustez y la calidad predictiva de cada modelo seleccionado han sido comprobadas con un test de validación. Las estrategias que se suelen adoptar como test de validación son: *a)* Crosvalidación o validación cruzada; *b)* validación interna dividiendo la serie de compuestos estudiada en un grupo de entrenamiento y un grupo test; *c)* validación externa con un grupo

test no usado en la búsqueda de la función de predicción; *d)* análisis de aleatoriedad y estabilidad [23]. En este trabajo utilizaremos las estrategias (a, y c) como criterio de validación.

El test de validación interna o bien de crosvalidation tipo leave-one-out, es un estudio que consiste en eliminar un compuesto (y su correspondiente propiedad) del resto del conjunto y volver a realizar las correlaciones, utilizando esta vez como grupo de entrenamiento al conjunto de N-1 compuestos. Se utilizan como variables independientes los índices topológicos obtenidos inicialmente.

Una vez obtenido el nuevo modelo se predice la propiedad para el compuesto eliminado. Este proceso se repite para todos los compuestos del conjunto, obteniéndose una predicción para cada uno de ellos.

Si los valores predichos y los residuales, calculados con la crosvalidacion son similares a lo de la ecuación original el modelo se considerará estable.

La estabilidad del modelo se determina calculando el coeficiente de correlación de la validación cruzada, Q^2, definido como:

$$Q^2 = \frac{SD - PRESS}{SD}$$

donde:

- *SD* es la desviación al cuadrado de cada valor respecto de la media: $SD = \sum_{i=1}^{n}(y_i - \bar{y})^2$,
- *PRESS* es la suma de los residuales predichos al cuadrado: $PRESS = \sum_{i=1}^{n}(y_i - y_{i(vc)})^2$,
- y_i es el valor experimental de la variable dependiente del compuesto *i* ,
- $\bar{y}$ es el valor medio experimental de la variable dependiente,
- $y_{i(vc)}$ es valor predicho en la validación cruzada de la variable dependiente del compuesto *i.*

Se considera aceptable un valor de Q^2 mayor a 0,5, [24] si bien se ha demostrado que un valor adecuado de dicho coeficiente es condición necesaria pero no suficiente para afirmar la capacidad predictiva del modelo.

Por esta razón también se realiza un test de validación externo, utilizando la ecuación de regresión obtenida con el grupo de entrenamiento.

El test de validación externo se refiere al grado con que pueden generalizarse los resultados de un experimento. Es un test que supera el objetivo de demostrar las relaciones funcionales entre las variables independiente y dependiente.

Además se hace un análisis de los *outliers*, es decir las moléculas que superan dos veces el error estándar de estimación (EEE), con el objetivo de ver donde falla la predicción. Esta análisis se hace tanto para el "training" como para el "test externo".

Con el propósito de mejorar y comparar los resultados de las regresiones multilineales se han construido, con los mismos índices topológicos, tres redes neuronales para las tres categorías mencionadas anteriormente. Las Redes Neuronales Artificiales (RNA) se utilizan para encontrar relaciones no lineales entre la propiedad física objeto del estudio, en este caso "log(BCF)", y los descriptores moleculares empleados.

A diferencia de los algoritmos de regresión multilineales que se basan sobre el concepto de función matemática, las redes neuronales se basan sobre el concepto de inteligencia artificial. Brevemente, cuando hablamos de las Redes Neuronales Artificiales (RNA) nos referimos a unas abstracciones de las estructuras nerviosas biológicas (cerebros) que tienen la característica de ser sistemas desordenados capaces de guardar información [25].
Como una neurona biológica, una sola neurona artificial recibe una información desde el exterior que, después de su elaboración, viene enviada como señal de salida.
Cada neurona artificial trabaja independientemente e influye en el conjunto total de neuronas que forma la red neuronal. El concepto puede ser fácilmente esquematizado en la siguiente figura:

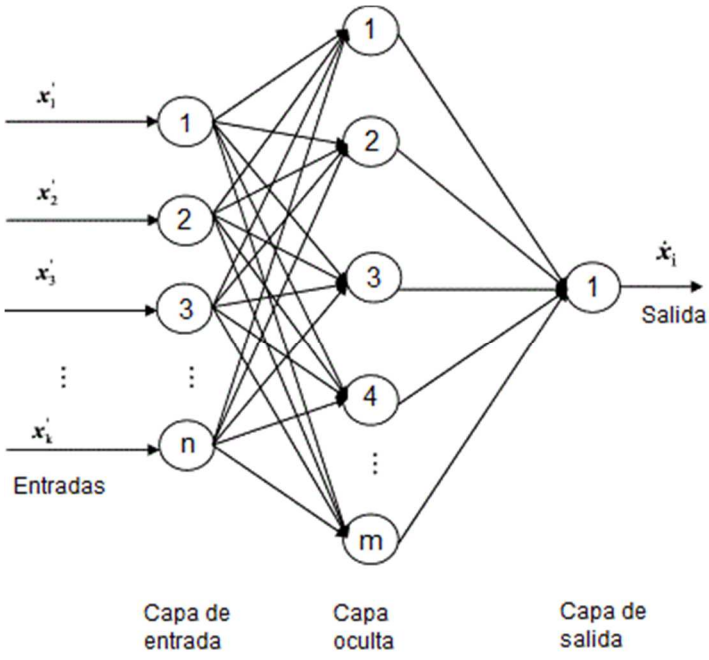


Figura 3.1 Esquema del funcionamiento de una red neuronal artificial

Como se aprecia de la imagen, una red neuronal tiene tres capas:

- La capa de entrada compuesta de 1…n neuronas,
- La capa oculta compuesta de 1…m neuronas,
- La capa de salida compuesta de 1 neurona.

Cada neurona de entrada procesa la información por su cuenta, de modo que si es excitada o inhibida da la señal externa, (x'_k), pero todas trabajan juntas para dar una única salida, ($\dot{x}_i$).
Entre la capas de entrada y de salida pueden haber muchas capas ocultas. Estas capas internas contienen muchas de las neuronas en diversas estructuras interconectadas. En nuestro caso, para todos los modelos, los datos empleados para elaborar las redes neuronales han sido divididos en:

- grupo training: formado por el 80% del total de moléculas,
- grupo test: formado por el 20% del total de moléculas.

Una vez obtenida la red neuronal se guarda en formado PMML script y se aplica a cada grupo externo de moléculas seleccionadas, ya utilizados en el análisis de regresión multilineal.
A continuación se comentan los resultados de los análisis.

4. RESULTADOS Y DISCUSIÓN

En el apartado anterior "Construcción de los modelos predictivos" se ha comentado como se han realizado los ensayos pertinentes a los tres escenarios diferentes:

- Predicción del factor de bioconcentración de todas las moléculas,
- Predicción del factor de bioconcentración de las moléculas con cloro,
- Predicción del factor de bioconcentración de las moléculas sin cloro.

A continuación se van analizando los resultados de cada una de las ecuaciones obtenidas. Los parámetros estadísticos que acompañan a cada ecuación y que son considerados al objeto de elegir el mejor modelo QSAR son:

- r^2: coeficiente de determinación
- EEE: error estándar de la estimación
- F: parámetro de Fisher-Snedecor
- p: significación estadística

La mejor ecuación predictiva se elige también considerando el número de índices topológicos, es decir que si al añadir una variable no produce una relevante mejora del coeficiente de determinación del grupo de entrenamiento, ($\Delta r^2 \leq 0{,}010$) la ecuación anterior es la mejor [1].
Debido a la gran cantidad de datos analizados se ha elegido poner las tablas de excesiva extensión en un CD adjunto a la copia impresa del proyecto. Todos los datos añadidos al CD y que se citan en el texto vienen indicados con la sigla "S" y un número romano entre paréntesis, para indicar correctamente la cronología del anexo. Por ejemplo la tabla S(I) es la primera tabla del anexo y en este caso se refiere a la base de datos.

4.1 Modelo de regresión multilineal para la predicción del factor de bioconcentración con todas las moléculas

Los resultados del análisis de regresión multilineal (RML), son resumidos en la siguiente tabla 4.1

Tabla 4.1 Análisis de regresión multilineal realizado con logBCF y diferentes números de variables

Variables	r^2	Δr^2
6	0,684	0
7	0,701	0,017
8	0,709	0,008

Ecuación de predicción seleccionada

	Coeficiente	EE	p
Intercepto	0,454	0,123	0,0003
nN	-0,273	0,045	0,0000
nO	-0,291	0,028	0,0000
nCL	0,142	0,024	0,0000
ATS6v	0,368	0,039	0,0000
MATS1v	0,609	0,083	0,0000
GATS6v	-0,317	0,065	0,0000
EEig03x	0,406	0,047	0,0000
N=381	r^2=0,701	Q^2=0,687	
F(7,373)=125	p=0,0000	EEE=0,725	

La Tabla 4.1 recoge los resultados estadísticos obtenidos en función del número de variables seleccionadas. La ecuación seleccionada tiene siete variables y ha sido elegida en base al principio de que la incorporación de una nueva variable apenas modifica el coeficiente de correlación múltiple ($\Delta r^2 \leq 0{,}010$). Para la agrupación de variables que mejor correlacionan el factor de bioconcentración tenemos un $r^2 = 0{,}701$ y un EEE = 0,725. Todas ellas son estadísticamente significativas con valores de $p \leq 0{,}0003$.

Los primeros tres índices topológicos que aparecen en la ecuación (tabla 4.1) evalúan la presencia de átomos de nitrógeno (nN), oxigeno (nO) y cloro (nCl) en las moléculas. Después encontramos los índices ATS6v, MATS1v y GATS6v, que son índices de autocorrelación y se refieren a como se distribuyen los volúmenes de Van der Waals entre las moléculas. Por ultimo hay el EEig03x, un índice puramente topológico.

El modelo seleccionado es capaz de explicar más del 70% de la varianza de la propiedad correlacionada ($r^2 = 0{,}701$) con un error estándar de estimación de cerca al 12,4% de la variabilidad en la que se mueve la propiedad (EEE = 0,725).

Las figuras 4.1 y 4.2 representan los datos de las columna 1 y 2 de la tabla S(II), y muestran los resultados de predicción obtenidos para cada compuesto. Todos ellos, a excepción de los compuestos marcados con puntos negros en la Figura 4.2, presentan residuales inferiores a ± 2EEE, lo cual es indicativo de la calidad de la ecuación seleccionada. Sobre un total de 381 moléculas tenemos 16 outliers es decir que el modelo no describe bien el 3,4% de los compuestos analizados. El primer estudio de validación de la ecuación seleccionada fue una crosvalidación interna. Para ello se elimina un caso del grupo y se realiza el análisis de regresión utilizando los N-1 restantes compuestos y prediciendo el valor de la propiedad del compuesto eliminado. El proceso se repite tantas veces como compuestos forme el grupo. El valor del coeficiente de predicción Q^2 nos informará sobre la calidad de la función

seleccionada y si el modelo es válido para fines predictivos (Q^2 ha de tomar un valor superior a 0.5 para poder considerar predictiva la función obtenida) [23,24]. El coeficiente de predicción obtenido fue de Q^2=0,687 ligeramente inferior a la varianza de la ecuación seleccionada (R^2=0.701) y bastante por encima del valor mínimo exigido ($Q^2 > 0.5000$). Los valores de logBCF predichos en la crosvalidación para cada compuesto aparecen en la columna 4 de la tablaS(II). Los resultados son similares a los mostrados en la columna 2.

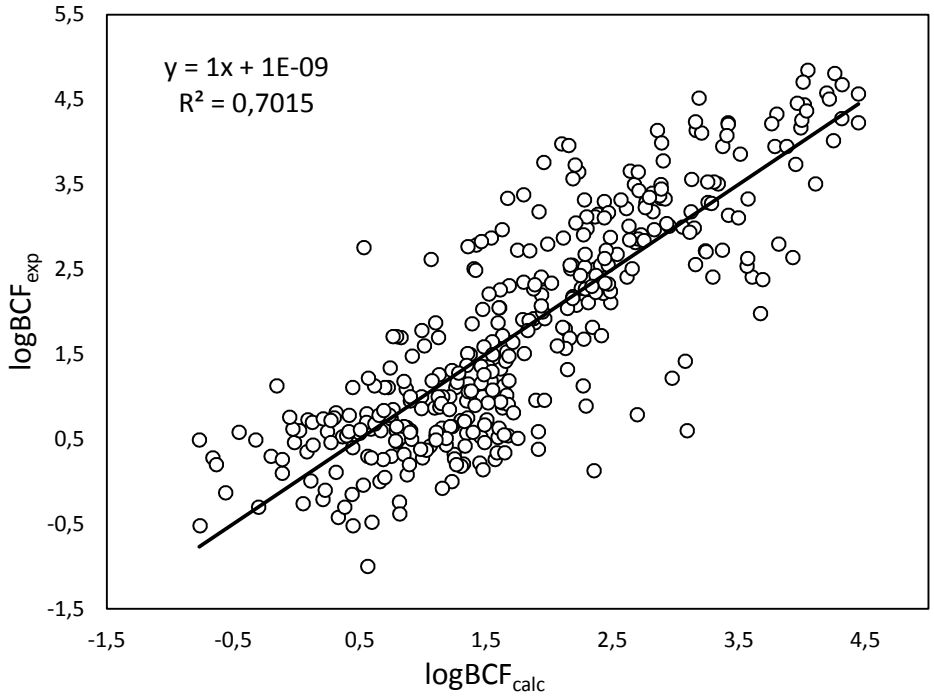


Figura 4.1 log BCF experimental frente al calculado para el modelo seleccionado con todas las moléculas

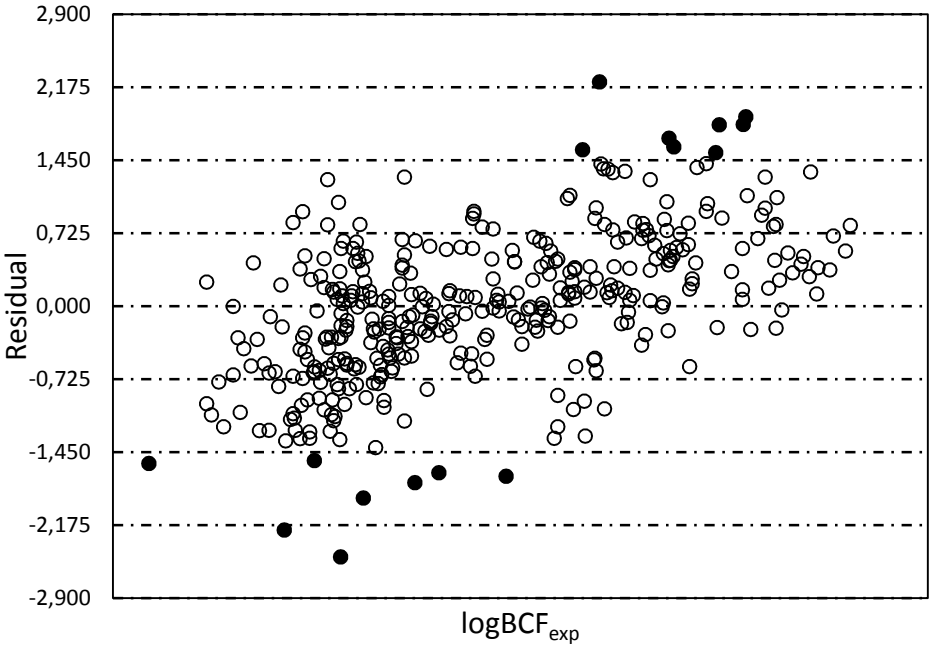
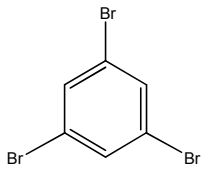
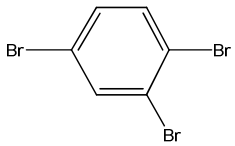
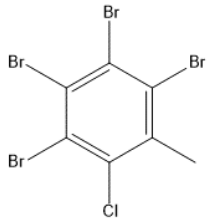
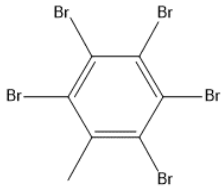
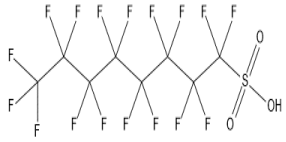
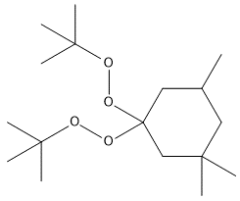
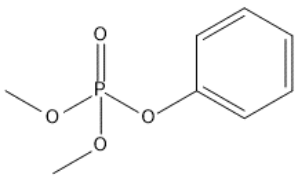
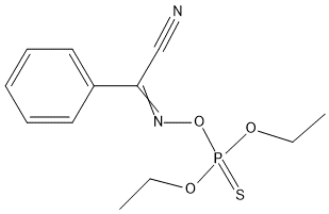


Figura 4.2 Residuales frente al experimental del log BCF del modelo seleccionado con todas las moléculas. Los puntos en negrita representan los outliers del conjunto de datos

Los outliers marcados en la figura 4.2 son respectivamente las moléculas identificadas en la tabla S(II) con el numero: bio57, bio63, bio171, bio172, bio175, bio176, bio288, bio289, bio296, bio400, bio419, bio426 bio471, bio476, bio486, bio488. La mitad de estas moléculas son falsos positivos, es decir que el modelo predice un valor de bioconcentración más elevado de lo real y el otro 50% son falsos negativos, es decir que el modelo subestima el valor de bioconcentración real, prediciendo un número menor respecto al valor experimental. Los casos críticos deben estudiarse uno a uno entre los falsos negativos que son los que el modelo no predice como compuestos peligrosos para el medioambiente.
Las moléculas críticas son representadas en la tabla siguiente:

Tabla 4.2 Identificación química de los falsos negativos pertenecientes a los outliers del grupo con todas las moléculas. Exp. = $logBCF_{exp}$ [tabla S(II)] y Calc. = $logBCF_{calc}$ (ecuación tabla 4.1)

bio171 1,3,5 dibromobenzene	bio172 1,2,4 dibromobenzene	bio175 tetrabromo-2-clorotoluene	bio176 pentabromotoluene
Exp.=3,38 Calc.=1,80	Exp.=3,34 Calc.=1,67	Exp.=3,98 Calc.=2,10	Exp.=3,98 Calc.=2,10
			
bio57 perfluorooctanoic sulfonic acid	bio476 1,1-Bis(tert-butyldioxy)-3,3,5-trimethyl cycloexano	bio486 Phosphoric acid dimethylphenyl ester	bio488 O,O-Diethyl-O(acyanobenzylideneamino) thiophosphate
Exp.=3,73 Calc.=2,20	Exp.=3,96 Calc.=2,15	Exp.=2,76 Calc.=0,53	Exp.=2,62 Calc.=1,07
			

Ahora haciendo referencia a los valores de las columas 1 y 2 de la tabla S(II), vamos a analizar entre las moléculas de la tabla 4.2, las que superan el umbral fijado por ley.

Se recuerda al lector los valores especificados en el reglamento REACH [3]:

- logBCF ≥ 3,3 molécula bioacumulable
- logBCF ≥ 3,7 molécula muy bioacumulable

Las moléculas bio486 y bio488 no pasan el umbral de bioacumulación, es decir que aunque sean falsos negativos, como el factor de bioconcentración experimental es muy bajo, el modelo falla la predicción pero no tanto como para superar el primer umbral legal.

Los compuestos bio171 y bio172 son bioacumulables mientras el modelo los clasifica como no bioacumulables.
Por último, las moléculas más críticas son el bio57, bio175, bio176 y el bio476 porque el modelo las clasifica como no bioacumulables siendo muy bioacumulables.

Después de este primer análisis, hemos que tener cuidado en el uso del modelo cuando nos enfrentamos sobre todo a bromuros aromáticos; de hecho existen estudios experimentales que demuestran como la masa molecular, así como el número y posición de los átomos de bromo pueden influir sobre el parámetro de bioacumulación.
Precisamente se ha visto que para masas moleculares superiores a 700 Da, disponemos de bajos valores de logBCF, por otro lado si tenemos masas moleculares entre los 450-700 la bioacumulación se ve afectada por el número y posición de los átomos de bromo [26].
La misma atención necesitan los ácidos sulfónicos y compuestos que son muy reactivos como un doble peróxido.
También el modelo falla cuando nos enfrentamos a tiofosfatos y ácidos fosfóricos, que no presentan valores de bioconcentración peligrosos para la contaminación del medioambiente, y por esto el fallo es menos importante.

4.1.1 Test de validación externo

En una segunda etapa se realiza un test de validación externo, utilizando la ecuación de regresión obtenida con el grupo de entrenamiento.

Tabla 4.3 Resultados del modelo de regresión multilineal (RML) con todas las moléculas aplicado al test externo. El asterisco (*) identifica los outliers

Compuesto	Nombre	$logBCF_{exp}$	$logBCF_{calc}$	Residuales
bio13	Bicyclo[4.3.0]nonane	2,91	2,09	0,82
bio14	t-Decalin	3,52	2,13	1,39
bio16	1,3,5-Trimethyl cyclohexane	3,34	2,23	1,11
bio21	2,3-Dimethylnaphthalene	2,71	2,75	-0,04
bio25	1,2,3,4-Tetramethyl benzene	2,82	2,36	0,46
bio29	Fluorene	2,78	2,82	-0,04
bio41	Chrysene	2,24	3,46	-1,22
bio42	Benzo[a]pyrene	2,69	3,63	-0,94
bio47	2,3-dichlorodibenzo-p-dioxin	2,88	2,46	0,42
bio50	1,2,3,7-Tetrachlorodibenzofuran	3,41	3,09	0,32
bio51	1,2,3,8-Tetrachlorodibenzofuran	3,11	3,12	-0,01
bio52	1,2,4-trichloroDibenzo-p-dioxin	2,97	2,69	0,28
***bio55**	Perflurohexane sulfonic acid	3,60	1,90	1,70
bio56	Perflurooctanic acid	3,12	2,24	0,88
bio68	Diisopentylether	2,24	1,77	0,47

bio77	Etylstyrene	2,57	2,68	-0,11
bio78	m-Cymene	2,73	2,21	0,52
***bio83**	p-tert-Butylphenol	1,83	-0,03	1,86
bio84	p-sec-Butylphenol	1,31	0,50	0,81
bio89	2,6-di-tert-Butyl-p-cresol	3,03	2,51	0,52
bio91	2,6-di-t-Butyl-4-ethylphenol	3,46	2,55	0,91
bio98	p-Bromophenol	1,17	1,23	-0,06
bio100	4-Chloro-m-cresol	0,92	1,52	-0,60
bio104	3,4-Dichlorophenol	1,69	1,60	0,09
bio117	1,5,9-Cyclododecatriene	3,92	3,05	0,87
bio119	Menthol	0,89	1,66	-0,77
bio124	2,6-Dicyclohexylphenol	2,89	2,20	0,69
bio127	Bis(2,3,5-trichloro-6-hydoroxyphenyl)methane	2,07	3,04	-0,97
bio132	2,4-Dichloro-1-hydroxynaphthalene	1,35	2,14	-0,79
bio134	2,6-Naphthalenedicarboxylic acid	0,72	0,92	-0,20
bio140	2-Naphthol-3,6-disulfonic acid	0,30	0,46	-0,16
bio144	Isopropyldecalin	3,58	2,53	1,05
bio151	2-Ethylanthraquinone	2,83	2,04	0,79
bio168	n-Hexyl cyclohexane	3,29	2,79	0,50
bio173	1,2,3,4-Tetrabromobenzene	3,18	1,75	1,43
bio178	2,3-Dichlorobiphenyl	3,72	3,44	0,28
bio188	2,3",4,4"-Tretrachlorobiphenyl	4,56	4,06	0,50
bio189	2,2",3,4"-Tretrachlorobiphenyl	4,53	3,99	0,54
bio194	2,3,4,5-Tretrachlorobiphenyl	4,39	3,98	0,41
bio196	2,3,4,5,6-Pentachlorobiphenyl	4,85	4,22	0,63
bio198	2,2",4,5,5"-Pentachlorobiphenyl	4,63	4,26	0,37
bio201	2,2",4,6,6"-Pentachlorobiphenyl	4,81	4,07	0,74
bio204	2,3",4,4",5-Pentachlorobiphenyl	4,38	4,32	0,06
bio219	Pentachlorobenzene	3,49	3,15	0,34
bio224	Tris(4-chlorophenyl)methanol	3,95	3,16	0,79
bio225	2,4,4"-Trichlorodiphenyl ether	3,79	2,94	0,85
bio230	3,3",4,4"-Tetrachlorodiphenyl ether	4,17	3,16	1,01
bio241	Trichlorometane	0,93	1,49	-0,56
***bio253**	1,3-Dibromo-2,2-bis(bromomethyl) propane	2,62	1,05	1,57
bio260	Trichloroethylene	1,00	1,49	-0,49
bio264	1,1,1-Trichloro-2-methyl-2-propanol	0,23	0,58	-0,35
bio271	1,2,3,4,5,6-Hexachlorocyclohexane	2,77	3,25	-0,48
***bio277**	Endrin	3,87	1,30	2,57
bio280	1-Methoxynaphthalene	2,21	1,67	0,54
bio293	2,2"-Methylenebis(6-t-buthyl-4-methylphenol)	1,97	2,40	-0,43
bio308	2-Nitropropane	0,92	0,31	0,61

bio317	N-Methylacetanilide	-0,30	0,64	-0,94
bio318	N,N-Diethylaniline	1,85	1,53	0,32
bio340	3-Nitroaniline	0,34	0,83	-0,49
bio343	N-Nitrosodiphenylamine	1,33	1,76	-0,43
bio344	Nitrobenzene	0,67	1,04	-0,37
bio355	1-Amino-2-methoxy-5-methyl-benzene	1,40	1,21	0,19
bio363	m-Nitroanisole	0,76	0,21	0,55
bio365	4-Nitro-m-cresol	1,06	1,02	0,04
bio369	Therephthalonitrile	0,24	0,63	-0,39
bio371	N-(3,4-Dichlorophenyl)-N''-methoxy-N''-methyl urea	1,26	1,39	-0,13
bio377	3,3''-Dichloro-4,4''-diaminodiphenylmethane	2,24	2,47	-0,23
bio381	1-(N-Phenylamino)naphthalene	3,23	2,60	0,63
bio383	1-Aminoanthraquinone	1,98	1,78	0,20
bio389	Quinoline	0,00	1,42	-1,42
bio395	3,3''-Dimethylbenzidine	1,67	1,90	-0,23
bio401	Tris(2-chloroethyl)phosphate	-0,18	1,17	-1,35
bio408	Ethanol,2-butoxy-,phosphate (3:1)	0,76	-0,14	0,90
bio411	2-Hydroxy-4-methoxybenzophenone	1,98	1,68	0,30
bio417	Benzothiazole	0,76	1,32	-0,56
bio424	2-Nitro-p-anisidine	0,82	0,30	0,52
bio431	alpha-Methylbenzylamine	0,66	1,38	-0,72
bio439	4-Vinylpyridine	1,86	1,16	0,70
bio453	2-(2''-Hydroxy-3'',5''-di-t-butylphenyl)-5-chlorobenzotriazole	1,00	2,23	-1,23
***bio466**	Bis(n-tributyltin)oxide	3,85	2,09	1,76
bio468	2,2-Dichloropropionic acid	0,85	0,30	0,55
bio492	7-Amino-4-hydroxy-2-naphthalenesulfonic acid	0,38	0,66	-0,28
bio494	4-amino-5-hydroxy-1,3-naphthalene disulfonic acid	0,90	0,55	0,35
bio497	2-Chloroanthraquinone	2,28	2,19	0,09
bio503	4,8-Diamino-9,10-dihydro-1,5-dihydroxy-9,10-dioxoanthracene-2,6-disulphonic acid	0,31	-0,28	0,59

A continuación se ven, en forma grafica, los resultados obtenidos a partir da los datos de la tabla 4.3.

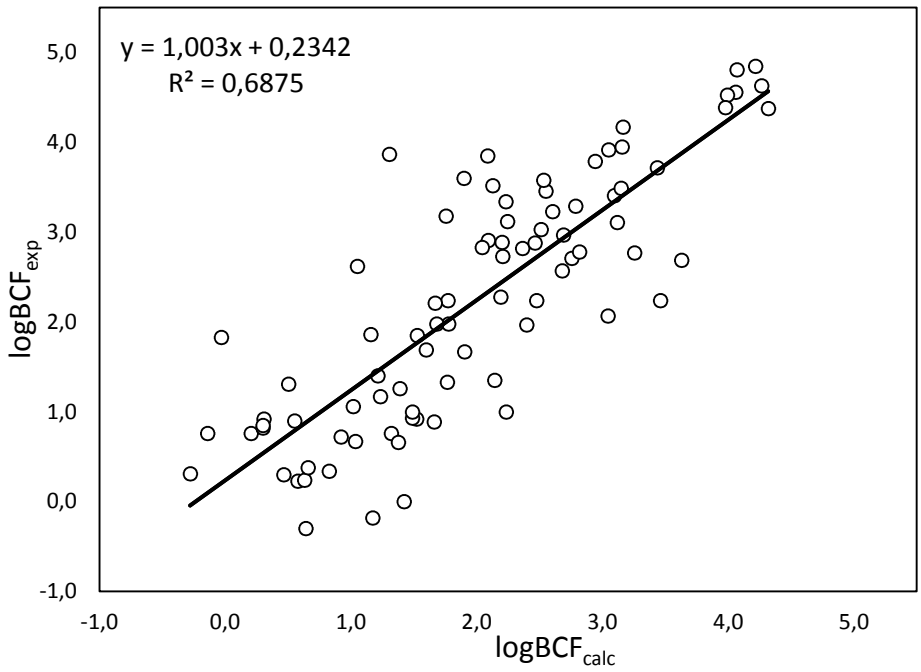


Figura 4.3 log BCF experimental frente al calculado del modelo seleccionado con todas las moléculas aplicado a un test externo

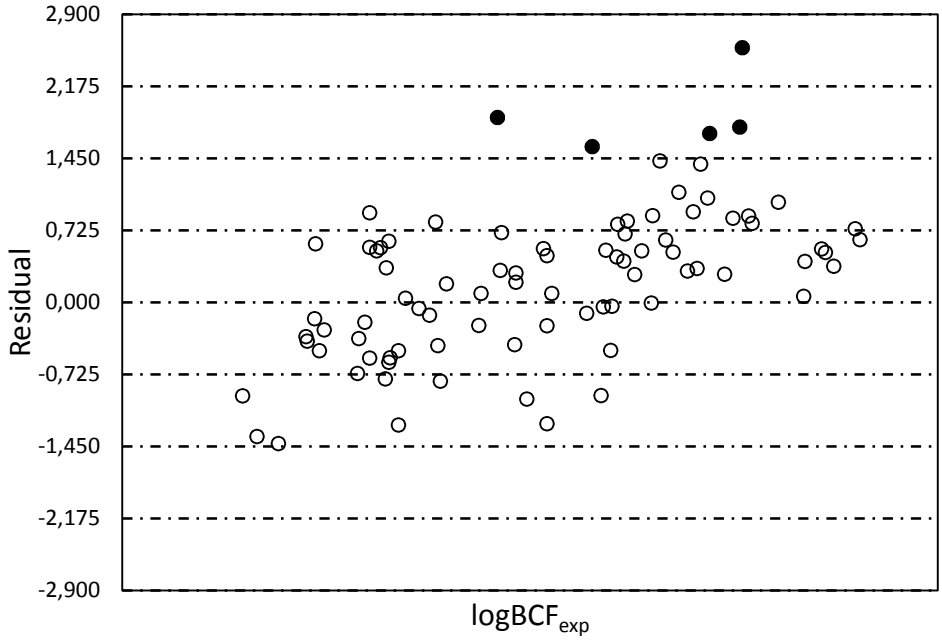
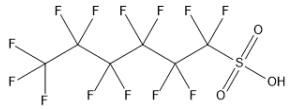
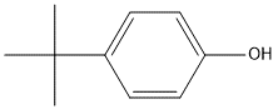
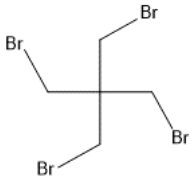


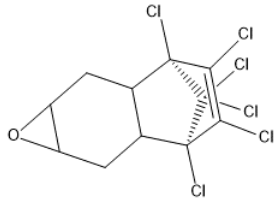
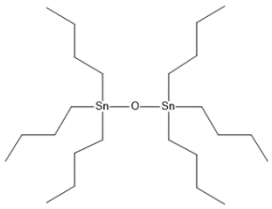
Figura 4.4 Residuales frente al experimental del log BCF del modelo seleccionado con todas las moléculas, aplicado a un test externo. Los puntos en negrita representan los outliers del conjunto de datos

Como muestran las gráficas de figura 4.3 y 4.4 el test externo tiene un valor de $R^2 = 0,685$ poco inferior respecto al valor del grupo de entrenamiento $R^2 = 0.701$. La variabilidad entre las dos correlaciones es pequeña, esto quiere decir que el modelo se aplica bien a un conjunto de datos independiente del grupo de entrenamiento.

La calidad del modelo se comprueba con el número de valores atípicos (outliers) encontrados (fig.4.4). A continuación se hace una tabla de las moléculas, que se salen del modelo de predicción, como ya se ha visto con el grupo de entrenamiento.

Tabla 4.4 Identificación química de los outliers pertenecientes al grupo test externo de todas las moléculas Exp. = $logBCF_{exp}$ (tabla 4.3) y Calc. = $logBCF_{calc}$ (ecuación tabla 4.1 aplicada al test externo)

bio55 Perfluorohexane sulfonic acid	bio83 p-tert-butylphenol	bio253 1,3-Dibromo-2,2-bis(bromomethyl) propane
Exp.=3,60 Calc.=1,90	Exp.=1,83 Calc= -0,03	Exp.=2,62 Calc= 1,05
		

bio277 Endrin	bio466 bis-(n-tributyltin) oxide
Exp.=3,87 Calc.=1,30	Calc.=2,09 Exp.=3,85
	

En este caso los ouliers (valores atípicos) son todos falsos negativos o bien presentan un valor de logBCF calculado más bajo del valor experimental. El bio55 que en realidad es biocumulable viene clasificado como poco bioacumulable por el modelo, mientras que el bio277 y el bio466 son los casos más críticos ya que son moléculas muy bioacumulables clasificadas como no bioacumulables.

Coherentemente con los outliers encontrados en el grupo de entrenamiento, el modelo no predice bien el comportamiento de los perfluorados (bio55), porque el átomo de flúor es muy electronegativo y muy poco susceptible a la polarización.

Ya que el modelo se basa en descriptores que caracterizan el volumen de Van der Waals, una baja polarizabilidad da un bajo valor del volumen y por esto la subestimación de la bioconcentración de estas moléculas. Además los otros descriptores llevan consigo una información estructural inherente a los átomos de Cloro, Oxigeno y Nitrógeno que no están presentes en estos compuestos [25].

El endrin (bio277), presenta un carbono cuaternario, como la estructura bio83 y además un grupo epóxido. La razón del fallo se puede encontrar en su complejidad estructural y por el hecho de ser un compuesto clorado, al igual que la molécula (bio466).

Para intentar mejorar el nivel de precisión de la predicción se desarrollaron dos modelos distintos, uno que tiene en cuenta solo las moléculas que contienen átomos de cloro, y otro modelo que analiza solo estructuras sin átomos de cloro.

4.2 Modelo de regresión multilineal para la predicción del factor de bioconcentración de las moléculas con cloro

El criterio de elección de la ecuación que mejor predice el modelo es el mismo utilizado en el caso anterior. Para facilitar la lectura a continuación incluimos una tabla de resumen:

Tabla 4.5 Análisis de regresión multilineal realizado con logBCF y diferentes números de variables

Variables	r^2	Δr^2
4	0,864	0
5	0,875	0,011
6	0,882	0,007

Ecuación de predicción seleccionada

	Coeficiente	EE	p
Intercepto	-1,044	0,418	0,0138
nX	0,537	0,036	0,0000
BAC	-0,043	0,003	0,0000
HVcpx	0,745	0,217	0,0008
MATS4v	0,625	0,086	0,0000
EEig05x	0,435	0,079	0,0000
N=126	r^2=0,875	Q^2=0,859	
F(5,120)=167	p=0,000	EEE=0,523	

La Tabla 4.5 recoge los resultados estadísticos obtenidos en función del número de variables seleccionadas. La agrupación de variables que mejor correlaciona el factor de bioconcentración es la formada por cinco descriptores (r^2 = 0,875 y EEE = 0,523). Todos ellos son estadísticamente significativos con valores de $p \leq 0,0000$. La incorporación de nuevas variables apenas modifican los parámetros anteriores ($\Delta r^2 \leq 0,010$).

Los índices topológicos que aparecen en la ecuación evalúan el número de átomos de halógenos (nX), la distribución del volumen de Van de Waals entre las moléculas (MATS4v) y otros valores puramente topológicos (BAC, HVcpx, EEig05x).

El modelo seleccionado es capaz de explicar más del 87% de la varianza de la propiedad correlacionada (r^2 = 0,875) con un error estándar de estimación cercano al 10,4% de la variabilidad en la que se mueve la propiedad (EEE = 0,523). Las figuras 4.5 y 4.6 representan los datos de las columna 1 y 2 de la tabla S(III) y muestran los resultados de predicción obtenidos para cada compuesto. Todos ellos, a excepción de los compuestos marcados en negro, presentan residuales inferiores a ± 2EEE, que se toma como valor límite para la aceptabilidad de la ecuación seleccionada. Sobre un total de 126 moléculas, tenemos 5 outliers es decir que el modelo no predice bien el 3,9% de los compuestos analizados. Como por el modelo precedente, se realiza una crosvalidación interna como ulterior prueba de verificación

de la calidad de la ecuación seleccionada. El resultado final es un coeficiente de predicción $Q^2 = 0{,}859$ ligeramente inferior a la varianza de la ecuación seleccionada ($R^2 = 0.875$) y muy por encima del valor mínimo exigido ($Q^2 > 0.5000$). Entonces la función seleccionada es válida también para fines legislativos, data su calidad de predicción [23,24]. Los valores de logBCF predichos en la crosvalidación para cada compuesto aparecen en la columna 4 de la tabla S(III), que se encuentra en el CD anexado. Los resultados son similares a los mostrados en la columna 2 de la misma tabla.

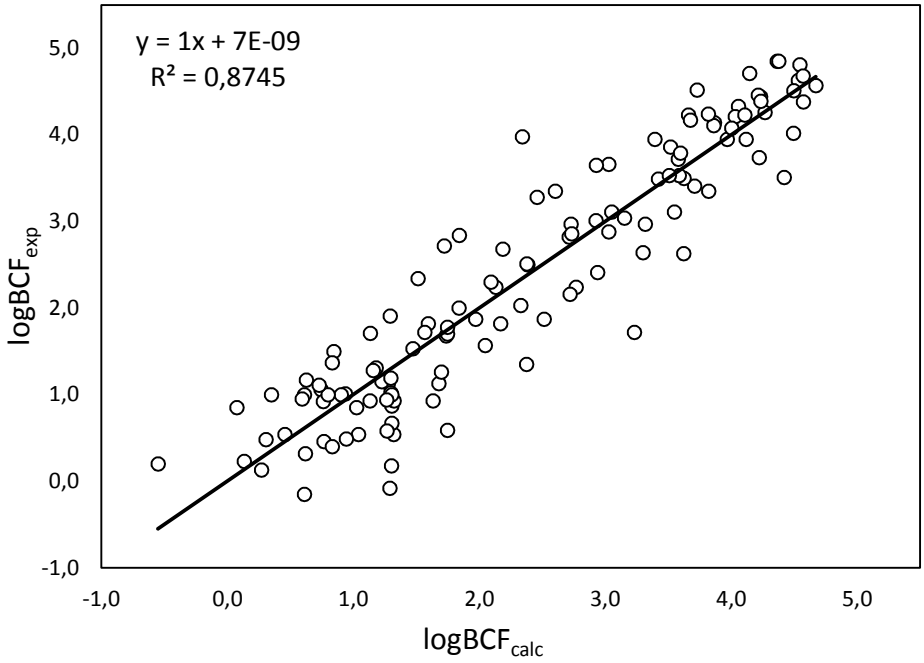


Figura 4.5 log BCF experimental frente al calculado para el modelo seleccionado de las moléculas con cloro

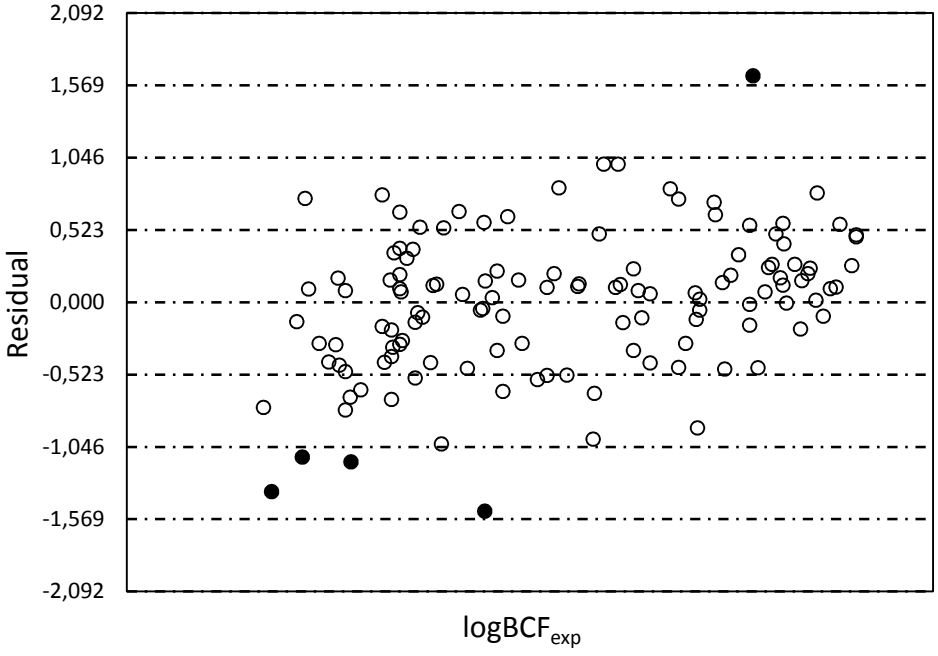
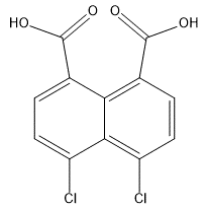
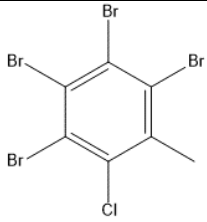
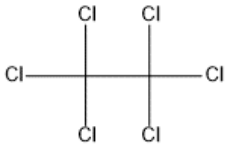
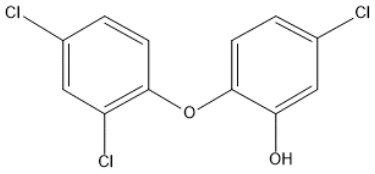
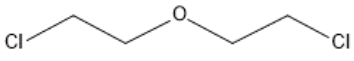


Figura 4.6 Residuales frente al experimental del log BCF del modelo seleccionado de las moléculas con cloro. Los puntos en negrita representan los outliers del conjunto de datos.

Siguiendo la misma línea de análisis ya vista en el modelo del apartado 4.1, vamos a identificar cuáles son las moléculas que presentan valores atípicos.
Dado el número exiguo de outliers, se decide reportar todos los compuestos involucrados y después centrarse sobre los clasificados como falsos negativos.

Tabla 4.6 Identificación química de los outliers pertenecientes al grupo de entrenamiento de las moléculas con cloro. Exp. = logBCF$_{exp}$ [tabla S(III)] y Calc. = logBCF$_{calc}$ (ecuación tabla 4.5)

bio135 4,5-dicloronaphalene 1,8-dicarboxylic acid	bio175 tetrabromo-2- clorotoluene	bio247 hexacloro-ethane
Exp.=0,18 Calc.=1,30	Exp.=3,98 Calc = 2,34	Exp.=0,59 Calc= 1,74
		

bio285 2,4,4'-trichloro-2'-hydroxydiphenyl ether	bio396 2,2'-Dichlorodiethylether
Exp.=1,72 Calc.=3,23	Exp.= -0,08 Calc.=1,28
	

Los outliers en la tabla 4.6 son respectivamente las moléculas marcadas en negro en la figura 4.6. Ahora vamos a seleccionar los falsos negativos, que como hemos visto, son los más críticos para hacer una correcta clasificación de bioacumulación ambiental.
La única molécula que tiene un valor de logBCF real superior a lo que predice el modelo es la bio175, compuesto ya encontrado en el modelo desarrollado en el apartado anterior.
La razón probablemente es debida al gran número de átomos de bromo respecto a los átomos de cloro. Además como ya se ha dicho en la sección precedente, en la literatura se encuentran estudios que demuestran como la posición y el número de átomos de bromo puede influir sobre el valor de bioconcentración [26]. Para las moléculas bio285 y bio396 se aprecia de nuevo la presencia de un epóxido, grupo que afecta la correcta predicción.

4.2.1 Test de validación externo

En la segunda etapa se realiza un test de validación externo, utilizando la ecuación de regresión obtenida con el grupo de entrenamiento.
Como ya se ha dicho el apartado anterior, el test de validación externo es fundamental si se quieren generalizar los resultados de un experimento. En este caso la ecuación del grupo de entrenamiento de las moléculas de cloro, construida partiendo de la base de 126 moléculas,

viene aplicada a un test externo independiente, para comprobar la robustez y la validez del modelo seleccionado. La tabla 4.7 muestra los valores de predicción de la bioconcentración para cada uno de los compuestos del test externo. Los resultados en forma gráfica aparecen reflejados en las figuras 4.7 y 4.8.

Tabla 4.7 Resultado del modelo de regresión multilineal (RML) aplicado al test externo de las moléculas con cloro. El asterisco (*) identifica los outliers

Compuesto	Nombre	$logBCF_{exp}$	$logBCF_{calc}$	Residuales
bio49	2,7-dichlorodibenzo-p-dioxin	2,22	2,66	-0,44
bio51	1,2,3,8-Tetrachlorodibenzofuran	3,11	3,77	-0,66
bio94	o-Chlorophenol	1,35	1,13	0,22
bio101	6-Chloro-m-cresol	0,53	1,18	-0,65
bio103	2,6-Dichlorophenol	1,03	1,48	-0,45
bio104	3,4-Dichlorophenol	1,69	1,59	0,10
bio126	2,2"-Dihydroxy-5,5"-dichlorodiphenylmethane	2,28	2,30	-0,02
bio127	Bis(2,3,5-trichloro-6-hydoroxyphenyl)methane	2,07	3,50	-1,43
bio185	2,4,5-Trichlorobiphenyl	4,22	3,92	0,30
bio187	2,2",3,3"-Tretrachlorobiphenyl	4,17	4,12	0,05
bio188	2,3",4,4"-Tretrachlorobiphenyl	4,56	4,40	0,16
bio189	2,2",3,4"-Tretrachlorobiphenyl	4,53	4,23	0,30
bio197	2,3,3",4,4"-Pentachlorobiphenyl	4,28	4,73	-0,45
bio200	2,2",4,5",6-Pentachlorobiphenyl	4,58	4,48	0,10
bio202	2,3",4,4",6-Pentachlorobiphenyl	4,81	4,64	0,17
bio210	1,3-Dichlorobenzene	2,33	2,13	0,20
bio212	3-Trifluoromethyl-chlorobenzene	2,35	2,93	-0,58
bio217	1,2,4,5-Tetrachlorobenzene	3,45	2,99	0,46
bio223	Tetrachloronaphthalene	3,33	3,71	-0,38
bio254	1,2-Dichloropropane	0,57	0,82	-0,25
bio255	1,2,3-Trichloropropane	0,96	1,45	-0,49
bio271	1,2,3,4,5,6-Hexachlorocyclohexane	2,77	2,73	0,04
bio277	Endrin	3,87	2,83	1,04
bio287	Heptachlor	3,95	3,73	0,22
bio331	o-Chloroaniline	1,36	1,27	0,09
bio334	2,4-Dichloroaniline	1,33	1,73	-0,40
***bio401**	Tris(2-chloroethyl)phosphate	-0,18	0,07	-0,25
bio402	Tris(1-chloro-2-propyl)phosphate	0,51	-0,29	0,80
bio487	2-Chloro-1-(2,4,5-trichlorophenyl) vinyl dimethylphosphate	1,60	2,21	-0,61
bio497	2-Chloroanthraquinone	2,28	2,52	-0,24
bio507	Disperse Red 206	1,22	0,04	1,18
bio509	2-Chloro-5-((2-hydroxy-1-naphthyl)azo)toluene-4-sulphonic acid	1,07	1,59	-0,52

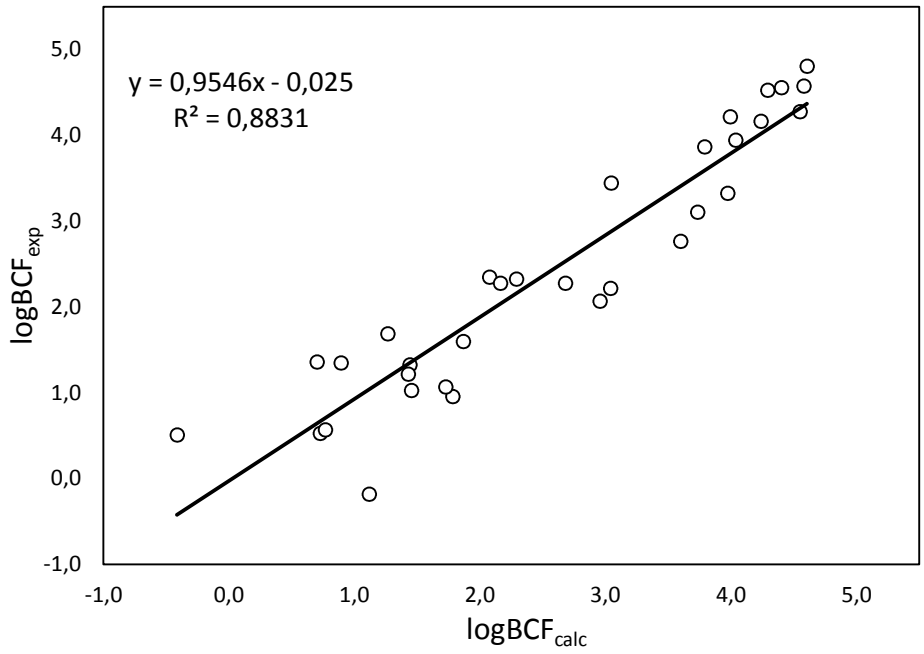


Figura 4.7 log BCF experimental frente al calculado del modelo seleccionado de las moléculas con cloro aplicado a un test externo

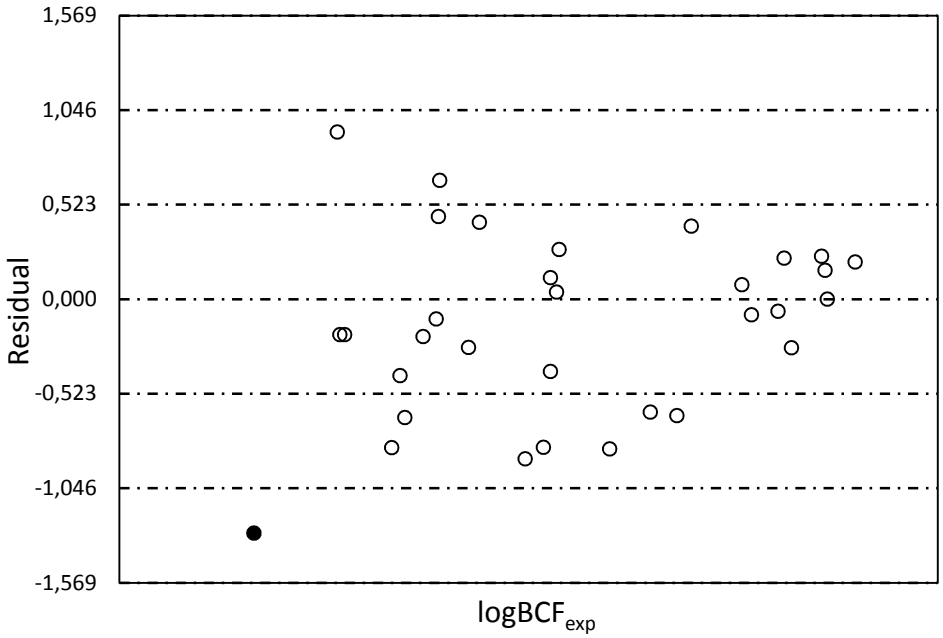
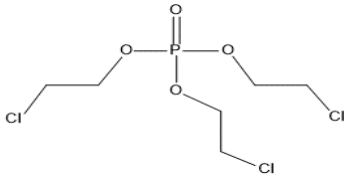


Figura 4.8 Residuales frente al experimental del log BCF del modelo seleccionado de las moléculas con cloro aplicado a un test externo. Los puntos en negrita representan los outliers del conjunto de datos

Como muestran los gráficos en figura 4.7 y 4.8, el test externo tiene un valor de R^2 =0,883, es decir un poco mayor respecto al valor del grupo de entrenamiento R^2 = 0.875. La variabilidad entre las dos correlaciones es pequeña, pero en este caso positiva, o sea que el modelo aplicado a un conjunto de datos independiente del grupo de entrenamiento, incluso mejora, aunque un poco, su predicción. La conveniencia de separar el conjunto de datos en moléculas que contienen cloro y que no lo contienen, se puede ver en la gráfica de predicción con la

representación de los ouliers. Como se ve en la figura 4.8 hay una sola molécula que se sale del modelo que además es un falso positivo, es decir que su valor está sobrecalculado.
La molécula en cuestión es la bio401 que, como se ve en la tabla 4.7, pertenece a los fosfatos, compuestos ya encontrados como atípicos en el modelo "general". Siendo un falso positivo y con valor muy bajo de bioacumulación, no presenta un carácter crítico a la hora de evaluar su comportamiento en el medioambiente.

Tabla 4.8 Identificación química de los outliers pertenecientes al grupo test externo de las moléculas con cloroExp. = $logBCF_{exp}$ (tabla 4.7) y Calc. = $logBCF_{calc}$ (ecuación tabla 4.5 aplicada al test externo)

bio401 tris(2 chloro-2-propyl) phosphate	
Exp.= -0,18	Calc.= 1,12
	

4.3 Modelo de regresión multilineal para la predicción del factor de bioconcentración de las moléculas sin cloro

Como en los apartados anteriores la tabla 4.9 recoge los resultados estadísticos obtenidos en función del número de variables seleccionadas.

Tabla 4.9 Análisis de regresión multilineal realizado con logBCF y diferentes números de variables

Variables	r^2	Δr^2
7	0,641	0
8	0,652	0,011
9	0,657	0,005

Ecuación de predicción seleccionada

	Coeficiente	EE	p
Intercepto	-0,475	0,408	0,246
SCBO	-0,048	0,014	0,001
nN	-0,146	0,065	0,025
nO	-0,168	0,033	0,000
T(O..F)	0,010	0,003	0,000
ATS2p	1,044	0,206	0,000
ATS6p	0,271	0,064	0,000
MATS1v	0,869	0,104	0,000
GATS5p	-0,199	0,060	0,001

N=276	r^2=0,652	Q^2=0,626
F(8,237)=55	p=0,000	EEE=0,712

En este caso, la agrupación de variables que mejor correlaciona el factor de bioconcentración es la formada por ocho descriptores (r^2 = 0,652 y EEE = 0,712). Todos ellos con valores de $p \leq 0,025$ o sea estadísticamente significativos, salvo el tèrmino independiente con $p \leq 0,246$. La incorporación de nuevas variables apenas modifica los parámetros anteriores ($\Delta r^2 \leq 0,010$).
Los índices topológicos que aparecen en la ecuación (tabla 4.8) evalúan el número de átomos de nitrógeno (nN), oxigeno (nO) y la suma del orden convencional de enlace sin considerar los hidrógenos (SCBO), presentes en las moléculas.
Después encontramos el índice T(O..F), que tiene en cuenta de la distancia topológica entre los átomos de oxígeno y de flúor. Por ultimo encontramos cuatro índices de autocorrelaacion ATS2p, ATS6p, MATS1v y GATS5p. Tres de ellos hacen referencia a la polizabilidad de las moléculas (subíndices p) y el MATS1v describe como se distribuyen los volúmenes de Van der Waals entre las moléculas.
El modelo seleccionado es capaz de explicar más del 65% de la varianza de la propiedad correlacionada (r^2 = 0,652) con un error estándar de estimación de cerca al 13,3% de la variabilidad en la que se mueve la propiedad (EEE = 0,712).
Las figuras 4.9 y 4.10 representan los datos de las columna 1 y 2 de la tabla S(IV) y muestran los resultados de predicción obtenidos para cada compuesto. Todos ellos, a excepción de los compuestos marcados con puntos negros, presentan residuales inferiores a ± 2EEE, lo cual es indicativo de la calidad de la ecuación seleccionada.
Sobre un total de 276 moléculas, tenemos solamente 13 outliers es decir que el modelo no predice bien el 4,7% de los compuestos analizados. Hasta ahora este modelo presenta valores estadísticos inferiores respecto al modelo general, lo que nos indica una calidad ligeramente inferior.

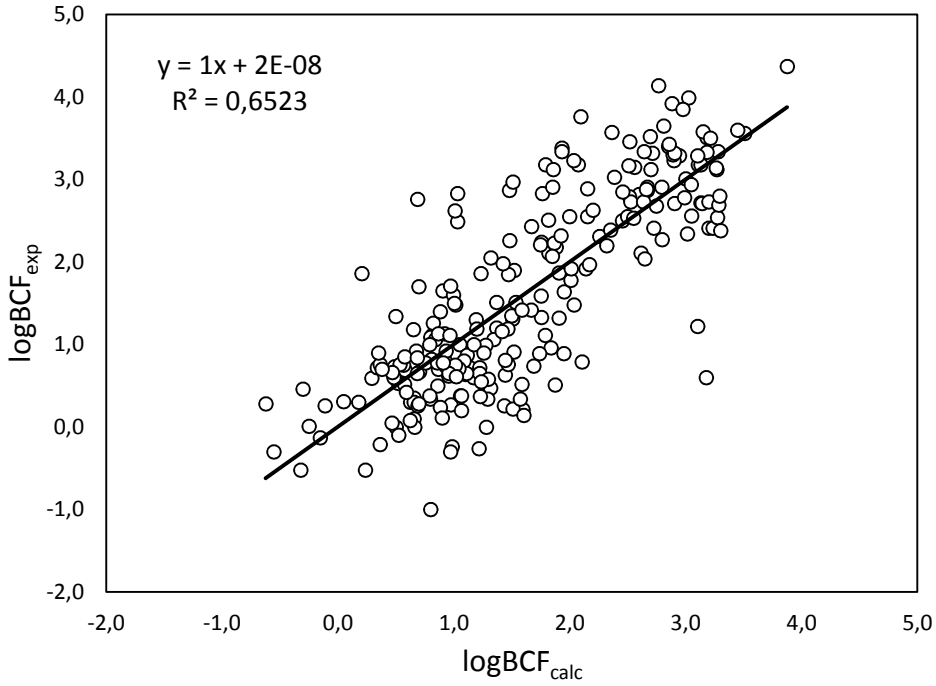


Figura 4.9 log BCF experimental frente al calculado para el modelo seleccionado de las moléculas sin cloro

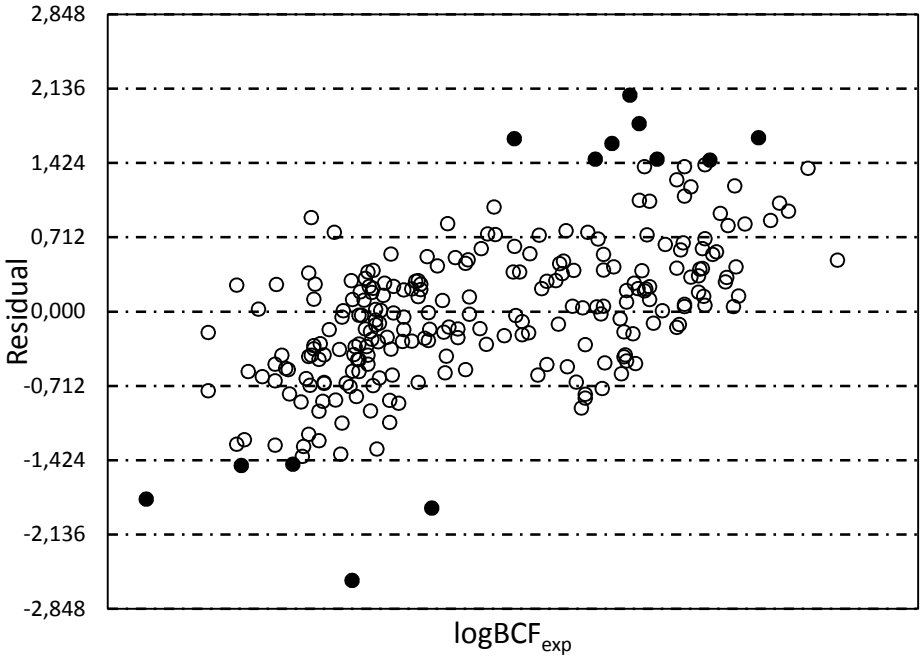
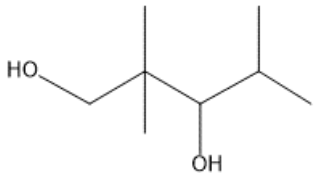
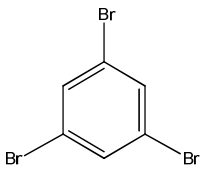
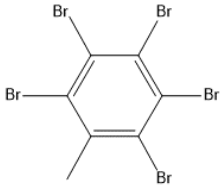
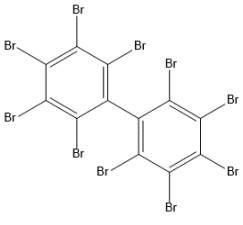
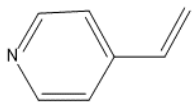
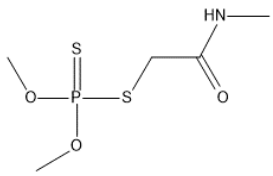
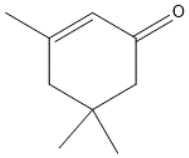
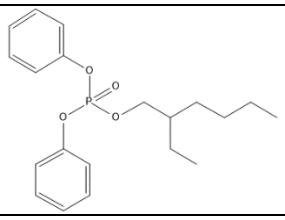
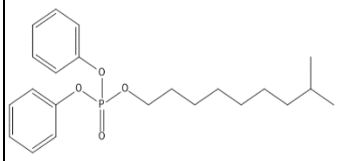
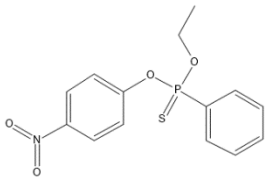
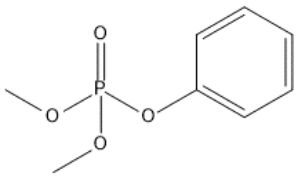
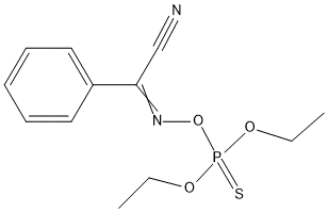


Figura 4.10 Residuales frente al experimental del log BCF del modelo seleccionado de las moléculas sin cloro. Los puntos en negrita representan los outliers del conjunto de datos

En la tabla 4.10, se ilustran las estructuras de las moléculas que presentan valores atípicos:

Tabla 4.10 Identificación química de los outliers pertenecientes al grupo de entrenamiento de las moléculas sin cloro. Exp. = logBCF$_{exp}$ [tabla S(IV)] y Calc. = logBCF$_{calc}$ (ecuación tabla 4.9)

bio63 2,2,4-Trimethyl-1,3-pentanediol	bio171 1,3,5 dibromobenzene	bio176 pentabromotoluene
Exp.= -1,00 Calc.=0,79	Exp.=3,38 Calc.=1,93	Exp.=3,98 Calc.=2,09
		
bio288 n-pentadecane	bio296 Decabromobiphenyl	Bio439 4-Vinylpyridine
Exp.=1,22 Calc.=3,10	Exp.=0,60 Calc.=3,17	Exp.=1,86 Calc.=0,20
		

Bio464 Dimethoate	bio477 Isophorone	bio479 2-Ethylhexyl diphenyl phosphate
Exp.= -0,26 Calc.=1,21	Exp.=0,14 Calc.=1,60	Exp.=2,49 Calc.=1,02
		
bio480 Isodecyl diphenyl phosphate	bio484 EPN	bio486 Phosphoric acid dimethylphenyl ester
Exp.=2,83 Calc.=1,02	Exp.=2,97 Calc.=1,50	Exp.=2,76 Calc.=0,68
		
bio488 O,O-Diethyl-O(acyanobenzylideneamino) thiophosphate Exp.=2,62 Calc.=1,00		

En este apartado se ha elegido mostrar todas las moléculas atípicas para identificar las estructuras químicas que efectivamente afectan una óptima construcción de la regresión multilineal.

Como se ha mostrado en la tabla 4.10, el modelo tiene problemas cuando debe predecir el comportamiento de moléculas como los tiofosfatos y derivados de los ácidos fosfóricos.

El mismo problema ya apareció en el modelo "general" (tabla 4.2) y de hecho algunas moléculas reaparecen con valores atípicos también en este nuevo modelo.

Las moléculas en cuestión son las: bio486, bio488 a las cuales se añaden las bio464, bio479, bio480, bio484, pertenecientes a las misma clase de compuestos que fueron bien predichas en el modelo "general".

Continuando el análisis de los outliers, encontramos otros puntos en común con el primer modelo desarrollado, en concreto la aparición de las moléculas bio63, bio171, bio176, bio288, y bio296.

Como ya se ha visto y comentado en el apartado 4.1, solo las bio171 y bio176 son un peligro para el medioambiente.

Es interesante apuntar que el nuevo modelo predice bien los perfluorados, (bio57 y bio419), lo que apoya la tesis precedente formulada (apartado 4.1) sobre una probable subestimación del valor del LogBCF calculado, debido a la baja polarizabilidad que implica un bajo volumen de Van der Waals.

De hecho entre los descriptores de este nuevo modelo también aparecen las funciones de autocorrelación, con la gran diferencia de que describen directamente la propiedad de polaridad de los átomos y por esto son más precisos para evaluar moléculas poco polarizables. La importancia de los índices topológicos en la descripción de la estructura molecular está comprobada por el hecho de que en este modelo como en el "general" aparece una molécula como el n-pentadecano (bio288), que está clasificado como outlier.
En efecto en ambos modelos faltan descriptores que describan el comportamiento de la flexibilidad molecular, de modo que moléculas flexibles y con baja capacidad de pasar la membrana celular [28], vienen predichas con un valor de bioconcentración más elevado del que presentan experimentalmente [1,2].

4.3.1 Test de validación externo

En una segunda etapa se realiza un test de validación externo, utilizando la ecuación de regresión obtenida con el grupo de entrenamiento.

Tabla 4.11 Resultado del modelo de regresión multilineal (RML) de las moléculas sin cloro aplicado al test externo. El asterisco (*) identifica los outliers

Compuesto	Nombre	$logBCF_{exp}$	$logBCF_{calc}$	Residuales
bio4	1-Decene	3,22	2,64	0,58
bio5	2,2,4,6,6-Pentamethyl-3-heptene	3,29	3,46	-0,17
bio10	Isodecanol 10	2,51	1,60	0,91
bio11	Isotridecanol	2,73	1,86	0,87
bio23	2,3,6-trimethylnaphthalene	3,00	3,10	-0,10
bio30	Anthracene	2,99	3,09	-0,10
bio41	Chrysene	2,24	3,23	-0,99
bio57	Perflurooctane sulfonic acid	3,73	4,81	-1,08
bio64	2,2-Dimethyl-1,3-propanediol	-0,42	1,00	-1,42
bio66	Diethyl ether	0,73	-0,05	0,78
bio69	2-Ethylhexylvinylether	2,77	1,44	1,33
bio74	alpha-Methylstyrene	1,80	2,36	-0,56
bio75	1,2,4-Trimethylbenzene	2,08	2,50	-0,42
bio77	Etylstyrene	2,57	2,69	-0,12
bio81	m- Divinylbenzene	2,55	2,54	0,01
bio83	p-tert-Butylphenol	1,83	1,53	0,30
bio98	p-Bromophenol	1,17	-0,01	1,18
bio115	Methylcyclohexane	2,27	2,35	-0,08
bio140	2-Naphthol-3,6-disulfonic acid	0,30	0,76	-0,46
bio146	Camphene	2,98	2,89	0,09
bio154	Benz anthracene-7,12-quinone	1,69	1,75	-0,06
bio156	3,5-Di-tert-Butylbiphenyl-4-ol	3,78	2,56	1,22
bio167	1-Isobutyl 2,5 dimethyl cyclohexane	3,37	3,21	0,16
bio253	1,3-Dibromo-2,2-bis(bromomethyl) propane	2,62	1,80	0,82

bio265	Dibromoneopentylglycol	-0,04	1,34	-1,38
bio280	1-Methoxynaphthalene	2,21	1,37	0,84
bio281	2-Naphthylisobuthyl ether	2,80	1,80	1,00
***bio289**	2,2,4,4,6,8,8-Heptamethylnonane	1,98	3,80	-1,82
bio294	2,2-Bis(4"-hydroxy-3",5"-dibromophenyl) propane	2,43	2,35	0,08
bio310	Triethanolamine	0,59	-0,06	0,65
bio312	2-Butanone oxime	0,62	0,96	-0,34
bio314	Bis(cyanoethyl)amine	-0,48	0,37	-0,85
***bio315**	1-Cyanoguanidine	0,49	-0,99	1,48
bio322	3,4-Dimethylaniline	0,42	0,52	-0,10
bio323	2,5-Dimethyl aniline	0,58	1,42	-0,84
bio337	1,3-Bis(aminomethyl)benzene	0,43	0,75	-0,32
bio338	N,N-Dimethylbenzylamine	0,63	1,31	-0,68
bio346	m-Nitrotoluene	0,88	0,82	0,06
bio351	3,4-Dinitrotoluene	0,43	0,74	-0,31
bio360	p-Nitrophenol	0,60	0,72	-0,12
bio363	m-Nitroanisole	0,76	0,83	-0,07
bio364	2-Nitro-p-cresol	0,95	0,61	0,34
bio366	2,4-Dinitrophenol	0,57	0,49	0,08
bio378	1-Naphthylenamine	1,26	1,12	0,14
bio383	1-Aminoanthraquinone	1,98	1,52	0,46
bio388	3-Aminopyridine	0,32	0,68	-0,36
bio390	1,4-Dioxane	-0,30	0,11	-0,41
bio392	2,4,6-Triamino-1,3,5-triazine	0,58	-0,69	1,27
bio395	3,3"-Dimethylbenzidine	1,67	1,97	-0,30
bio398	Dimethyl sulfoxide	0,60	0,08	0,52
bio416	Dibenzothiophene	3,05	2,18	0,87
bio418	Thiophene	0,86	1,05	-0,19
bio426	N,N-Diethyl-m-toluamide	0,38	1,52	-1,14
bio429	3,5-Xylyl N-methylcarbamate	0,28	1,17	-0,89
bio438	3-Amino-1,2,4-triazole	0,49	-0,45	0,94
bio444	1,3,5-Tris(2"-hydroxyethyl)isocyanuric acid	0,20	0,13	0,07
bio455	2-Isopropyl-4-methyl-6-hydroxypyrimidine	-0,38	0,61	-0,99
***bio476**	1,1-Bis(tert-butyldioxy)-3,3,5-trimethyl cyclohexane	3,96	2,16	1,80
bio481	Diphenylmonotridecylphosphite	2,05	1,21	0,84
bio485	Benzenesulfonamide	0,68	0,60	0,08
bio493	1-Amino-8-naphthol-3,6-disulfonic acid	0,46	0,69	-0,23
bio506	Disperse Yellow 64	1,08	1,56	-0,48

A continuación se ven los resultados de los análisis obtenidos a partir de los datos de la tabla 4.11.

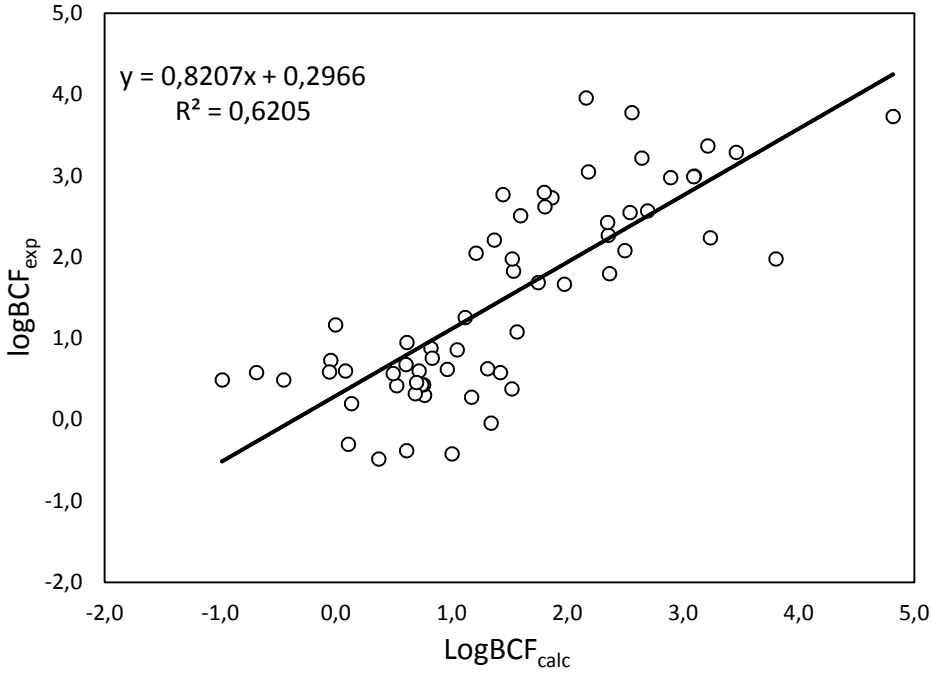


Figura 4.11 log BCF experimental frente al calculado del modelo seleccionado de las moléculas sin cloro aplicado a un test externo

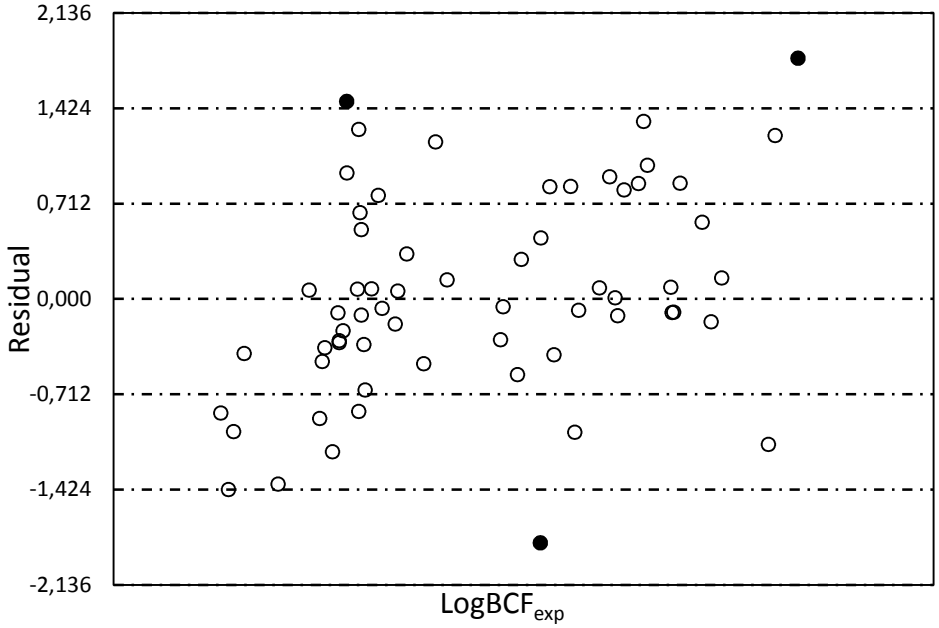
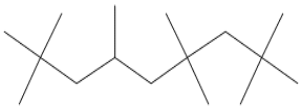
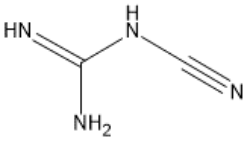
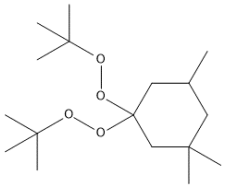


Figura 4.12 Residuales frente al experimental del log BCF del modelo seleccionado de las moléculas sin cloro aplicado a un test externo. Los puntos en negrita representan los outliers del conjunto de datos

Como muestran los gráficos en figura 4.10 y 4.11, el test externo tiene un valor de R^2=0,62 que es algo inferior respecto al valor del grupo de entrenamiento (R^2=0.652). La variabilidad entre las dos correlaciones es pequeña, lo que quiere decir que el modelo funciona bien en un conjunto de datos independiente del grupo de entrenamiento.
La confirmación de lo que acabamos de decir se puede encontrar analizando la tabla 4.11 de los outliers del grupo test. Si nos referimos a la gráfica logBCF$_{exp}$ vs Residual (fig.4.11), se

nota que se destacan solo tres moléculas del conjunto. En la tabla 4.12 se identifican las estructuras químicas de los compuestos atípicos.

Tabla 4.12 Identificación química de los outliers pertenecientes al grupo test externo de las moléculas sin cloro. Exp. = $logBCF_{exp}$ (tabla 4.11) y Calc. = $logBCF_{calc}$ (ecuación tabla 4.9 aplicada al test externo)

bio289 2,2,4,4,6,8,8- Heptamethylnonane	bio 315 1-Cyanoguanidine	bio476 1,1-Bis(tert-butyldioxy)- 3,3,5-trimethyl cycloexano
Exp.=1,98 Calc.=3,80	Exp.=0,49 Calc.= -0,99	Exp.=3,96 Calc.=2,16
		

Mirando la columna 1 de la tabla 4.11 vemos que el bio476 resulta muy bioacumulable mientras el bio289 y el bio315 no resultan bioacumulables. El primero es subestimado y, como se señalaba anteriormente, es un problema a la hora de predecir correctamente su valor de bioacumulación en el ambiente acuático. Además el bio476 ya se encontró en el modelo "general", donde ya se explicó que su mala predicción es probablemente debida a su carácter de molécula muy reactiva.
El otro compuesto, el bio289, es un falso positivo, por lo que no hay que preocuparse desde el punto de vista ambiental. Su análisis es interesante a nivel topológico porque es una molécula larga, muy ramificada y con una cierta flexibilidad, características que los índices del modelo no tienen en consideración y por esto es sobreestimado. Además tiene tres carbonos cuaternarios, característica común con las otras moléculas atípicas, ya encontradas en los demás modelos.
El bio 315, también es un falso negativo pero su valor real de bioconcentración es muy bajo por lo que no da preocupaciones ambientales.

4.4 Modelos de regresión no lineal con redes neuronales artificiales

La tercera etapa es realizar, sobre la base de los índices topológicos utilizados para describir el factor de bioconcentración, una red neuronal para cada modelo de regresión multilineal.
Cada red ha sido construida considerando el grupo de entrenamiento empleado en el análisis de regresión multilineal (RML) como base de datos y tomando el 80% para la construcción de la red y el restante 20% como entrenamiento de la misma.

Una vez obtenida la red, se guarda en formato PMML script y sucesivamente se aplica al mismo test externo del análisis RML, para ver si hay una mejora en la predicción.

4.4.1 Red neuronal artificial con todas las moléculas

La RNA seleccionada tiene una arquitectura de multicapas, MLP 7-10-1, con 7 neuronas en la capa de entrada, una para cada descriptor seleccionado, diez neuronas en la capa interna y una en la capa de salida (ya que queremos predecir solo el logBCF).
La función de activación utilizada en esta red es la sigmoidal logística para las neuronas de la capa interna mientras para la neurona de la capa de salida es la función identidad (la activación de la neurona pasa directamente como señal de salida).

Tabla 4.13 Parámetros estadísticos del modelo de regresión multilineal (RML) y de la red neuronal artificial (RNA) con todas las moléculas

Grupo	r^2	EEE
Regresión Multilineal (RML)		
Training	0,701	0,725
Test	0,687	-
Red Neuronal Artificial (RNA)		
Training	0,816	0,565
Test	0,825	-

Al comparar los resultados obtenidos con la red neuronal artificial (RNA), se resalta una mejora en todos los parámetros estadísticos.
Como ya se ha dicho antes, se ha trabajado con los 7 índices topológicos resultantes del modelo de regresión multilineal (RML), de modo que pudiéramos tener una mejor comparación de los resultados.
Desde el momento que el training de la RNA se procesa sobre el 80% de las 381 moléculas totales empleadas, es decir 305 moléculas, no parece oportuno comparar las gráficas de los dos modelos construidos con técnicas distintas.
Lo que sí que podemos comparar son las gráficas del test externo independiente, que es igual para ambos tipos de análisis. En la tabla 4.13 se ve un aumento significativo del r^2 y una reducción del EEE.
A continuación se ponen las dos graficas en comparación, añadiendo también las gráficas que resaltan los outliers, para ver las eventuales mejoras.

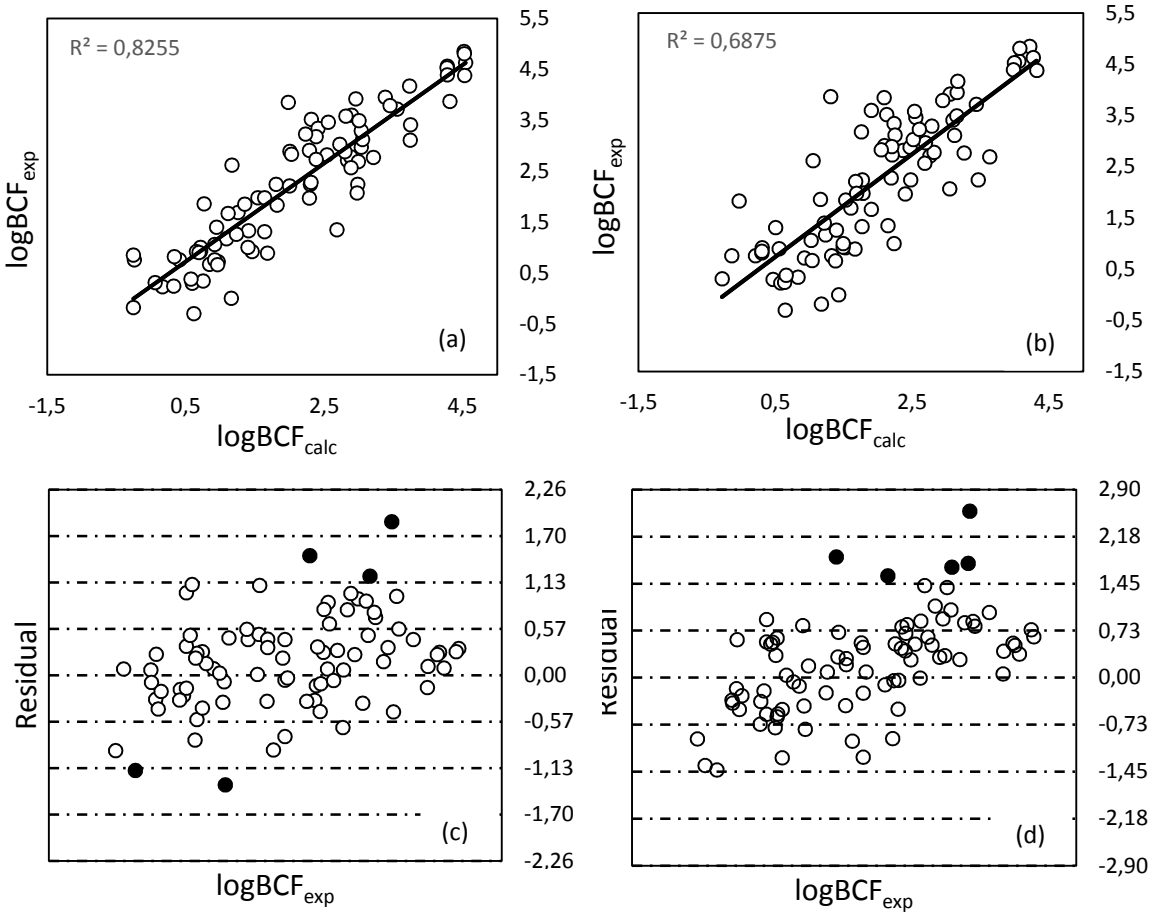


Figura 4.13 Log BCF experimental frente al calculado con todas las moléculas: (a) aplicando la RNA al test externo, (b) aplicando el RML al test externo. Residual frente al log BCF experimental: (c) aplicando la RNA al test externo, (d) aplicando el RML al test externo. Los puntos en negrita representan los outliers del conjunto de datos.

Comparando los gráficos de regresión de figura 4.13 (a) y (b) se nota una gran diferencia en los ajustes de los datos, como se podía esperar solo viendo la diferencia entre los cuadrados de los coeficientes de correlación (tabla 4.13).
El dato más interesante es la diferencia entre los ouliers en los gráficos de figura 4.13 (c) y (d). La comparación es relativa y cualitativa porque cada gráfico tiene su orden de magnitud debido al diferente error estándar de estimación (EEE).
Los outliers del modelo de RNA aplicado al test externo son el bio14, bio132, bio253, bio389 y el bio466, mientras que, como ya hemos visto antes, para el modelo de RML tenemos las moléculas bio55, bio83, bio253, bio277 y bio466.
Como ya hemos señalado anteriormente, el problema surge cuando una molécula sale del umbral fijado por la ley y se bioacumula en el medioambiente.
Analizando los outliers de ambos casos, tomando como umbral el logBCF≥3.3, las moléculas bio132, bio253 y la bio83, no preocupan mientras que las demás presentan toxicidad.

La mejora que aporta el modelo construido con redes neuronales es la detección de dos moléculas biocumulables en el medio ambiente como el Endrin (bio277) y el ácido sulfónico perfluorohexano (bio55) que vienen predichos perfectamente (tabla 4.4).
El modelo de red neuronal falla sobre un compuesto bioacumulable, bien predicho por el modelo de regresión multilineal, hablamos de la molecula bio14, (t-Decalin), hidrocarburo alifático que consta en dos anillos de ciclohexano unidos por dos átomos de carbono y que comparten un enlace. Es una molécula muy apolar.

La única molécula que presenta toxicidad en ambos los modelos es el óxido de tributilestaño (bio466), que como en el modelo estructurado en algoritmos de regresiones multilineales (RML), no viene bien detectada resultando como no bioacumulables cuando experimentalmente lo es. Esto tiene su lógica considerando que es un compuesto organometálico, muy diferente por tanto al resto de los compuestos analizados.
Ahora en base a los datos de la tabla 4.3 y de la tabla S(V), se hace un análisis cuantitativo de todas las moléculas predichas, con el fin de individualizar los falsos negativos que se salen del límite legal.
El modelo RML falla en 13 casos sobre 85 moléculas, es decir tenemos un error del 15,3% mientras el modelo construido con RNA no predice bien 9 casos sobre 85, es decir que el error es del 10,6%.
Podemos concluir que los dos son modelos aceptables, aunque sin duda la red neuronal mejora bastante la predicción.
Como ya se tiene una información estructural que dice *a priori* las moléculas a las cuales poner atención, no es incorrecto quitar los valores atípicos que son superiores respecto el umbral de ley. Así pasamos a 10 casos para el RML y a 7 casos para la RNA.
En este sentido, el error disminuye hasta un 11,7% para el RML y a un 8,2% para la RNA.

4.4.2 Red neuronal artificial de las moléculas con cloro

El análisis de los datos inherentes del grupo clorado da como resultados una red neuronal que tiene una arquitectura de multicapas, MLP 5-11-1, con cinco neuronas en la capa de entrada, una para cada descriptor seleccionado, once neuronas en la capa interna y una en la capa de salida.
La función de activación utilizada en esta red es la sigmoidal logística sea para las neuronas de la capa interna como para las neuronas de la capa de salida.

Tabla 4.14 Parámetros estadísticos del modelo de regresión multilineal (RML) y de la red neuronal artificial (RNA) de las moléculas con cloro

Grupo	r^2	EEE
Regresión Multilineal (RML)		
Training	0,875	0,523
Test	0,883	-
Redes Neuronales Artificial (RNA)		
Training	0,916	0,428
Test	0,899	-

Aunque el modelo de redes neuronales sea mejor no se ve una marcada diferencia entre las dos metodologías de predicción. La comparación de las gráficas de los tests externos utilizados en ambos casos confirma esto último.

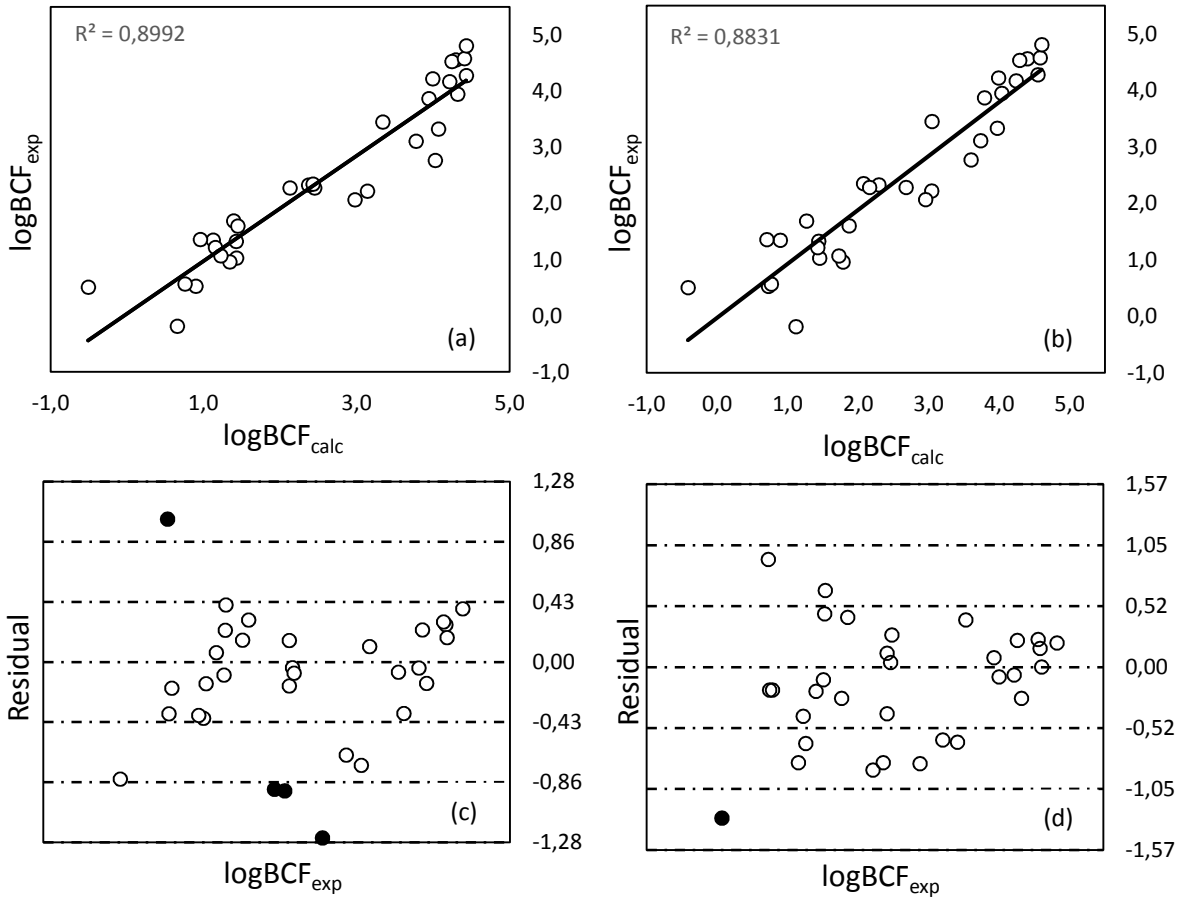


Figura 4.14 Log BCF experimental frente al calculado de las moléculas con cloro: (a) aplicando la RNA al test externo, (b) aplicando el RML al test externo. Residual frente al log BCF experimental: (c) aplicando la RNA al test externo, (d) aplicando el RML al test externo. Los puntos en negrita representan los outliers del conjunto de datos.

Como se puede apreciar de las gráficas de la fig.4.14, no hay una gran diferencia de resultados de una técnica respecto a la otra. La única gran diferencia es que el modelo RNA genera más outliers que el modelo RML. Los outliers del modelo RNA que se ven en la figura 4.14(c) son las moléculas bio402, bio127, bio49 y bio271, mientras el modelo RML tiene solo la bio401 (véase fig.4.14(d)). Ninguna de ellas son bioacumulables y ni siquiera peligrosas para el medioambiente.

Analizando más detalladamente los resultados se resalta una diferente distribución de los datos de fig.4.14 (c) y (d).

En la gráfica residual vs $logBCF_{exp}$ de las redes neuronales, los datos parecen más "condensados" entre la banda de ± EEE, quiere decir que la predicción es más precisa.

Haciendo un análisis cualitativo tomando como referencia los datos de tabla (4.7), a fin de detectar los falsos negativos que se salen del umbral de ley ($logBCF \geq 3,3$), se nota que el modelo RML falla en 1 caso sobre 32 moléculas (error del 2,8%), mientras el modelo construido con RNA no falla en ningún caso [tabla S(VI)].

4.4.3 Red neuronal artificial de las moléculas sin cloros

En este grupo de moléculas se sigue la misma línea de análisis tomada desde el principio del capítulo.
La red neuronal seleccionada tiene una arquitectura de multicapas, MLP 8-5-1, con ocho neuronas en la capa de entrada, una para cada descriptor seleccionado, cinco neuronas en la capa interna y una en la capa de salida.
La función de activación utilizada en esta red es la sigmoidal logística sea para las neuronas de la capa interna que para las neuronas de la capa de salida.

Tabla 4.15 Parámetros estadísticos del modelo de regresión multilineal (RML) y de la red neuronal artifical (RNA) de las moléculas sin cloro

Grupo	r^2	EEE
Regresión Multilineal (RML)		
Training	0,652	0,712
Test	0,62	-
Redes Neuronales Artificial (RNA)		
Training	0,695	0,675
Test	0,706	-

Los resultados finales obtenidos (tabla 4.15) aplicando la red neuronal aumentan mucho la calidad de la predicción. Todos los parámetros estadísticos mejoran. De hecho el error estándar de estimación se reduce mientras el coeficiente de correlación aumenta significativamente.
Como en los casos precedentes, se lleva a cabo una comparación entre las gráficas de cada test externo.
Como se puede apreciar en la figura 4.15, que encontramos en la página siguiente, el modelo construido con red neuronal ajusta mucho mejor los datos que el modelo de regresión lineal múltiple (fig.4.15 (a) y (b)).
Además la distribución de los datos calculados se concentra en la banda ± EEE (fig 4.15 (c)), mientras que en el modelo de regresión lineal múltiple se aprecia una distribución más dispersa (fig.4.15 (d)).

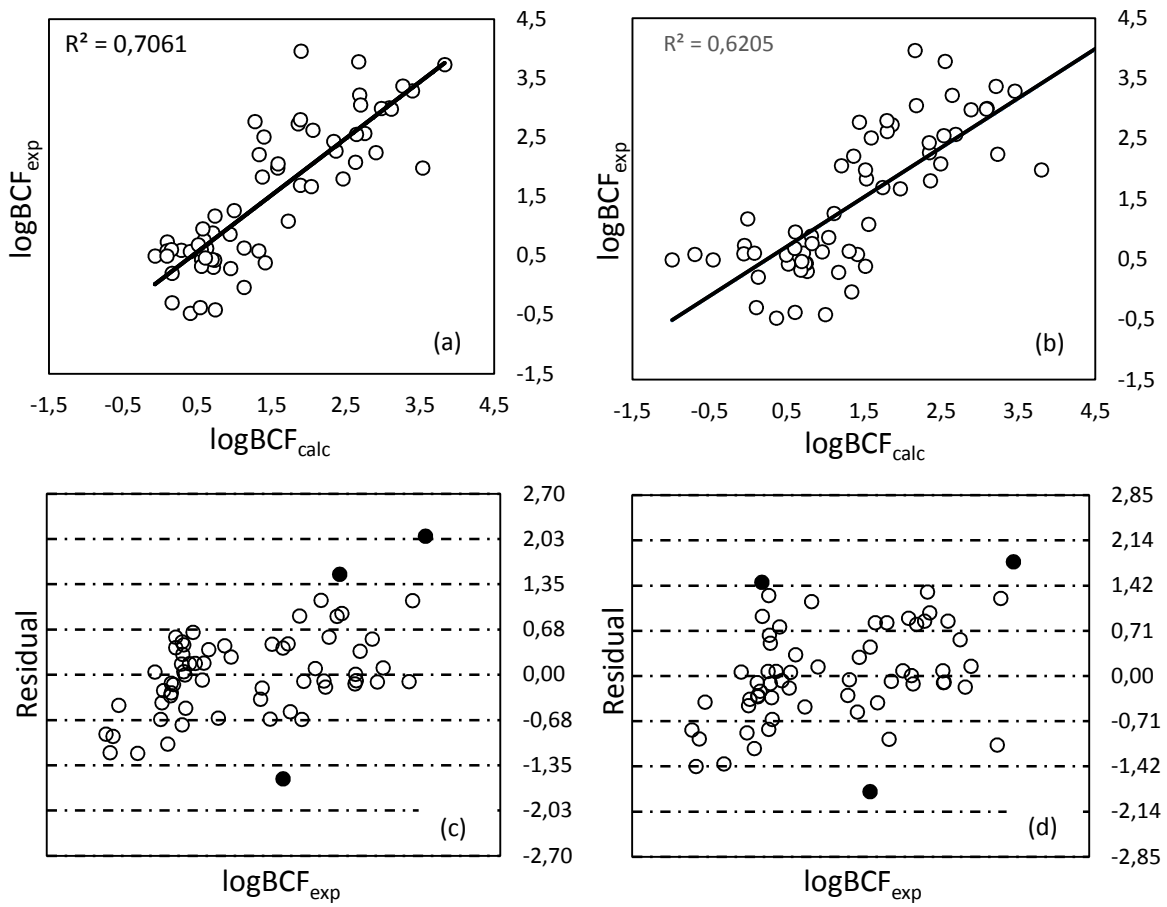


Figura 4.15 Log BCF experimental frente al calculado de las moléculas sin cloro: (a) aplicando la RNA al test externo, (b) aplicando el RML al test externo. Residual frente al log BCF experimental: (c) aplicando la RNA al test externo, (d) aplicando el RML al test externo. Los puntos en negrita representan los outliers del conjunto de datos.

Si nos fijamos en los outliers, vemos que el modelo de RNA tiene tres outliers que corresponden a las moléculas: bio69, bio289, bio476. Dos de ellas, precisamente la bio289 y la bio476, son las mismas encontradas en el modelo RML, esto quiere decir que dejando de lado los outliers, que son los mismos, el conjunto de datos es más representativo en el modelo RNA que en el modelo RML dado que tiene mejores parámetros estadísticos.
Haciendo un análisis cualitativo para individualizar los falsos negativos que se salen del umbral de ley (logBCF $\geq$ 3,3), [véase tabla 4.11 y tabla S(VII)], ambos los modelos tienen un error porcentual del 4,8% (fallan en 3 casos sobre 62), lo que no es muy representativo de la bondad de la RNA probablemente por el reducido número de moléculas bioacumulables que componen el test.

5. CONCLUSIONES

1. La topología molecular usada conjuntamente con la metodología QSAR, ha demostrado ser una metodología útil para la predicción de valores de bioconcentración (logBCF) en un grupo heterogéneo de moléculas orgánicas. Como técnicas para obtener los modelos se emplearon el análisis de regresión multilineal y las redes neuronales artificiales.

2. Los mejores modelos QSAR se obtuvieron separando los compuestos en dos grupos: clorados y no clorados. Otros factores, tales como la presencia o no de heteroátomos o el carácter aromático/alifático, no ejercieron una influencia relevante.

3. De las variables que aparecen en los modelos, se puede concluir que propiedades electrónicas, así como factores topológico-estructurales relativamente simples, como por ejemplo las posiciones relativas de heteroátomos tales como O, Cl y N juegan un papel destacado en el valor de BCF.

4. A la vista de lo anterior se demuestra que el método QSAR basado en la topología molecular aquí empleado, puede ser usado de modo eficaz para el seguimiento de la normativa europea REACH.

6. BIBLIOGRAFÍA

[1] Piir G, Sild S, Roncaglioni A, Benfenati E, Maran U. QSAR model for the prediction of bio-concentracion factor using aqueous solubility and descriptors considering various electronic effects. SAR and QSAR in Environmental Research. 2010; 21:711-729.

[2] Zhao C, Boriani E, Chana A, Roncaglioni A, Benfenati E. A new hybrid system of QSAR models for predicting bioconcentration factors (BCF) Chemosphere. 2008; 73:1701-1707.

[3] Arnot JA, Gobas FAPC. A review of bioconconcentration factor (BCF) and bioaccumulation factor (BAF) assessments for organic chemicals in aquatic organisms. Environmental Review. 2006; 14:257-297.

[4] Dimitrov S, Dimitrova N, Parkerton T, Comber M, Bonnell M & Mekenyan O. Base-line model for identifying the bioaccumulation potential of chemicals. SAR and QSAR in Environmental Research. 2005; 16:531-554.

[5] C.C. Ellsberry, Development of a Tiered Approach to Assess Bioaccumulation of Chemicals Project Committee, The Procter & Gamble Company.

[6] Pavan M, Netzeva T, Worth AP. Review of Literature-Based Quantitative Structure-Activity Relationship Models for Bioconcentration. QSAR Comb. Sci. 2008; 27:21-31.

[7] Cayley A. On the theory of the analytical forms called trees. Philosophical Magazine. 1857; 13:172-176.

[8] Amigo JM, et al. Topologia Molecular. Bol. Soc. Esp. Mat. Apl. 2007; 39:135–149.

[9] Anton Fos GM, et al. Predicción De Propiedades Fisico-Quimicas De Benzodiazepinas Por Topología Molecular. Anales de la Real Academia de Farmacia. 1995; Pag:485-495.

[10] Gago, F. Metodos Computacionales De Modelado Molecular y Diseno De Farmacos. Monografias de la Real Academia Nacional de Farmacia. 2009; Pag:253-306.

[11] R.Garcia Domenech y J.Galvez Alvarez. Química sostenible y Topología Molecular Parte primera: Conceptos, Master en Química Sostenible Universidad de Valencia.

[12] Balaban AT. Applications of graph theory in chemistry. J Chem Inf Comput Sci. 1985; 25:334-343.

[13] Hansen P, Jurs P. Chemical applications of graph theory. J Chem Ed. 1987; 65:574-580.

[14] Randić M. On characterization of molecular branching. J Am Chem Soc. 1975; 97:6609-6615.

[15] Kier LB, Murray WJ, Randić M, Hall LH. Molecular Connectivity I: Relationship to nonspecific local anesthesia. J Pharm Sci. 1975; 64:1971-1974.

[16] Hall LH, Kier LB, Murray WJ. Molecular Connectivity II: Relationship to water solubility an boiling point. J Pharm Sci. 1975; 64:1974-1978.

[17] Sylvester JJ. Application of the new atomic theory to the graphical representation of the invariant and covariants to binary quantics. Am. J. Math. 1874; 1:64-83.

[18] García Domenech R, Gálvez J, de Julián-Ortiz JV and Pogliani L. Some New Trends in Chemical Graph Theory. Chem. Rev. 2008; 108:1127-1169.

[19] Wiener H. Structural determination of paraffin boiling points J. Am. Chem. Soc. 1947; 69:17-20.

[20] Gálvez J, García-Domenech R, Salabert MT, Soler-Roca R. Charge Indexes. New topological descriptors. J. Chem. Inf. Compu. Sci. 1994; 34:520-525.

[21] Lilienblum W, Dekant W, Foth H, Gebel T, Hengstler JG, Kahl PR, et al. Alternative methods to safety studies in experimental animals: role in the risk assessment of chemicals under the new European Chemicals Legislation (REACH). Arch Toxicol. 2008; 82:211-236.

[22] Todeschini, R. Consonmi, V., Pavan, M. Dragon software versión 5.4, 2006.

[23] García-Doménech R, Montealegre MC, Nagham EG, Sandoval N, Santana M, Gálvez J. Aplicación de la topología molecular para la predicción de la actividad anti-VIH-1 de un grupo de compuestos análogos del aciclovir y ganciclovir. An. R. Acad. Nac. Farm. 2010; 76:45-57.

[24] Tichý M, Rucki M. Validation of QSAR models for legislative purposes. Interdisc Toxicol. 2009; 2:184-186.

[25] http://info.fisica.uson.mx/arnulfo.castellanos/archivos_html/quesonredneu.htm (última visita 23 junio 2013).

[26] Hardy ML, A comparison of the fish bioconcentration factors for brominated flame retardants with their nonbrominated analogues. Environ. Toxicol. Chem. 2004; 23:656-661.

[27]http://es.wikibooks.org/wiki/Disolventes_en_la_Industria_Qu%C3%ADmica/La_sustituci%C3%B3n_de_los_disolventes_en_la_Industria (última visita 25 junio 2013).

[28] Dimitrov S, Dimitrova NC, Walker JD, Veith GD, and Merkenyan OG. Predicting bioconcentration factors of highly for hydrophobic chemicals. Effects of molecular size. Pure Appl. Chem. 2002;74:1823-1830.