



UNIVERSIDAD CATÓLICA
SAN ANTONIO

UCAM

FACULTAD DE CIENCIAS DE LA SALUD,
LA ACTIVIDAD FÍSICA Y EL DEPORTE

CÁTEDRA DE INGENIERÍA Y TOXICOLOGÍA AMBIENTAL

“Nuevas aportaciones al desarrollo de modelos
QSAR/QSPR para la predicción de la mutagenicidad
de contaminantes ambientales y su interacción con
sustancias activas presentes en el medio.”

Lic. Alfonso Pérez Garrido

Directores:

Dr. Amalio Garrido Escudero

Dra. Aliuska Morales Helguera

Murcia, Mayo 2010

Índice general

1. INTRODUCCIÓN	7
1.1. Mutagénesis inducida por compuestos químicos	7
1.1.1. Ensayos toxicológicos usados para identificar y clasificar sustancias mutagénicas	12
1.2. Estudios QSAR en Toxicología	14
1.2.1. Consideraciones Generales	14
1.2.2. Principios de la OECD para la validación de modelos QSAR con fines regulatorios	15
1.2.3. Estudios QSAR para la predicción de mutagenicidad	17
1.3. Problemática de sustancias mutagénicas desconocidas presentes en el medio ambiente	22
1.3.1. Ácidos haloacéticos	23
1.3.2. Carbonilos α, β -insaturados	28
1.4. Complejación con la β -ciclodextrina	31
1.5. Importancia de la complejación de sustancias químicas con β -ciclodextrina	32
1.6. Estudios teóricos para la determinación de la complejación con β -ciclodextrina	35
2. PLANTEAMIENTO Y OBJETIVOS	39
2.1. Planteamientos	39
2.2. Objetivos	40

3. RESULTADOS Y DISCUSIÓN	43
3.1. Ácidos haloacéticos	43
3.2. Carbonilos α, β -insaturados	48
3.2.1. Mutagenicidad en el ensayo de Ames	49
3.2.2. SAs y comparación con el sistema de expertos TOXTREE	56
3.2.3. Mutagenicidad en células de mamífero	57
3.3. Complejación con la β -CD	60
3.4. Predicciones de complejación con la β -CD y mutagenicidad para los ácidos haloacéticos y los monómeros dentales	65
4. CONCLUSIONES	69
4.1. Recomendaciones	70
I ANEXO	73

INTRODUCCIÓN

1.1. Mutagénesis inducida por compuestos químicos

La interacción de ciertos compuestos químicos presentes en el medio ambiente con el ácido desoxiribonucleico (DNA) puede provocar cambios genéticos debido a modificaciones en su estructura, afectando a uno o más genes¹. Estas mutaciones químicamente inducidas se conocen como mutagénesis química, a los productos químicos capaces de inducir éstas se denominan sustancias mutagénicas o genotóxicas. Muchos tipos de cáncer² y otras enfermedades degenerativas son resultado de mutaciones genéticas adquiridas debido a la exposición al medio ambiente, y no como un resultado de los rasgos hereditarios.

El poder mutagénico de una sustancia depende de su capacidad para penetrar en la célula, su reactividad con el DNA, su toxicidad general, y la probabilidad de que el tipo de cambio químico que introduce en el DNA sea corregido por un sistema de reparación. Hay cientos de mutágenos químicos conocidos pudiendo ser éstos indirectos o directos, dependiendo de si requieren o no de activación metabólica por las enzimas celulares para producir la especie final, la cual interacciona con el DNA. Entre las sustancias de acción directa se encuentran: N-metil-N-nitrosourea, epóxidos, etc y entre las de acción indirecta está el benzo[α]pireno, aflatoxina B1, etc. Las sustancias mutagénicas se pueden dividir también en varias clases dependiendo de su interacción con el DNA^{3,4}:

- Análogos de bases. Son moléculas cuya estructura química es similar a una de

las cuatro bases del DNA (adenina, timina, citosina y guanina). Debido a esta semejanza, pueden ser incorporados en la hélice durante la replicación del DNA. Una característica clave de estas sustancias es que forman pares de bases con más de una base. Esto puede causar mutaciones en la próxima ronda de replicación, cuando la maquinaria de replicación intenta emparejar una base con el incorporado mutágeno. Por ejemplo, 5-bromo-deoxiuridina (5-BU) existe en dos formas diferentes, una con similitud a la timina y por lo tanto emparejada con la adenina durante la replicación, mientras que la otra, es parecida a la citosina y por lo tanto emparejada con la guanina. En su forma de timina, 5-BU puede ser incorporada a través de una adenina. Si se convierte en la forma ionizada (parecida a la citosina) durante la siguiente ronda de replicación, esto causará que una guanina entre en la cadena opuesta en lugar de la correcta adenina ocasionando la transición de AT a GC (Figura 1.1).

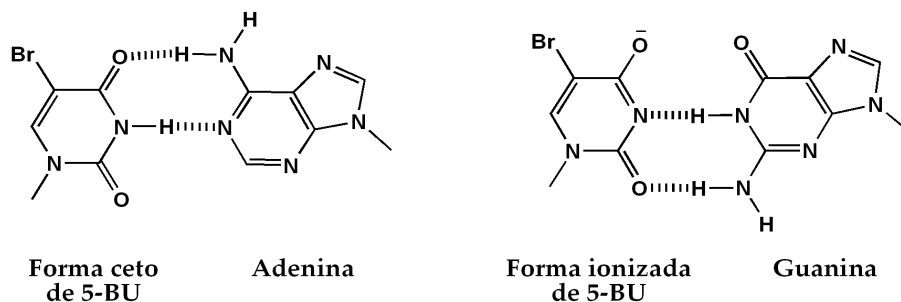


Figura 1.1: Pares de bases posibles con 5-BU

- Alteran las bases. Estos mutágenos provocan cambios químicos en las bases que forman parte del DNA. La gran mayoría de estos cambios se producen de tres maneras:

1. Deaminación. La deaminación consiste en la eliminación del grupo amino de la adenina o citosina para dar hipoxantina o uracilo, respectivamente. Debi-

Los agentes alquilantes con dos o más centros electrófilos (cross-linking agents) pueden generar enlaces cruzados entre dos o más centros nucleófilos del DNA⁹⁻¹¹. Un ejemplo de agentes alquilantes lo tenemos en etil metano sulfonato (EMS) que introduce un metilo en la guanina que ya no se aparea con la citosina provocando la transición GC → AT. En este grupo también tenemos a los hidrocarburos aromáticos policíclicos, los cuales se forman en grandes cantidades en el humo del tabaco. Entre éstos el más representativo es el benzo[α]pireno¹², el cual actúa combinándose con el DNA para formar grupos voluminosos que interrumpen la replicación (Figura 1.3).

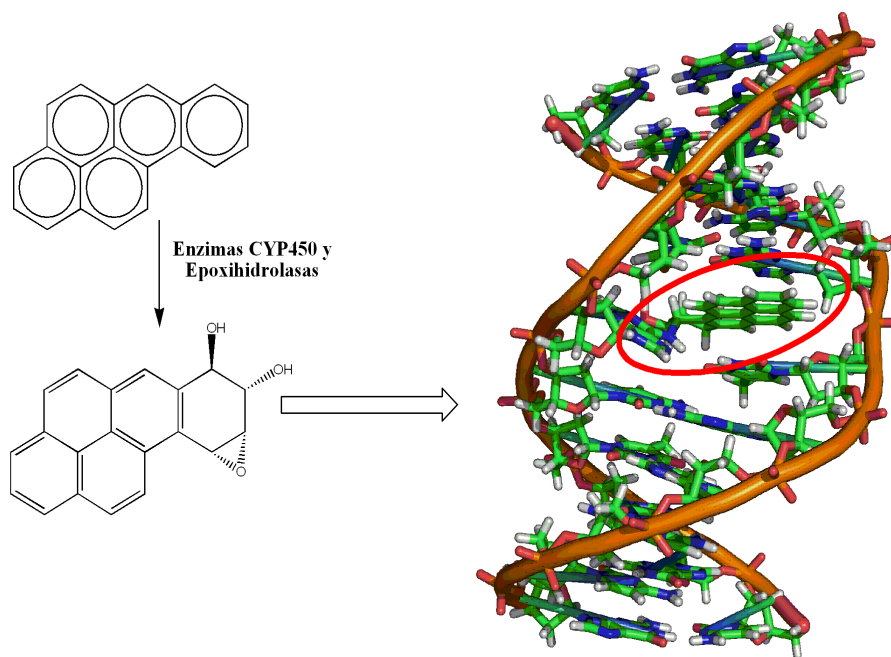


Figura 1.3: Mecanismo de mutagenicidad del benzo[α]pireno

3. Hidroxilantes. Los N-hidroxycarbamatos y las hidroxioreas son agentes reductores que forman radicales libres en presencia de oxígeno y trazas de metales. Los radicales libres son compuestos en los que un átomo, generalmente

de oxígeno, tiene un electrón desapareado en capacidad de aparearse, por lo que son muy reactivos y pueden causar varios tipos de daño al DNA¹³. Reaccionan preferentemente con la citosina produciendo prioritariamente los derivados N-hidroxilados en las posiciones 3 y 4.

- Agentes de intercalantes. Son moléculas planas que se insertan entre las bases adyacentes de la doble hélice distorsionándolas e interfiriendo en la replicación, transcripción, reparación y recombinación del DNA. Cuando esto ocurre, la DNA polimerasa puede añadir una base adicional frente al agente intercalante. Si esto ocurre en un gen, induce una mutación por corrimiento de lectura (es decir, que altera la lectura de la transcripción del gen, cambiando los aminoácidos que serán añadidos a la proteína codificada de acuerdo con el código genético). El bromuro de etidio es un agente de este tipo, ampliamente utilizado en la investigación de DNA. A menudo se utiliza en los laboratorios de bioquímica para visualizar fragmentos de DNA que han sido separados en geles. La molécula de etidio es fluorescente y cuando se excita con luz ultravioleta emite en el rango del visible (Figura 1.4)¹⁴.

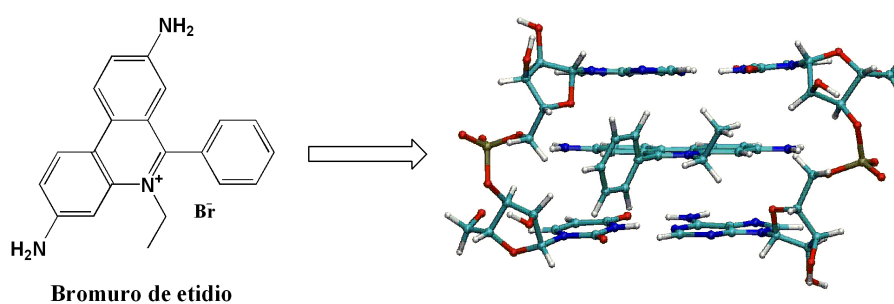


Figura 1.4: Bromuro de etidio intercalado entre dos pares de bases adenina-uracilo

1.1.1. ENSAYOS TOXICOLÓGICOS USADOS PARA IDENTIFICAR Y CLASIFICAR SUSTANCIAS MUTAGÉNICAS

Para la identificación de sustancias mutagénicas existen varios ensayos definidos para uso regulatorio por la Organización para la Cooperación Económica y el Desarrollo (OECD). Entre los más utilizados están: Mutagénesis *in vitro* en bacterias¹⁵, Mutagénesis *in vitro* en células de mamífero¹⁶, aberraciones cromosómicas *in vivo*¹⁷ e *in vitro*¹⁸ y el test de micronucleos *in vivo*¹⁹.

- Mutagénesis *in vitro* en bacterias. Este ensayo es ampliamente utilizado para propósitos de *screening* de sustancias mutagénicas y carcinogénicas. Combina una alta sensibilidad con una relativa facilidad técnica, rapidez y economía. En este ensayo se utilizan diferentes cepas mutantes (incapaces de sintetizar histidina) de *Salmonella typhimurium*. Cada una de estas cepas tiene diferentes mutaciones que desactiva el gen que codifica la enzima requerida en la síntesis de este aminoácido vital. De manera que no pueden crecer en un cultivo a no ser que el medio esté suplementado con este aminoácido. Si el gen afectado es mutado se produce una reversión al estado salvaje u original y entonces la bacteria será capaz de crecer en ausencia del aminoácido. Este fenómeno es conocido como reversión y las colonias como revertantes.
- Mutagénesis *in vitro* en células de mamífero. Estos ensayos son utilizados para confirmar si un presunto mutágeno lo es para mamíferos superiores como los humanos. Las células de mamífero presentan un mayor grado de organización que las bacterias y su capacidad metabólica y de reparación del DNA también es mucho más compleja. Están admitidas las células de hámster chino (CHO, AS52 y V79), células de linfoma de ratón (L5178Y) y células linfoblastoides humanas TK6. Células deficientes en timidinaquinasa (TK) debido a la mutación $TK^{+/-} \rightarrow TK^{-/-}$ son resistentes a los efectos citotóxicos de la trifluorotimidina (TFT), mientras que las células normales son sensibles a la TFT, que inhibe el metabo-

lismo celular y detiene la división celular. Así, las células mutantes son capaces de proliferar en presencia de TFT, mientras que las células normales, que contienen timidinaquinasa, no lo son. El procedimiento consiste en exponer a las células al compuesto en estudio con y sin activación metabólica e incubarlas por un período de tiempo que permita la expresión de cualquier mutación que conlleve a la transformación a células homocigóticas $TK^{-/-}$ (enzima infuncional), las cuales son resistente a la TFT y permanecen viables en presencia de esta sustancia.

- Aberraciones cromosómicas *in vitro*. El propósito del ensayo de aberraciones cromosómicas *in vitro* tiene por objeto detectar agentes que provocan aberraciones cromosómicas estructurales en los cultivos de células de mamífero. Las aberraciones estructurales pueden ser cromosómicas o cromatídicas. El test consiste en exponer a los cultivos celulares a la sustancia en ensayo, con y sin activación metabólica. A intervalos predeterminados, después de la exposición, son tratados con una sustancia que detenga la metafase (por ejemplo, colchicina). Se recolectan las células, se tiñen y se observan al microscopio en metafase para detectar la presencia de aberraciones cromosómicas.
- Aberraciones cromosómicas *in vivo*. En esta prueba los animales (roedores) se exponen a la sustancia de ensayo por una vía adecuada y son sacrificados a intervalos apropiados tras el tratamiento. Antes de sacrificar los animales, éstos son tratados con un agente que detiene la metafase (por ejemplo, colchicina). Se realizan preparaciones de cromosomas de las células de la médula ósea en metafase que, después de teñir, se observan al microscopio para detectar aberraciones cromosómicas.
- Test de micronucleos *in vivo*. Este ensayo se utiliza para la detección de lesiones provocadas por la sustancia en los cromosomas o el aparato mitótico de eritroblastos, mediante el análisis de eritrocitos tomados de la médula ósea o la sangre periférica de animales, por lo general roedores.

Todos ellos identifican sustancias capaces de producir alguna alteración en el material genético. En los ensayos *in vitro* suele ser necesario el uso de una fuente exógena de activación metabólica a modo de imitar las condiciones *in vivo*. El sistema más comúnmente utilizado es un cofactor suplementario de una fracción post-mitocondrial (S9) a partir de hígados de roedores tratados con agentes inductores enzimáticos como el pesticida Aroclor 1254²⁰⁻²⁴ o una combinación de fenobarbital y β -naphthoflavone²⁵⁻²⁸.

1.2. Estudios QSAR en Toxicología

1.2.1. CONSIDERACIONES GENERALES

Los estudios QSAR (Quantitative Structure-Activity Relationship) permiten establecer una relación entre la estructura de una familia de sustancias y su actividad biológica/toxicológica a través de una expresión matemática.

En las últimas décadas se han desarrollado multitud de modelos QSAR relacionando la actividad biológica de sustancias que tienen el mismo mecanismo de acción a través de descriptores moleculares simples. De ahí, que tengan una gran importancia como herramienta para predecir actividades de compuestos desconocidos por su rapidez y simplicidad. El desarrollo y aplicación de los modelos QSAR se consideró necesario para resolver los problemas de bienestar animal, para sustituir, reducir y perfeccionar el uso de animales en las evaluaciones toxicológicas, como exige la directiva europea sobre la protección de los animales de laboratorio²⁹. Por otro lado, en Octubre de 2003, la Comisión Europea (CE) aprobó una propuesta legislativa³⁰ para un nuevo sistema de gestión de productos químicos denominada REACH (Registro, Evaluación y Autorización de Productos Químicos), destinada a armonizar los requisitos de información de los productos químicos nuevos y existentes. Dada la cantidad de sustancias que se caracterizan y las dificultades encontradas en los ensayos, existe un interés considerable en la predicción de las actividades toxicológicas a través de la estructura molecular, además de, como medio de sustitución y reducción de los ensayos experimentales con

animales, como screening virtual de compuestos rápido y sencillo. Por lo tanto, en noviembre de 2004, la Comisión Europea y la OECD de los Países Miembros acordaron el desarrollo de una serie de principios para la validación de los modelos QSAR para su uso en la Reglamentación REACH³¹.

1.2.2. PRINCIPIOS DE LA OECD PARA LA VALIDACIÓN DE MODELOS QSAR CON FINES REGULATORIOS

La estrategia a seguir para el desarrollo de un modelo consiste en varios pasos iterativos, basados en diseños experimentales estadísticos y análisis de datos multivariante. La OECD publicó en 2007 una guía para la validación de los modelos QSAR³², según la cual, para facilitar el examen de un modelo QSAR para fines de regulación, éste debería tener presente la siguiente información:

- Punto final definido. Punto final se refiere a cualquier efecto fisicoquímico, biológico o medioambiental que pueda ser medido y por lo tanto modelado. La intención de este principio es garantizar la transparencia en el punto final predicho por el modelo, dado que un mismo punto final puede ser determinado por diferentes protocolos experimentales y bajo diferentes condiciones. Idealmente, un QSAR debe ser desarrollado a partir de conjuntos de datos homogéneos en los que los datos experimentales han sido generados por un único protocolo. Sin embargo, esto rara vez es viable en la práctica, y los datos producidos por los diferentes protocolos a menudo se combinan.
- Algoritmo no ambiguo. Un QSAR debe expresarse en forma de un algoritmo que no presente más de una interpretación. La intención de este principio es garantizar la transparencia en la descripción del algoritmo de modelo. En el caso de los modelos desarrollados comercialmente, esta información no siempre está a disposición del público.
- Dominio de aplicación definido. Un modelo QSAR debe estar asociado a un do-

minio determinado de aplicabilidad. La necesidad de definir un ámbito de aplicabilidad expresa el hecho de que los QSAR son modelos reduccionistas que están inevitablemente asociados con limitaciones en cuanto a los tipos de estructuras químicas, propiedades físico-químicas y los mecanismos de acción para la que los modelos pueden hacer predicciones fiables. Este principio no implica que un determinado modelo sólo deba ser asociado a un dominio de aplicación único. Los límites del dominio pueden variar de acuerdo al método utilizado en su definición y el equilibrio deseado entre la amplitud de la aplicabilidad del modelo y la fiabilidad global de las predicciones.

- Medidas apropiadas de bondad de ajuste, robustez y predictividad. Un QSAR debe estar asociado con las medidas adecuadas de bondad de ajuste, robustez y valor predictivo. Este principio expresa la necesidad de proporcionar dos tipos de información: a) el rendimiento interno de un modelo (representado por la bondad de ajuste y robustez), determinado mediante un conjunto de entrenamiento, y b) el valor predictivo de un modelo determinado mediante el uso de un conjunto de pruebas adecuadas. No hay ninguna medida absoluta de valor predictivo que sea conveniente para todos los efectos, el valor predictivo puede variar de acuerdo a los métodos estadísticos y los parámetros utilizados en la evaluación.
- Interpretación del mecanismo, si es posible. Un QSAR debe estar asociado con una interpretación del posible mecanismo de actuación, siempre que dicha interpretación se pueda hacer. Evidentemente, no siempre es posible dar una interpretación del mecanismo de un determinado modelo QSAR. La intención de este principio es asegurar que existe una asociación entre el mecanismo, los descriptores utilizados en el modelo y el punto final predicho, y que se documenta dicha asociación. Cuando una interpretación mecanista es posible, puede formar parte también del dominio de aplicación definidos.

En estos últimos años la Unión Europea está creando una base de datos de modelos QSAR (http://qsar.db.jrc.it/qmrf/?order=qmrf_number&changedirection=true), para su uso regulatorio bajo el marco del Reglamento REACH, donde la información está estructurada de acuerdo a los principios de validación que acabamos de explicar. La base de datos presenta unos QMRF (QSAR Model Reporting Format) para cada modelo. Estos informes resumen y presentan información clave sobre los modelos (Q)SAR, incluidos los resultados de los estudios de validación.

1.2.3. ESTUDIOS QSAR PARA LA PREDICCIÓN DE MUTAGENICIDAD

Entre los modelos teóricos de predicción *in silico*, podemos encontrar las relaciones (cuantitativas) estructura actividad: (QSAR) y sistemas de expertos basados en reglas como: CASE³³, MULTICASE^{34-36, 36-41}, TOPKAT^{42, 43}, ADAPT⁴⁴⁻⁴⁶, DEREK⁴⁷⁻⁴⁹ y de más reciente creación el TOXTREE con un módulo específico para carcinogénesis y mutagénesis^{50, 51}.

- En Computer Automated Structure Evaluation (CASE) cada estructura molecular es descompuesta por el programa en todos los fragmentos constituyentes posibles de 2-10 átomos pesados contiguos de longitud, con todos sus hidrógenos y una posible cadena lateral. El análisis estadístico del sistema de fragmentos generados por la descomposición de todas las moléculas implica el análisis de la distribución de cada fragmento único entre las moléculas activas e inactivas e identificación de los fragmentos cuya distribución se desvíe de una distribución binomial simétrica ideal: cada uno de los fragmentos que se desvían perceptiblemente de la distribución de referencia se etiqueta como bioforo (fragmento que activa) o un biofobo (fragmento que inactiva). Bioforos y los biofobos son los descriptores moleculares primarios del modelo CASE QSAR.
- MULTIPLE Computer Automated Structure Evaluation (MULTICASE) es un desarrollo del programa CASE, se construyó o a partir de los problemas expuestos

por CASE. Particularmente, MULTICASE responde al problema de distinguir entre los fragmentos que provocan la actividad y los fragmentos que modulan la actividad. En términos más generales, procura hacer frente a la presencia de jerarquías y de la no linealidad dentro de modelos SAR en relación a sistemas no congénicos de productos químicos. Como CASE, MULTICASE comienza creando su propio diccionario de descriptores directamente de la base de datos. En este punto, y en contraste con CASE, MULTICASE selecciona como bioforo el estadístico más importante de estos fragmentos, creyéndosele responsable de la actividad observada de las moléculas que lo contienen, y separa de la base de datos restante todas las sustancias que contienen este bioforo. Este proceso se repite en la base de datos restante con el bioforo más significativo siguiente, y así sucesivamente, hasta que la base de datos se divide en segmentos de clases químicas que contienen un importante bioforo. El análisis CASE se aplica entonces a cada clase de bioforo por separado para determinar modificaciones subestructurales en la actividad de este.

- Toxicity Prediction by Komputer Assisted Technology (TOPKAT) es un sistema informático automatizado que consiste en diversos módulos para la predicción de una variedad de efectos tóxicos agudos y crónicos (carcinogénesis en roedores y mutagénesis en *S. typhimurium*, pero también toxicidad de desarrollo, irritación de la piel y de los ojos, toxicidad oral aguda, LD50, etc). Cada modelo se ha derivado de una base de datos específica. Mientras que en versiones anteriores dependen de la presencia o ausencia de fragmentos estructurales, versiones más recientes usan descriptores con valores continuos. Estos incluyen índices topológicos y electrotopológicos (E-state) en fragmentos de uno o dos átomos. TOPKAT realiza el análisis de la sustancia en cuatro pasos. En el primer paso, TOPKAT identifica fragmentos en el producto químico en análisis que no se encuentra en los compuestos del sistema de entrenamiento. En el segundo paso comprueba si el producto químico se ajusta dentro del espacio óptimo de predicción (OPS) de la ecuación de estima-

ción. Esto permite al usuario determinar si la estructura en cuestión está contenida en el espacio del descriptor del modelo: un producto químico que está fuera del OPS tendrá poca confianza a la hora de predecir su toxicidad. El tercer paso determina la toxicidad del producto químico. En el cuarto paso, TOPKAT permite que el usuario realice otra prueba independiente con una búsqueda de similitudes en la base de datos. La base de datos de entrenamiento se explora para moléculas similares a la molécula en cuestión, para determinar independientemente la posible significancia química o biológica de las asociaciones del modelo. En la práctica, el usuario puede comprobar la toxicidad real de los productos químicos *similares* y la exactitud de la predicción de TOPKAT generada para determinar la confiabilidad del modelo.

- En Automated Data Analysis and Pattern Recognition Toolkit (ADAPT) cada compuesto es representado por descriptores moleculares calculados que codifican aspectos topológicos, electrónicos, geométricos o fisicoquímicos. Los descriptores topológicos utilizan solamente la tabla de las conexiones de una molécula y por lo tanto no requieren las estructuras tridimensionales optimizadas. Estos descriptores codifican información acerca de tipos de átomos, tipos de enlaces, índices de conectividad, y distancias interatómicas. Ellos se correlacionan con el tamaño molecular, forma, y grado de ramificación. Los descriptores geométricos, los cuales codifican la información sobre el tamaño y la forma total de una molécula, requieren geometrías tridimensionales. Ejemplos de descriptores electrónicos son las cargas atómicas parciales, momentos dipolares, energías de repulsión electrón-núcleo, y áreas superficiales parciales cargadas. Finalmente, el modelo QSAR para la actividad biológica de interés se encuentra aplicando por separado diversos métodos de reconocimiento de patrones (en artículos más recientes, análisis discriminante lineal, redes neuronales, etc) y seleccionando los modelos que mejor se ajusten.
- Deductive Estimation of Risk from Existing Knowledge (DEREK), está basado en reglas (del tipo *if-then-else*) asociadas a grupos funcionales particulares, o alarmas

estructurales, con varias formas de toxicidad. Las características estructurales utilizadas en la predicción se le llaman toxicóforos. Los límites toxicológicos cubiertos actualmente por el sistema DEREK incluyen carcinogénesis, mutagénesis, sensibilización de la piel, irritación, teratogénesis, y neurotoxicidad. Cada punto final de toxicidad tiene una serie de reglas y un sistema de toxicóforos. Se introduce la estructura del compuesto en el programa, y este mediante un sistema de reglas busca comparando las características estructurales de la molécula en cuestión con los toxicóforos descritos en su base de datos. Cualquier alarma estructural situada dentro de la estructura de la molécula en cuestión se destaca, y se proporciona un mensaje que indica la naturaleza del peligro toxicológico.

- TOXTREE posee varios módulos de predicción, de entre los cuales nos vamos a centrar en el dedicado a mutagenesis y carcinogénesis realizado por Benigni *et al.*^{50, 51}. Este módulo consiste en un sistema de reglas basado en la presencia o no de una serie de alertas estructurales (SAs). Las SAs se definen como aquellos grupos funcionales o subestructuras que están ligadas con la actividad genotóxica o carcinogénica. En total contiene un listado de 33 SAs, cinco de ellas referidas a mecanismos de acción no-genotóxicos. El procesamiento de una sustancia por el módulo es el siguiente: a) presencia de SAs para carcinogénesis; b) una o mas SAs son reconocidas; c) si se reconocen SAs relativas a aminas aromáticas o aldehídos α , β -insaturados, la sustancia es analizada mediante modelos QSAR los cuales dan un resultado positivo o negativo. Los resultados finales son uno o combinaciones de unas pocas etiquetas del tipo:
 - Ninguna alerta de actividad carcinogénica
 - Alerta estructural de carcinogénesis genotóxica.
 - Alerta estructural de carcinogénesis no-genotóxica.
 - Mutagénico en *S. typhimurium* cepa TA100 basado en QSAR (potencial carcinógeno o improbable)

Los propios autores, en un análisis realizado para obtener el rendimiento del módulo empleando para ello una extensa base de datos⁵², encontraron una baja selectividad carcinogénica para las SAs correspondientes a los bencenos y dibenzodioxinas halogenadas y una baja selectividad mutagénica para las SAs correspondientes a los carbonilos α , β -insaturados y aldehídos simples, recomendándose un estudio exhaustivo de sus factores de modulación.

La reactividad de una SA o toxicóforo puede ser modulada o suprimida por el resto de la molécula en la que está inmersa. A grandes rasgos, el efecto de modulación puede ser representado por otras subestructuras moleculares (por ejemplo, grupos voluminosos en orto a un grupo amino aromático) que se sabe que tienen una influencia sobre la reactividad de la SA. Una generalización de gran alcance es la que proporcionan los análisis QSAR, que producen un modelo matemático que vincula la actividad biológica a un número limitado de propiedades físicas, químicas u otras propiedades moleculares (descriptores).

Los estudios QSAR realizados para mutagenicidad los podemos dividir en dos grupos: aquellos basados en series congenéricas de sustancias (p. ej. Aminas aromáticas, aldehídos, etc..) y los basados en series no congenéricas. Unas buenas revisiones al respecto las tenemos en los trabajos de Benigni *et al.*^{53, 54}. En general, los modelos basados en series no congenéricas de sustancias son inferiores en calidad a los QSARs clásicos de series congenéricas. Esto no es sorprendente, si tenemos en cuenta que un modelo QSAR para series congenéricas está dirigido a modelar un solo fenómeno, mientras que los modelos generales de series no congenéricas de sustancias, aunque poseen un mayor dominio de aplicación, intentan modelar al mismo tiempo, varios mecanismos de acción, cada uno en relación a una clase determinada de agentes mutagénicos, obteniendo así una menor precisión.

Entre las familias de sustancias más estudiadas a través de estos modelos QSAR se encuentran las aminas aromáticas y los nitroaromáticos. Las aminas aromáticas tienen una importancia medioambiental e industrial grande, por lo que hay disponible

una gran base de datos de resultados experimentales y QSAR diferentes⁵⁵⁻⁷¹. En líneas generales las conclusiones obtenidas con estos modelos QSAR relacionan la hidrofobicidad y la facilidad de ser oxidadas y, por lo tanto, bioactivadas, con la mutagenicidad, en concordancia con los mecanismos de acción mutagénica de esta familia de sustancias. Los nitroarenos han sido de importancia como intermediarios en la síntesis de varios tipos de sustancias químicas industriales, además, están presentes en diversas matrices de importancia ambiental (por ejemplo, humos de automóviles). Los QSAR para nitroarenos relacionan, igualmente, la hidrofobicidad y, al contrario de las aminas, la mayor facilidad para ser reducidas de acuerdo con los mecanismos conocidos de bioactivación reductiva que presentan estas sustancias⁷²⁻⁷⁷. Otras sustancias de interés toxicológico y escasamente estudiadas son los ácidos haloacéticos y los carbonilos α , β -insaturados de los cuales hablaremos más adelante en los siguientes apartados.

1.3. Problemática de sustancias mutagénicas desconocidas presentes en el medio ambiente

Diariamente estamos expuestos a sustancias químicas desconocidas derivadas de la actividad humana. Se ha estimado que hay más de cinco millones de productos químicos artificiales conocidos, de los cuales 140.000 sustancias diferentes están en uso hoy en día y pre-registradas de acuerdo al cumplimiento del Reglamento REACH. La aplicación de la mejora continua de los métodos de análisis ha revelado que muchos de estos productos químicos pueden entrar en la cadena alimentaria y dar lugar a exposiciones humanas. Con el paso del tiempo y debido a la preocupación en este sentido, se han ido identificando sustancias en diversas matrices medioambientales y determinando su toxicidad. Aún así, la investigación en este sentido está todavía en pañales. Es conocido que la introducción de los procesos de desinfección de agua fue un importante éxito en el control de enfermedades transmitidas por el agua⁷⁸. La desinfección del agua potable es necesaria para eliminar los contaminantes dañinos, incluidos los microorganismos patógenos. Hace algunas décadas se desconocía que en el agua potable

existieran sustancias que pudieran producir efectos tóxicos a largo plazo. Posteriormente, estudios realizados en extractos concentrados de agua potable desinfectada con cloro fueron tóxicos en muchos bioensayos *in vivo* e *in vitro*⁷⁹⁻⁸² y, hasta ese momento, las sustancias responsables de dicha actividad no habían sido identificadas. Además, estudios epidemiológicos demostraron que las personas que consumen agua potable con cloro están expuestas a un mayor riesgo de desarrollar un cáncer de estómago, páncreas, riñones, vejiga y recto, así como de Hodgkin y el linfoma no Hodgkin⁸³⁻⁸⁵. Hoy en día se conocen varias familias de compuestos responsables de dicha toxicidad, entre ellos los ác. haloacéticos. Otra preocupación importante en este sentido la presentan los carbonilos α , β -insaturados, más concretamente los empleados como monómeros dentales. Éstas sustancias presentaban valores negativos en los ensayos de mutagenicidad realizados en bacterias⁸⁶ pero, hace pocos años, se vió que uno de los más comúnmente empleados, el trietilenglicol dimetacrilato (TEGDMA), causa deleciones en el DNA de células de mamífero^{87,88}. Obviamente, la toxicología experimental no es una herramienta práctica para hacer frente a situaciones que requieren una rápida toma de decisiones. Además, incluso si se dispone de las instalaciones suficientes para llevar a cabo pruebas toxicológicas en un plazo pertinente, todavía puede cuestionarse si los ensayos de un gran número de sustancias sería un planteamiento racional y práctico. En este contexto, los modelos de predicción *in silico* tienen ventajas evidentes en términos de tiempo, costo, y la protección de los animales.

1.3.1. ÁCIDOS HALOACÉTICOS

Los subproductos de la desinfección del agua potable (DBPs) representan una importante clase de sustancias químicas peligrosas para la salud y el medio ambiente a largo plazo. Las cuestiones epidemiológicas para la evaluación de riesgos para la salud sobre la exposición humana a DBPs fueron revisadas recientemente⁸⁹.

La mayoría de los subproductos de la desinfección se forman como resultado de la reacción entre la materia orgánica del agua cruda y los desinfectantes químicos co-

mo el cloro. Estos compuestos orgánicos provienen de productos de la descomposición de materiales naturales (NOM), que incluyen ácidos húmicos, microorganismos y sus metabolitos, y algunos derivados del petróleo de alto peso molecular, como los hidrocarburos alifáticos y aromáticos.

Después de los trihalometanos, los ácidos haloacéticos son el segundo mayor grupo de subproductos de la desinfección del agua potable (Tabla 1.1).

Tabla 1.1: Subproductos formados por la cloración del agua potable⁹⁰.

Sub-producto de oxidación	Concentración media ($\mu\text{g/L}$)	Concentración máxima ($\mu\text{g/L}$)
TRIHALOMETANOS		
cloroformo	25	240
bromodiclorometano	9.5	90
dibromoclorometano	1.6	36
bromoformo	<0.2	7.1
ACIDOS HALOACÉTICOS		
ácido dicloroacético	15	74
ácido tricloroacético	11	85
ácido bromocloroacético	3.2	49
ácido cloroacético	1.3	5.8
ácido dibromoacético	<0.5	7.4
ácido bromoacético	<0.5	1.7
ácido tribromoacético	–	–
ácido bromodicloroacético	–	–
ácido clorodibromoacético	–	–
HALOACETONITRILOS		
dicloroacetoniitrilo	2.1	10
bromoacetoniitrilo	0.7	4.6
bromocloroacetoniitrilo	0.6	1.1
dibromoacetoniitrilo	<0.5	9.4
tricloroacetoniitrilo	<0.02	0.02
tribromoacetoniitrilo	–	–

Estos compuestos están basados en la molécula del ácido acético ($\text{CH}_3\text{CH}_2\text{COOH}$),

en la que uno o más átomos de hidrógeno unidos a los átomos de carbono son reemplazados por un elemento halógeno (cloro, bromo, flúor y/o yodo, ver Figura 1.5). Existen treinta y cuatro especies de ácidos haloacéticos (HAA) entre las que se incluyen el ácido cloroacético (CA), el ácido dicloroacético (DCA), y el ácido dibromoacético (DBA).

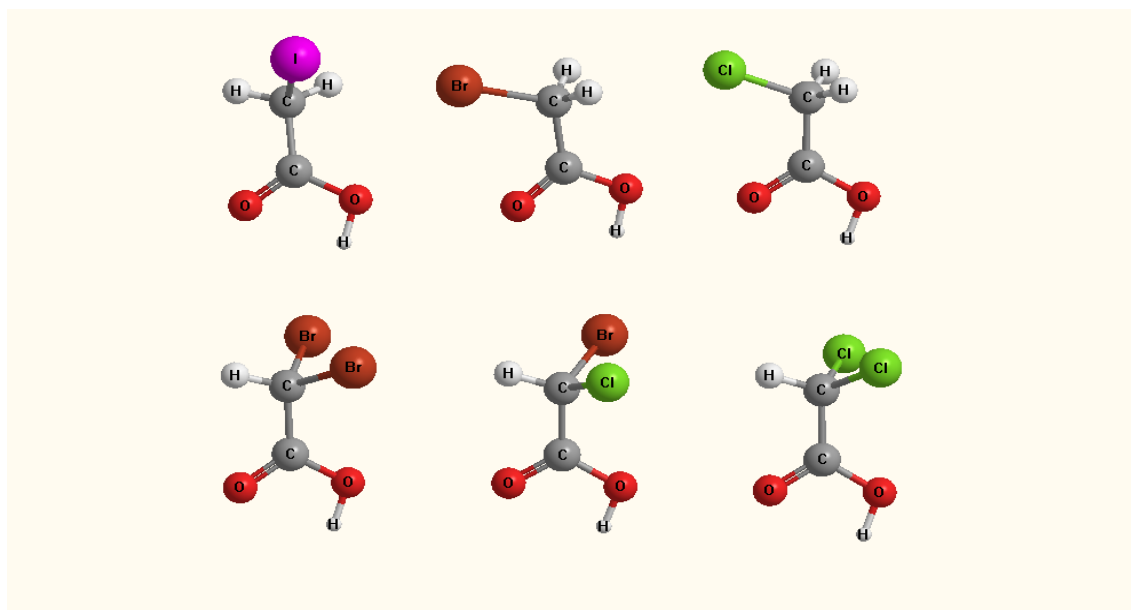


Figura 1.5: Representación estructural de algunos ácidos haloacéticos.

De entre éstos podemos encontrar el ácido dicloroacético (DCA) y tricloroacético (TCA) que son hepatocarcinogénicos en roedores⁹¹⁻⁹³. El ácido cloroacético (CA) no es mutagénico en la cepas de *S. typhimurium* TA1535, pero si en TA98 y TA1538⁹⁴. DCA, pero no TCA⁹⁵, es débilmente mutagénico con o sin activación metabólica S9 en *S. typhimurium* TA100⁹⁶. Giller *et al.*⁹⁷ con el test de fluctuación en cepas TA100 demostraron genotoxicidad de CA, DCA, TCA, Ac. Bromo- (BA), Dibromo- (DBA) y Tribromoacético (TBA). Estudios recientes han demostrado que los ácidos haloacéticos bromados son más citotóxicos y genotóxicos en *S. typhimurium*^{98,99} y en células ováricas de hamsteres chinos (CHO) que sus análogos clorinados¹⁰⁰⁻¹⁰². También se ha encontrado la presen-

cia de ácidos iodoacéticos en aguas con altos contenidos de bromuro/ioduro^{100, 103} y se ha visto que el ácido iodoacético es 2.6 y 523.3 veces más mutagénico en *S. typhimurium* TA100 que BA y CA, respectivamente.

Entre los modelos QSAR para los ácidos haloacéticos podemos encontrar estudios de Richard *et al.*¹⁰⁴ que relacionan la concentración de Benchmark BC_m (calculada como el límite de confianza menor del 95 % de concentración que produce un incremento del 5 % en el número de embriones con defecto en el tubo neural), la constante de disociación del ácido pK_a y la energía E_{LUMO} a través de la ecuación:

$$\log \frac{1}{BC_m} = 1.406 pK_a - 42.772 E_{LUMO} + 5.774 \quad (1.1)$$

$$N = 10; R^2 \text{ ajustada} = 0.922; s = 0.38$$

En este caso se puede observar que el log P no juega un papel importante en la toxicidad embrionaria de los ácidos haloacéticos y se observa que el mecanismo que puede estar implicado estriba en cambios del pH intercelular y del metabolismo red-ox. Este mismo autor¹⁰⁵ utilizó posteriormente los ácidos haloacéticos para estudiar las diferencias entre los distintos programas TOPKAT, CASE, MULTI-CASE, DEREK y Oncologic¹⁰⁶, obteniendo los resultados de la Tabla 1.2.

Asumiendo que ningún modelo individual reproduce satisfactoriamente los seis resultados experimentales para los HAA, cabe destacar que solamente TOPKAT es el más preciso en predecir la carcinogenicidad para los di- y trihaloacéticos.

Un comunicado de Venkatapathy *et al.*¹⁰⁷ utilizan las predicciones de mutagenicidad y toxicidad realizadas con el programa TOPKAT, obteniendo los siguientes modelos:

Para los ácidos monohaloacéticos:

$$E_X = 2.81(\pm 0.110) [EFD - \text{halogen}] + 7.84(\pm 0.0986) \quad (1.2)$$

$$N = 4; r^2 = 0.997; Q^2 = 0.928; F = 614.8$$

Para los dihaloacéticos:

Tabla 1.2: Sumario de los datos experimentales y predichos para la carcinogenesis en roedores (ratones machos) para los ácidos haloacéticos.

HAA ⁵	Experimental	CASE ¹	TOPKAT	TOPKAT MM	DEREK (combinado)
CA	-	-	na	na	+
BA	NT ²	marginal	na ³	na	+
DCA	+	-	+	+	-
DBA	+	marginal	+	+	-
BCA	+	+++	NP ⁴ +	+	-
TCA	+	-	+	+	+
BDCA	+	-	na	na	+
DBCA	NT	-	-	+	+
TBA	NT	-	+	+	+

¹Modelo combinado de CASE y MULTICASE en ratas y ratones. ²No Testeado.

³EL modelo resultante no es válido para utilizarlo en este estudio.

⁴No hay predicción posible debido a que los parámetros moleculares están fuera del espacio óptimo de predicción (OPS).

⁵ BDCA = ácido bromodicloroacético, DBCA = dibromocloroacético y BCA = ácido bromocloroacético.

$$\sum EH = -4.48(\pm 0.482) [nHBa] + 9.83(\pm 0.458) \quad (1.3)$$

$$N = 10; r^2 = 0.983; Q^2 = 0.983; F = 458.6$$

Para los trihaloacéticos:

$$E_{C_\alpha} = -0.360(\pm 0.240) \rho LUMO - 1.91(\pm 0.241) N_F + 1.25(\pm 0.209) q_{max}^{Cl} - 5.70(\pm 0.180) \quad (1.4)$$

$$N = 20; r^2 = 0.959; Q^2 = 0.961; F = 125.6$$

Donde E_X es el efecto de cada halógeno en la mutagenicidad, $E_{FD} - halogen$ es la densidad frontera electrofílica del halógeno (nos indica la susceptibilidad de una sustancia de ser atacada por un electrófilo), $\sum EH$ es la suma de los efectos de los halógenos, $nHBa$ es el número de aceptores fuertes de Hidrógeno (el enlace de hidrógeno ocurre entre un dador de hidrógenos y un heteroátomo fuertemente electronegativo como el oxígeno o nitrógeno que es llamado aceptor de Hidrógenos), E_{C_α} es el efecto del

carbono α , ρ LUMO es la densidad del orbital LUMO, N_F número de átomos de fluor y q_{max}^{Cl} la carga parcial máxima en un átomo de cloro.

Densidades en el carbono α , Halógenos y la máxima carga parcial en los átomos de cloro generalmente promocionan reacciones nucleofílicas con ataque en el carbono α al Halógeno¹⁰⁸.

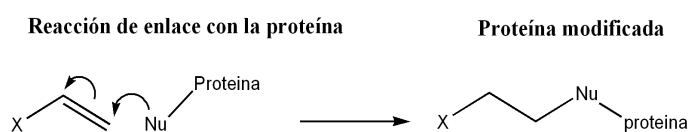
Según los autores y basándose en los resultados obtenidos se puede incisar que la mutagenicidad depende de la habilidad de los ácidos haloacéticos de experimentar reacciones electrofílicas con fragmentos específicos o formar enlaces con dadores de Hidrógeno del DNA, sugiriendo que la habilidad para donar electrones es esencial para que estas sustancias puedan llegar a ser mutagénicas.

Como hemos visto, existe poca información acerca de la potencia mutagénica o carcinogénica de estas sustancias. El uso de métodos que sean capaces de predecir esos valores es importante para la evaluación de los riesgos toxicológicos. Además, como comentamos en el apartado anterior y en el marco del Reglamento REACH, se alienta el uso de QSAR^{109,110} y otros métodos alternativos con el fin de reducir costes y el número de ensayos con animales.

1.3.2. CARBONILOS α , β -INSATURADOS

Otro grupo de sustancias importantes y presentes en el medio ambiente son los carbonilos α , β -insaturados. Estas sustancias suelen usarse en la síntesis de otros compuestos, disolventes, aditivos alimentarios, desinfectantes y en materiales dentales de restauración¹¹¹⁻¹¹³. Una matriz hecha de resina a base de monómeros acrílicos o metacrílicos son foto y/o químicamente polimerizables y se utilizan generalmente como materiales de relleno dentales y adhesivos. Estos materiales de restauración dental, se preparan *in situ* y, como la polimerización a menudo no es ideal, algunos monómeros que no han reaccionado pueden ir filtrándose hacia el exterior en contacto con tejidos blandos a lo largo del tiempo^{114,115}. Entre estas sustancias está el trietilenglicol dimetacrilato (TEGDMA) que causa delecciones en el DNA de células de mamífero^{87,88}.

El doble enlace que caracteriza a los carbonilos α , β -insaturados los hace más reactivos que el carbonilo *per se*, incrementando su habilidad para interactuar con las macromoléculas biológicas ricas en electrones. Este hecho da como resultado una serie de efectos adversos, siendo uno de ellos la mutagenicidad. En general, estas sustancias actúan mediante el mecanismo de adición de Michael (Figura 1.6) por el cual los sustituyentes en el carbón alfa o beta de la insaturación vecinal del carbonilo afectan significativamente en la efectividad de la reacción¹¹⁶.



Dobles o triples enlaces con sustituyentes retiradores de electrones X, como $-\text{CHO}$, $-\text{COR}$, $-\text{CO}_2\text{R}$, $-\text{CN}$, $-\text{SO}_2\text{R}$, $-\text{NO}_2$, etc. incluidos orto- o para-quininas, formadas por la oxidación de orto- o para-dihidroxi aromáticos actúan como aceptores pro-Michael. X puede ser también un grupo heterocíclico como las orto- o para-piridinas.

Figura 1.6: Mecanismo de adición de Michael extraído de Aptula *et al.*¹¹⁶

Existen varios estudios que modulan la mutagenicidad de los carbonilos α , β -insaturados, los cuales se pueden subdividir en dos grupos: aquéllos que modulan la potencia mutagénica^{117–120} y los que modulan la actividad^{121, 122}.

Los modelos que predicen la potencia no suelen ser útiles para predecir la actividad mutagénica y además algunos estudios muestran que los efectos estructurales de la potencia mutagénica deberían distinguirse de los efectos derivados de la actividad⁵⁶.

Los intentos para predecir la actividad mutagénica de esta familia de sustancias están basados solamente en el test de Ames en *Salmonella typhimurium* cepa TA100^{121, 122}. Benigni *et al.*¹²¹ desarrollaron un modelo predictivo para 20 aldehídos α , β -insaturados mediante un análisis discriminante lineal, los resultados de este estudio (Tabla 1.3) indican una dependencia entre la mutagenicidad y parámetros globales como la hidrofobicidad ($\log P$), la voluminosidad a través del parámetro de refractividad molar (MR) y la electrofilia a través del descriptor cuántico LUMO (Lowest Unoccupied Molecular

Orbital).

Tabla 1.3: Análisis discriminante lineal para 20 aldehídos α , β -insaturados realizado por Benigni *et al.*¹²¹.

Variable	Actividad	
	-	+
Constante	-47.13331	20.52153
MR	38.24641	25.41469
log P	-31.77763	-21.45102
LUMO	30.46799	19.77513

Recientemente ha aparecido un estudio para 45 carbonilos α , β -insaturados¹²². En este estudio, el grupo de compuestos fué dividido en varios subgrupos para predecir la mutagenicidad en *Salmonella typhimurium* cepa TA100, (1) derivados halogenados, (2) nitro-derivados de cinamaldehído y (3) acroleínas. Así mismo se desarrolló un sistema de reglas para su clasificación basado en su reactividad y mecanismo de acción, empleando valores de corte de parámetros globales como la masa molecular (MW), hidrofobicidad (log P) y la presencia o no de determinados sustituyentes como grupos alquilo en α o en β .

Estos QSAR solo consideran el test de Ames como punto final exclusivo para la estimación de la mutagenicidad, pero es sabido que es necesario el uso de información a través de múltiples puntos finales para obtener predicciones más reales¹²³. Además solamente emplean descriptores globales (p. ej. el logaritmo del coeficiente de partición octanol/agua, etc) y muchos de ellos son incapaces de observar la influencia de cada fragmento o grupo funcional dentro de la estructura de interés ya que solo pueden dar una comprensión global y no una explicación subestructural.

1.4. Complejación con la β -ciclodextrina

Las ciclodextrinas (CDs) son oligómeros cíclicos de β -D glucosa producidos a partir de almidón por medio de conversiones enzimáticas, con formas parecidas a conos truncados con los hidroxilos primarios y secundarios coronando los bordes más estrechos y más anchos, respectivamente¹²⁴ (Figura 1.7). Según el número de unidades de glucosa que forman la ciclodextrina, ésta se nombra con una letra griega diferente:

- α -CD: 6 moléculas de glucosa (Diámetro/Altura de la cavidad: 4,7..5,3/7,9A).
- β -CD: 7 moléculas de glucosa (Diámetro/Altura de la cavidad: 6,0..6,5/7,9A).
- γ -CD: 8 moléculas de glucosa (Diámetro/Altura de la cavidad: 7,5..8,3/7,9A).
- δ -CD: 9 moléculas de glucosa.

En la bibliografía se describen ciclodextrinas hasta con 17 unidades de glucosa pero no tienen importancia económica ya que los homólogos superiores son difíciles de separar y sus propiedades como huésped de moléculas orgánicas son malas. Hoy en día en Europa la más empleada es la β -Ciclodextrina (β -CD)^{125, 126} ya que fueron las primeras en aprobarse como aditivo alimentario y excipiente, por lo tanto hemos elegido esta familia para realizar nuestro estudio.

Estas sustancias tienen la característica de ser anfipáticas, el interior del toroide es más hidrofóbico y, por tanto, capaces de acoger otras moléculas hidrófobas. En contraste, el exterior es suficientemente hidrofílico para hacer a las ciclodextrinas (o sus complejos) solubles en agua. Estudios anteriores han sugerido cinco principales tipos de interacción ciclodextrina-huésped¹²⁷⁻¹³⁶: (i) las interacciones hidrofóbicas, (ii) las interacciones de van der Waals, (iii) puentes de hidrógeno entre grupos polares del huésped y grupos hidroxilo de la CD, (iv) la relajación por la liberación de agua de alta energía desde la cavidad de la CD hasta la inclusión del sustrato, y (v) el alivio de la tensión conformacional en un aducto CD-agua. La formación de complejos con la CD usualmente es el resultado de las diferentes combinaciones de estas fuerzas. Las CDs son de

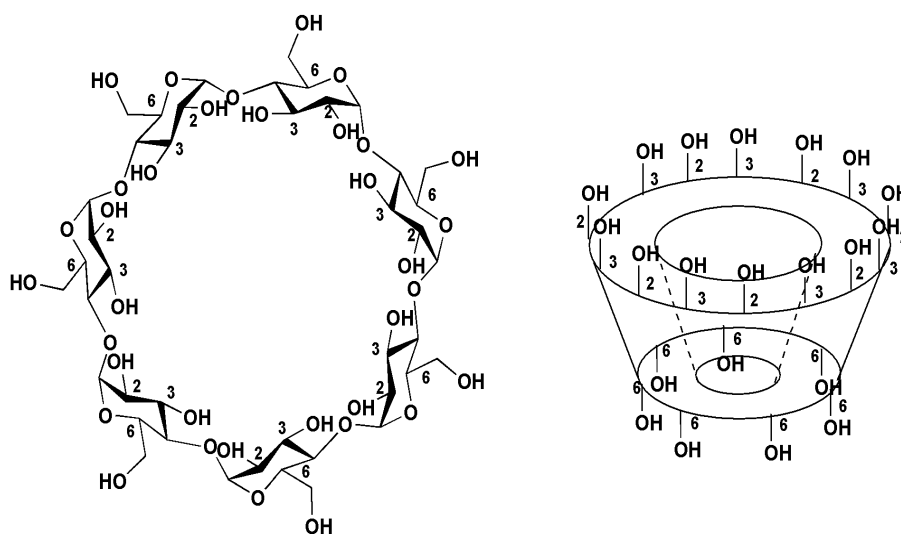


Figura 1.7: Estructura química de la β -CD.

un gran interés en muchos campos, porque son capaces de formar complejos huésped-anfitrión con moléculas hidrófobas y de modificar notablemente sus propiedades físicas y químicas, principalmente en términos de la solubilidad en agua. Por ejemplo, la complejación con CDs aumenta fuertemente la solubilidad de las drogas, haciendo que estén disponibles para una amplia gama de aplicaciones farmacéuticas.

1.5. Importancia de la complejación de sustancias químicas con β -ciclodextrina

La pobre solubilidad de los fármacos continúa siendo de gran importancia para el desarrollo de un gran número de posibles candidatos a drogas¹³⁷. Por esta razón, diferentes fármacos se comercializan actualmente como sólidos o como una solución a base de complejos con CDs^{125, 138, 139}. En estos productos farmacéuticos, las CDs se utilizan principalmente como agentes complejantes para aumentar la solubilidad acuosa de fármacos insolubles, para aumentar su biodisponibilidad y estabilidad^{140–142}. Estos factores han tenido un impacto significativo en lo que se requiere a los formuladores,

dado que el número de opciones de formulación y, por extensión, excipientes, se ha de incrementar para atender el mayor número de retos que se presentan¹⁴³. Las CDs representan un valor añadido en este contexto.

Además, las CDs también puede promover la absorción dérmica, nasal o intestinal de los fármacos, mediante la extracción de colesterol, fosfolípidos o proteínas de membranas¹⁴⁴. A su vez, pueden reducir o prevenir la irritación ocular y gastrointestinal, reducir o eliminar olores o sabores desagradables^{145, 146}, prevenir las interacciones droga-droga o droga-aditivo, así como, convertir sustancias oleosas y líquidas en polvos microcristalinos o amorfos¹⁴⁷. Además, los complejos CD-fármaco a menudo aumentan la biodisponibilidad de la sustancia activa y permiten su liberación de una manera controlada¹⁴⁸. Un ejemplo de esto último podemos verlo en la encapsulación de transplatinum con CDs donde se ha constatado que la citotoxicidad *in vitro* del complejo de inclusión indica que posee una mayor actividad¹⁴⁹.

La β -CD también sirve como soporte y estabilizador de aromas, colorantes y algunas vitaminas en alimentos. Además, la ingesta de β -CD resultante de los aditivos alimentarios se ha estimado en 1-1,4 g/día^{150, 151}.

Las CDs al entrar en contacto con otras sustancias presentes en el medio (como los subproductos de desinfección del agua potable, monómeros dentales, etc..) podrían de alguna manera interactuar con éstas, derivándose de esta interacción una serie de efectos para la salud. La interacción podría darse de dos maneras: puede ser que el tóxico posea una constante de complejación con la CD superior al fármaco o sustancia que complejaba en un principio, y en cierta medida, pueda desplazar al huésped original; o que haya CD libre en la formulación que pueda complejar al tóxico. Los efectos sobre la salud de esta interacción han generado controversia. En un principio se especuló con que esta interacción podría seguir la misma línea que con los fármacos y aumentar, también, la biodisponibilidad de las sustancias tóxicas (que generalmente tienen un carácter hidrofóbico). Esta hipótesis la plantearon Horský y Pitha¹⁵² acompañándola por los resultados de experimentos *in vitro*, que establecieron, por un lado, que las ciclodextrinas

(a saber, (2-hidroxi propil)- β -ciclodextrina (HPBCD)) aumentan la solubilidad en agua de benzo[α]pireno (BaP) y de aflatoxina B1 y por otro lado, que la HPBCD facilita la disolución y el transporte a través del agua de los compuestos lipofílicos poco solubles, refiriéndose a la solubilización de BaP por HPBCD en presencia de sales de ácidos biliares. Estos resultados se utilizaron a este respecto en el trabajo para apoyar dicha hipótesis¹⁵².

Posteriormente los autores, Westerberg y Wiklund¹⁵³, diseñaron un estudio *in vivo* para poner a prueba la hipótesis de que las CDs, al ser buenos solubilizadores, pueden aumentar la biodisponibilidad oral de los agentes carcinógenos ambientales lipofílicos después de ser consumidas. Estos autores investigaron el efecto de la (β -CD) sobre la absorción oral y/o la cinética de eliminación de benzo[α]pireno (BaP) en la rata utilizando BaP radiactivo. Observaron que la biodisponibilidad oral de éste, en dosis que simulan los niveles esperados de exposición sistémica en los seres humanos, es, de hecho, reducida por la administración continua de β -CD en la alimentación. Los resultados son excepcionales, ya que se obtuvieron con una dosis de lipófilo superior a su solubilidad en agua. El efecto negativo de la β -CD en la absorción intestinal de soluciones insaturadas ha sido observado con anterioridad¹⁵⁴. De hecho se sabe que la complejación con CD disminuye la concentración del compuesto libre que pueda ser absorbido y disminuye la fuerza impulsora de la permeabilidad de la pared del intestino. Apoyando esta hipótesis Zheng *et al.*¹⁵⁵ demostró con ensayos *in vitro* que la presencia de β -CD no perturba la integridad del epitelio, ni aumenta la permeabilidad de éste.

La diferencia entre la HPBCD, usada en el estudio de Horský y Pitha¹⁵², y la β -CD, usada en el estudio de Westerberg y Wiklund¹⁵³, se encuentra en la solubilidad. La β -CD posee la menor solubilidad en agua entre las ciclodextrinas naturales y sus derivados comunes, y, los huéspedes lipofílicos son capaces de inducir la precipitación, incluso en soluciones acuosas diluidas de β -CD. Por lo tanto, los datos de Horský y Pitha¹⁵² hacen menos probable que las observaciones de Westerberg y Wiklund¹⁵³ puedan aplicarse sólo a β -CD. En cuanto al hecho de que la biodisponibilidad fue disminuida por

la co-administración de β -CD, Westerberg y Wiklund¹⁵³ proponen que la elevada constante de complejación entre BaP y la β -CD aumenta la solubilidad aparente de BaP, pero al mismo tiempo ralentiza la liberación de BaP del complejo por lo que se elimina con mayor facilidad.

Por lo tanto, ya sea que las ciclodextrinas aumenten la biodisponibilidad oral de las sustancias tóxicas o, por el contrario ralenticen la liberación del complejo facilitando la eliminación, hacen falta estudios que determinen estas constantes de complejación de la ciclodextrina con los tóxicos más comunes presentes en el medio, como pueden ser los ácidos haloacéticos (subproductos de la desinfección del agua potable) y los carbonilos α , β -insaturados (monómeros dentales), además de estudios de biodisponibilidad *in vivo* para poder determinar el alcance de esta interacción.

1.6. Estudios teóricos para la determinación de la complejación con β -ciclodextrina

La determinación experimental de la constante de complejación con las CDs es a menudo difícil y consume mucho tiempo debido a la baja solubilidad de las moléculas huéspedes en solución acuosa. Recientemente se han usado métodos computacionales para predecir las constantes de complejación y para estudiar las fuerzas que intervienen en el proceso. Un exhaustivo conjunto de estas aplicaciones computacionales ha sido excelentemente revisado por Lipkowitz¹⁵⁶.

Para aclarar los factores que más influyen en las interacciones huésped-anfitrión y para predecir la estabilidad termodinámica de los complejos de inclusión con CDs se han aplicado modelos de contribuciones por grupos, métodos QSAR/QSPR (2D-QSAR, 3D-QSAR, CoMFA), cálculos de modelización molecular (utilizando la mecánica cuántica, Monte Carlo/simulaciones de dinámica molecular, mecánica molecular, etc), herramientas de análisis estadísticos y redes neuronales artificiales^{127-132, 134-136, 157}. Sin embargo, es evidente que el conocimiento de las capacidades de complejación de las moléculas huésped se considera necesario para decidir si una complejación huésped-anfitrión es útil en una aplicación en particular usando el conocimiento de qué tipo de

enlaces contribuyen positivamente a este fenómeno.

En este sentido, Katritzky *et al.*¹³⁶ presentó un estudio QSAR para predecir las energías libres de complejos de inclusión entre diversas moléculas huésped y las CDs empleando (i) descriptores del CODESSA y (ii) contaje de los diferentes fragmentos moleculares. El primero de ellos (enfoque tipo Hansch¹⁵⁸) utiliza como parámetros fisicoquímicos determinados descriptores calculados por métodos de mecánica cuántica o por algunas de las técnicas empíricas. El segundo (el enfoque tipo Free-Wilson¹⁵⁹), utiliza el contaje de diferentes fragmentos moleculares como variables en un análisis de regresión múltiple. Ambas técnicas tienen sus ventajas y desventajas¹³⁶. En general, los fragmentos como descriptores (enfoque tipo Free-Wilson) son más interpretables que los descriptores del CODESSA (enfoque tipo Hansch). Sin embargo, la principal desventaja de los métodos de QSPR basados en contaje de diferentes fragmentos moleculares se relaciona con el hecho de que, por lo general, se usan más variables que descriptores del CODESSA, lo que conduce a menores valores de criterio de Fisher (modelos menos robustos).

Otro problema del enfoque basado en fragmentos está relacionado con las moléculas que contienen fragmentos *raros* (es decir, que se encuentran en una sola molécula), los cuáles deberían ser excluidos del grupo de entrenamiento, lo que reduce el número de compuestos tratados¹³⁶.

El último problema se plantea cuando se intenta estudiar los conjuntos de datos heterogéneos de moléculas orgánicas. En este caso no hay necesariamente un patrón atómico/enlace que se repite en todas las moléculas en estudio y, como consecuencia, es más adecuado el empleo de descriptores moleculares tales como, el potencial químico electrónico, la electronegatividad molecular, la dureza química u otros índices globales moleculares.

Este hecho plantea la cuestión de si es posible obtener información estructural a escala local de los modelos desarrollados usando descriptores moleculares globales. Como hemos comentado con anterioridad con los descriptores TOPS-MODE, la única información que necesitamos para transformar el modelo global en contribuciones

atómicas/de enlace es la relación matemática entre el descriptor molecular global y la actividad/propiedad¹⁶⁰.

PLANTEAMIENTO Y OBJETIVOS

2.1. Planteamientos

La mutagenicidad es uno de los primeros pasos para el desarrollo del cáncer. Debido al costo en recursos y el tiempo requerido en los ensayos para determinar la mutagenicidad de una sustancia, ha habido un aumento notable en el interés por el uso de técnicas alternativas para acelerar el establecimiento de prioridades y evaluación de riesgos toxicológicos de las sustancias químicas, alentado bajo el marco de la Unión Europea a través del Reglamento REACH. Una de estas herramientas son los modelos QSAR/QSPR, los cuáles han sido incluidos como punto final válido junto a los ensayos *in vitro*. En las últimas décadas se ha multiplicado enormemente el empleo de modelos QSAR como herramientas válidas para la predicción de la actividad biológica y propiedades de las sustancias químicas, por su rapidez y simplicidad.

Los ácidos haloacéticos están presentes en el agua potable como subproductos de desinfección y para alguno de ellos se han obtenido resultados positivos en ensayos de mutagenicidad en *S. typhimurium*. Además, con el paso del tiempo se han ido encontrando en el agua potable nuevas sustancias pertenecientes a esta familia con valores superiores de mutagenicidad. Al mismo tiempo, las predicciones realizadas de mutagenicidad para esta familia de sustancias no han sido de gran alcance y no existe ningún estudio QSAR dedicado a este propósito.

Los carbonilos α , β -insaturados están presentes en el medio ambiente, sobre todo como monómeros empleados para la elaboración *in situ* de materiales dentales de res-

tauración. Estas sustancias poseen la característica de poder interactuar con las macromoléculas biológicas ricas en electrones, dando como resultado una serie de efectos adversos, siendo uno de ellos la mutagenicidad. Los QSAR existentes para estas sustancias o para una subfamilia de estas (aldehídos α , β -insaturados) solo consideran el test de Ames como punto final exclusivo para la estimación de la mutagenicidad empleando para ello descriptores globales. Debido a ello, la mayoría de estos estudios son incapaces de observar la influencia de cada fragmento o grupo funcional dentro de la estructura de interés ya que solo pueden dar una comprensión global y no una explicación subestructural. Además, es necesario el uso de información a través de múltiples puntos finales para obtener predicciones más reales.

Unido a todo esto, la β -CD está presente como ingrediente alimentario y excipiente farmacéutico, pudiendo llegar a interactuar con los tóxicos más comunes presentes en el medio (p. ej. los ác. haloacéticos y los carbonilos α , β -insaturados). Los resultados de esta interacción no están del todo claros ya que es necesario conocer las constantes de complejación de la CD con estas sustancias para empezar a conocer el alcance de ésta. Los modelos QSPR realizados para predecir la complejación están basados por un lado en descriptores globales (que solo pueden dar una comprensión global y no una explicación subestructural) y, por otro lado, en la presencia de fragmentos (elevado número de variables, modelos menos robustos). Este hecho plantea la cuestión de si es posible obtener información estructural a escala local de los modelos desarrollados usando descriptores moleculares globales.

2.2. Objetivos

El objetivo general de esta tesis es desarrollar modelos QSAR/QSPR para predecir la mutagenicidad de los ácidos haloacéticos y los carbonilos α , β -insaturados; así como, la constante de complejación con β -CD para diferentes familias de sustancias. Con este fin se han establecido los siguientes objetivos específicos:

1. Desarrollar modelos QSAR para predecir la potencia mutagénica en *S.typhimurium*

- cepa TA100 para la familia de ác. haloacéticos.
2. Desarrollar modelos QSAR para predecir la mutagenicidad para los carbonilos α , β -insaturados a través de distintos puntos finales.
 3. Desarrollar modelos QSPR para predecir la constante de complejación con la β -CD.
 4. Formular hipótesis acerca de los mecanismos de complejación con la β -CD para distintas familias de sustancias; así como, acerca de los mecanismos de actuación mutagénica para los ácidos haloacéticos y los carbonilos α , β -insaturados.
 5. Determinar una serie de fragmentos que favorezcan o interfieran el fenómeno de complejación con la β -CD y una serie de alertas estructurales (SAs) de mutagenicidad para los ácidos haloacéticos y los carbonilos α , β -insaturados.

RESULTADOS Y DISCUSIÓN

Comenzamos modelando la potencia mutagénica de un total de 42 derivados halogenados (nitrohaloalcanos, haloácidos, haloaldehídos, halocetonas, haloalcohols, haloepóxidos y haloalcanos) muchos de ellos conocidos agentes alquilantes y que podemos encontrar como subproductos de desinfección del agua potable, con el fin de obtener predicciones para los ác. haloacéticos debido a la problemática que presentaban éstos (ver Tabla 3.1).

En la Tabla 3.1 se muestra un resumen de los datos más importantes (familia de sustancias, punto final evaluado, descriptores moleculares, etc.) tenidos en cuenta para el desarrollo de los modelos QSAR/QSPR, así como las referencias bibliográficas donde fueron publicados los mismos y que forman parte de los resultados de esta tesis doctoral.

3.1. Ácidos haloacéticos

El modelo QSAR generado para esta familia de sustancias es el referido en la Tabla 3.1 como QSAR1, cuyos parámetros estadísticos una vez ortogonalizado y eliminado las desviaciones se muestra en la Tabla 3.2, junto con los resultados con otras familias de descriptores del Dragon, los cuáles fueron presentados en el primer artículo¹⁶¹.

El modelo QSAR derivado del uso de los momentos espectrales (QSAR1) poseen una mayor bondad de ajuste por presentar unos mayores valores del coeficiente de determinación (R^2), el parámetro de Fisher (F) y la función de Kubinyi (FIT), además

Tabla 3.1: Listados de los modelos QSAR desarrollados

Código	Familia sustancias	Punto final ¹	Tipo	Descriptorios moleculares	Mut. ²	No-mut. ²	N ³	Análisis estadístico ⁴	Ref.
QSAR1	Derivados halogenados	TA100	Potencia	TOPS-MODE	-	-	40	RLM	161
QSAR2	Carbonilos α, β -insaturados	AMES	Actividad	TOPS-MODE	103	116	219	ALD	162
QSAR3	Carbonilos α, β -insaturados	MCM	Actividad	TOPS-MODE	34	14	48	ALD	162
QSAR4	Carbonilos α, β -insaturados	AMES	Actividad	DRAGON	104	116	220	ALD	163
QSAR5	Carbonilos α, β -insaturados	MCM	Actividad	DRAGON	34	14	48	ALD	163
QSAR6	Carbonilos α, β -insaturados	AMES	Actividad	DRAGON	104	116	220	Consenso	163
QSPR1	no-cogénica	log K	Potencia	DRAGON	-	-	232	RLM	164
QSPR2	no-cogénica	log K	Potencia	TOPS-MODE	-	-	232	RLM	165

¹ TA100= ensayo de Ames en cepa TA100 sin activación; AMES= ensayo de Ames; MCM= ensayo de mutagenicidad en células de mamífero y log K= logaritmo de la constante de complejación.

² Mut= n° sustancias mutagénicas; No-mut= n° sustancias no mutagénicas.

³ N= n° sustancias totales.

⁴ RLM= Regresión lineal multivariada; ALD= Análisis discriminante lineal.

de por unos menores valores del criterio de información de Akaike (AIC) y de la desviación estandar (s). Al mismo tiempo este modelo presenta el mejor poder predictivo, dado por los altos valores en los coeficientes de determinación de la validación cruzada: Q^2 y Q_{boot}^2 .

En el modelo QSAR1 podemos observar (Tabla 3.3), que la contribución de las variables ponderadas con los momentos dipolares explican un 66.9% de la varianza. La ausencia de la hidrofobicidad nos confirma la característica propia de las sustancias mutagénicas de acción directa¹⁶⁶. Estos momentos dipolares son debidos a diferencias de electronegatividades entre los átomos que constituyen el enlace, luego, densidades de carga en el carbono α al halógeno dan lugar a reacciones nucleofílicas con ataque en ese carbono, corroborándose las conclusiones obtenidas por Venkatapathy *et al.*¹⁰⁷ en su modelo.

Las contribuciones de enlace para la mutagenicidad obtenidas del modelo desarro-

Tabla 3.2: Parámetros estadísticos de los mejores modelos QSAR obtenidos por regresión lineal para las distintas familias de descriptores.

Descriptores	Variables	R^2	F	s	Q^2	AIC	FIT	Q^2_{boot}	$a(R^2)$	$a(Q^2)$
(QSAR1) TOPS-MODE	$\mu_{15}^{Std}, \mu_1^{Pols}, \mu_7^{Mol},$ $\mu_0 \mu_{13}^{Dip}, \mu_0 \mu_5^{Pols}, \mu_1 \mu_2^{Dip2},$ $\mu_1 \mu_{14}^{Dip2}, \mu_1 \mu_9^{Pol}$	0.90	35.73	0.54	0.84	0.72	1.52	0.71	0.12	-0.39
RDF	RDF055v, RDF010u, RDF030v, RDF055p, RDF025u, RDF025p, RDF020e, RDF035u	0.81	16.79	0.75	0.68	0.83	1.28	0.17	0.15	-0.49
Geométricos	G(O..CI), FDI, HOMT, G(O..I), SPH, SPAN, G2, ASP	0.83	18.73	0.72	0.66	0.76	1.43	0.44	0.16	-0.48
Índices basados en autovalores	SEigZ, SEige, AEigp, VRA1, LP1, SEigv, AEigm, Eig1e	0.78	14.21	0.81	0.64	0.96	1.09	0.50	0.14	-0.55
WHIM	L2s, L3u, E3m, G2e, G3u, E3e, Dp, G3s	0.77	13.64	0.82	0.57	0.99	1.04	0.29	0.15	-0.52
Información	IC1, IDDE, IC2, CIC4, HVcpx, Yindex, Uindex, TIC5	0.81	16.61	0.76	0.56	0.84	1.27	0.43	0.16	-0.55
Autocorrelaciones 2D	GATS5v, ATS6e, MATS5e, ATS1m, GATS2e, ATS2m, GATS4m, MATS5p	0.79	15.33	0.78	0.39	0.90	1.17	0.22	0.15	-0.64

Tabla 3.3: Contribución de los momentos espectrales al modelo

Variables	R^2 global	R^2 para cada variable (paso a paso)
$\Omega \mu_1 \mu_2^{Dip2}$	0.421	0.421
$\Omega^2 \mu_0 \mu_{13}^{Dip}$	0.595	0.174
$\Omega^7 \mu_1 \mu_9^{Pol}$	0.675	0.08
$\Omega^6 \mu_1^{Pols}$	0.753	0.078
$\Omega^8 \mu_1 \mu_{14}^{Dip2}$	0.827	0.074
$\Omega^4 \mu_7^{Mol}$	0.86	0.033
$\Omega^5 \mu_0 \mu_5^{Pols}$	0.888	0.028
$\Omega^3 \mu_{15}^{Std}$	0.902	0.014

llado vienen expresadas en la Figura 3.1.





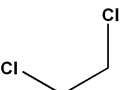
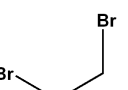
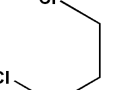
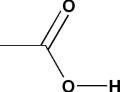






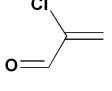
				Fragmento	Contribución
				F ₁	0.14
				F ₂	0.45
				F ₃	0.69
				F ₄	2.20
				F ₅	0.90
				F ₆	1.60
				F ₇	2.15
				F ₈	-0.63
				F ₉	0.67
				F ₁₀	0.50
				F ₁₁	-1.09
				F ₁₂	1.07
				F ₁₃	0.95
				F ₁₄	1.47
				F ₁₅	2.20

Figura 3.1: Estructuras y contribuciones de los fragmentos seleccionados a la mutagenicidad.

Observando los fragmentos F₁, F₂, F₃ y F₄ vemos que el orden de mutagenicidad es I>Br>Cl>F en concordancia con el orden de reactividad de los derivados halogenados y su potencial como grupo saliente. Los dihalogenados, si son vecinales, aumentan la mutagenicidad puesto que pueden actuar, bien por sí mismos o a través de la conjugación con glutathion (GSH) formando el ion episulfonio (potente electrófilo), como agentes de entrecruzamiento (*cross-linking*) entre dos centros nucleófilos del DNA^{167, 168} (ver mecanismo en la Figura 3.2); ésto lo podemos observar a través de los valores de los fragmentos F₁₃, F₁₄, F₆ y F₇.

De acuerdo a los valores positivos de las contribuciones de los fragmentos F₉, F₁₀ y

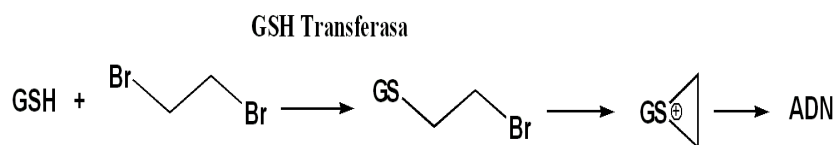


Figura 3.2: Mecanismo de activación de los dihalogenados a través de la Glutation-S-Transferasa.

F12 podemos decir que subestructuras con grupos epóxidos, aldehídos y carbonilos α , β -insaturados aumentan la mutagenicidad. Esto está en consonancia con los hallazgos encontrados en la literatura^{169–171} donde se informa que los epóxidos y los aldehídos son potentes agentes alquilantes, especialmente los de tamaño pequeño.

Los carbonilos α , β insaturados, como veremos en la siguiente sección, tienen un mecanismo de acción del tipo adición de Michael¹⁷² (Figura 3.3). Además, la sustitución de un un halógeno en el carbono β aumenta la mutagenicidad¹⁷³ como podemos apreciar por los fragmentos F₁₂ y F₁₅, debido al potencial de entrecruzamiento con otros centros nucleófilos del DNA o proteínas¹⁷⁴.

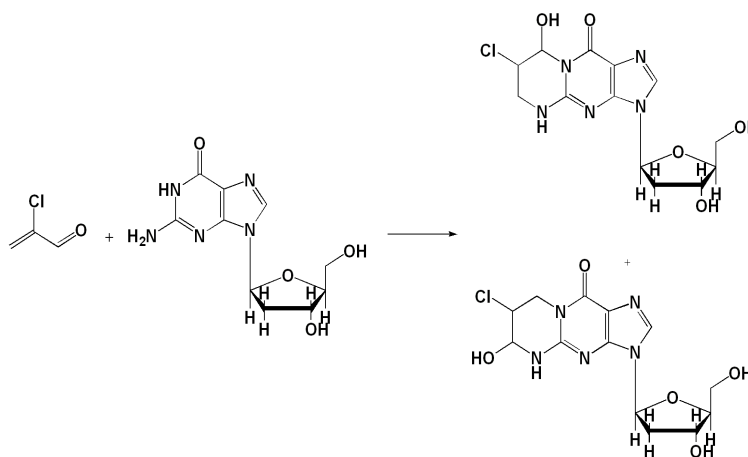


Figura 3.3: Mecanismo de adición tipo Michael para los grupos carbonilo α , β insaturados con cloro en posición 2.

Tabla 3.4: Mejores modelos obtenidos para el ensayo de mutagenicidad de Ames.

Familia	N ^o variables	Entrenamiento (%)			λ	Validación (%)			FIT (λ)	Kappa ² (K)
		Sens.	Espec.	Precis.		Sens. ²	Espec. ²	Precis. ²		
TOPS-MODE ¹ (QSAR2)	7	87	91	89	0.45	85	86	85	0.87	0.70
Conteo de grupos funcionales (QSAR4)	8	81	91	86	0.49	86	91	89	0.71	0.77
Fragmentos centrados en átomos	8	80	90	85	0.50	74	85	79	0.69	0.58
3DMORSE	8	81	84	82	0.53	90	81	85	0.61	0.70
GETAWAY	7	75	88	82	0.54	75	87	81	0.62	0.62
Autocorrelaciones 2D	7	78	88	84	0.58	76	83	80	0.53	0.58
Geométricos	6	80	80	80	0.58	75	82	79	0.57	0.57
RDF	8	81	78	80	0.62	86	80	83	0.42	0.65
WHIM	7	71	86	79	0.63	81	75	78	0.44	0.56
Constitucionales	5	72	87	80	0.64	78	67	72	0.47	0.44
Autovalores de Burden	7	71	81	76	0.65	80	74	77	0.40	0.53

¹ Obtenido con 175 compuestos eliminando la desviación correspondiente a la sustancia: cinamato de cinamilo.

² Resultados obtenidos teniendo en cuenta sólo las sustancias que están dentro del dominio de aplicación.

Grupos como los ácidos carboxílicos y alcoholes (F₈ y F₁₁) disminuyen la mutagenicidad. Tanto para ácidos como para alcoholes alifáticos, estos últimos de bajo número de átomos de carbono, se han obtenido resultados negativos¹⁷⁵⁻¹⁷⁹.

Es importante destacar que el modelo QSAR1 está incluido por la Unión Europea en su base de datos para la evaluación regulatoria de productos químicos con arreglo al Reglamento REACH (QMRF Q8-26-8-155).

3.2. Carbonilos α, β -insaturados

La mutagenicidad de los carbonilos α, β -insaturados se modeló con QSARs basados en la aproximación TOPS-MODE y en los descriptores del programa Dragon. En las Tablas 3.4 y 3.5 se recopilan los mejores modelos obtenidos, junto a sus parámetros estadísticos para los dos puntos finales estudiados (ensayo de Ames y mutagenicidad en células de mamífero).

Tabla 3.5: Mejores modelos obtenidos para el ensayo de mutagenicidad en células de mamífero.

Familia	N ^o variables	Entrenamiento (%)			λ	Validación (%)			FIT (λ)	Kappa ² (K)
		Sens.	Espec.	Precis.		Sens. ¹	Espec. ¹	Precis. ¹		
Fragmentos centrados en átomos (QSAR5)	4	89	83	87	0.36	86	100	89	1.10	0.72
Conteo de grupos funcionales	4	100	67	90	0.39	86	100	89	0.93	0.72
Topológicos	4	96	75	90	0.40	86	100	89	0.92	0.72
TOPS-MODE (QSAR3)	4	86	91	89	0.41	86	88	87	0.88	0.72
Conectividad	4	100	75	92	0.42	86	100	89	0.85	0.72
GETAWAY	4	93	83	90	0.42	86	0	67	0.83	0.72
RDF	4	93	83	90	0.43	86	50	78	0.82	-0.17
Autovalores de Burden	4	85	100	90	0.44	83	100	88	0.76	0.35
Autocorrelaciones 2D	4	93	83	90	0.45	71	100	75	0.75	0.71
Constitucionales	4	96	67	87	0.46	86	100	89	0.70	0.38
Contaje de caminos y pasos	4	93	75	87	0.47	71	100	78	0.68	0.72
Basados en autovalores	4	89	75	85	0.49	86	100	89	0.64	0.52
Geométricos	4	93	67	85	0.50	71	0	56	0.60	0.72
WHIM	4	89	83	87	0.52	71	100	75	0.57	0.38
Perfiles moleculares de Randić	4	85	58	77	0.62	86	50	78	0.38	0.35

¹ Resultados obtenidos teniendo en cuenta sólo las sustancias que están dentro del dominio de aplicación.

3.2.1. MUTAGENICIDAD EN EL ENSAYO DE AMES

La Tabla 3.4 muestra que los mejores modelos son QSAR2 y QSAR4, obtenidos con la aproximación TOPS-MODE y con el uso de los descriptores: conteo de grupos funcionales, estos últimos implementados en el programa Dragon. El mejor desempeño de estos modelos viene reflejado por la calidad de sus estadígrafos; específicamente los pequeños valores de la lambda de Wilk (λ), elevados valores de FIT(λ) y de los porcentajes de clasificación, tanto para la serie de entrenamiento como de predicción. Estas familias de descriptores son muy atractivos desde el punto de vista de la modelación QSAR ya que se pueden obtener directamente de las estructuras empleando bajos recursos computacionales. Para una comparación visual fácil, nuestros resultados se expresaron como gráficos de características operativas recibidas (ROC) (Figura 3.4). Un gráfico ROC informa de la razón de verdaderos positivos (sensibilidad) en el eje Y y la razón de falsos positivos (1 - especificidad) en el eje X, dónde un rendimiento perfecto se encuentra en la esquina superior izquierda y la línea diagonal representa resultados clasificatorios

aleatorios¹⁸⁰. Además de los resultados de la Tabla 3.4 hemos incluido los obtenidos por el módulo de carcinogenicidad/mutagenicidad del sistema de expertos TOXTREE para el mismo grupo de sustancias.

Un análisis de los valores de las variables más significativas ($nArNO_2$, $nCconjX$, nCH_2RX , $nRCHO$ y $nArC=N$) del modelo QSAR4 y de las contribuciones de enlace obtenidas del QSAR2 (representadas en la Figura 3.5), muestra que la presencia de conocidas SAs, como son los grupos nitro-aromáticos ($nArNO_2$ y F_9) y aminas aromáticas, específicamente iminas ($nArC=N$), incrementan la mutagenicidad. Esa mutagenicidad se produce a través de rutas enzimáticas de activación metabólica (nitroreductasas para los nitroaromáticos¹⁸¹ y citocromo P-450 para las aminas aromáticas⁵⁹) a especies N-hidroxi que se transforman en iones nitronio, que pueden reaccionar con el DNA formando aductos^{182, 183}.

Otras SAs conocidas son los derivados halogenados alifáticos primarios (nCH_2RX y F_{10}) y los grupos epóxido (F_8) ambos reconocidos agentes alquilantes. Además, la presencia de halógenos en la posición *alpha* o *beta* del doble enlace ($nCconjX$, F_6 , F_7 y F_{17}) aumenta la mutagénesis en el test de Ames^{173, 184}, debido a la formación de enlaces cruzados (crosslink) con otro centro nucleófilo del DNA o proteína¹⁷⁴. A su vez, como era de esperar, los aldehídos ($nRCHO$ y F_2) son más mutagénicos que otros grupos carbonilo, como las cetonas (F_1). Esta misma relación ha sido obtenida recientemente por Koleva *et al.*¹²², quienes afirman que la reactividad del grupo carbonilo en los procesos de adición electrofílica está influenciada por el tamaño y los efectos electrónicos de los sustituyentes. Ambos factores estéricos y electrónicos favorecen a los aldehídos a ser más reactivos; además este grupo posee un mayor efecto electro-aceptor en el doble enlace que el grupo cetónico, incrementando así su reactividad a través del mecanismo de adición de Michael¹¹⁶. De la misma manera la presencia del grupo nitrilo en los acrilatos aumenta la mutagenicidad del grupo (comparando los valores de F_{17} y F_4) debido, posiblemente, a este efecto aceptor de electrones.

En resumen, la información extraída de los mejores modelos individuales (QSAR2

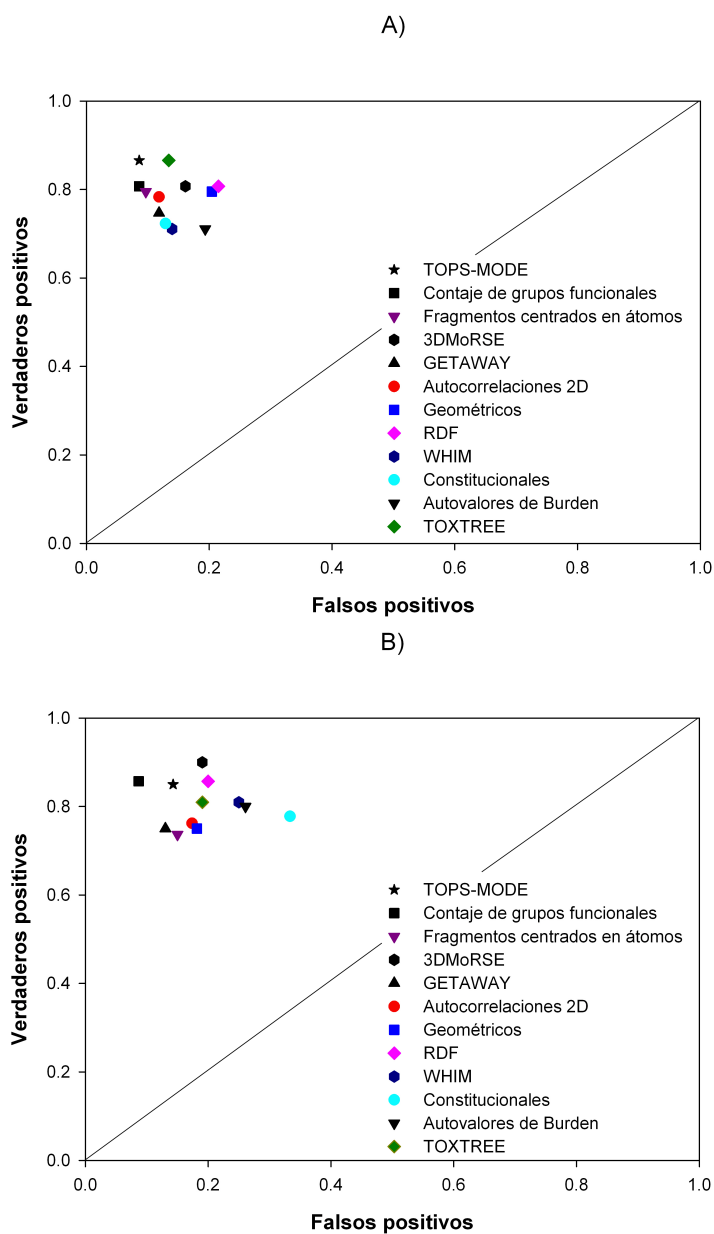


Figura 3.4: Gráfico ROC correspondiente a los resultados obtenidos en la Tabla 3.4 para los modelos QSAR desarrollados para el ensayo de mutagenicidad de Ames junto a las predicciones hechas por el programa TOXTREE. A) grupo de entrenamiento, B) grupo de predicción.

A)

Fragmento	Variable	Contribución
	$nArNO_2$	Positiva
	$nCconjX$	Positiva
	nCH_2RX	Positiva
	$nRCHO$	Positiva
	$nArC=N$	Positiva

B)

Fragmento	Contribución
	F ₁
	F ₂
	F ₃
	F ₄
	F ₅
	F ₆
	F ₇
	F ₈
	F ₉
	F ₁₀
	F ₁₁
	F ₁₂
	F ₁₃
	F ₁₄
	F ₁₅
	F ₁₆
	F ₁₇

Figura 3.5: Fragmentos y contribuciones obtenidas por los modelos. A) QSAR4, B) QSAR2.

y QSAR4) puso de manifiesto que el principal mecanismo de acción para esta familia de sustancias en el ensayo de Ames es del tipo adición de Michael. Como hemos visto, sustituyentes en α o β del doble enlace tienen una fuerte influencia, positiva o negativa, al igual que ocurre con la reactividad de los receptores tipo Michael (Figura 3.5). Aunque los receptores de Michael son electrófilos suaves esto no significa que no reaccionen con nucleófilos duros como el DNA¹¹⁶.

Si bien, la capacidad predictiva de los modelos individuales fue buena, aplicamos el desplazamiento *dotefilide*^{163, 185} en un intento de mejorar la calidad en las predicciones. Mediante este método se combinan los mejores modelos para crear otros más útiles, ya que las sustancias son clasificadas de distinta manera por cada una de las familias de descriptores. Los modelos elegidos y los mejores resultados para las combinaciones anteriores se muestran en la Tabla 3.6 y en la Figura 3.6.

Tabla 3.6: Resultados de los modelos seleccionados para realizar los consensos basados en el ensayo de mutagenicidad de Ames teniendo en cuenta todas las sustancias.

Familia	Sensibilidad	Especificidad	Precisión	Kappa
	%	%	%	
Conteo de grupos funcionales (QSAR4)	86	91	89	0.77
Fragmentos centrados en átomos	67	74	70	0.51
3DMoRSE	86	74	80	0.64
GETAWAY	71	87	80	0.60
Recuperación positiva ^a	95	78	86	0.73
Recuperación negativa ^b	76	100	89	0.77
Modelo consenso general ^c	71	70	70	0.60
Modelo consenso teórico ^d (QSAR6)	94	100	97	0.94

^aModelo combinado entre conteo de grupos funcionales y 3DMoRSE

^bModelo combinado entre conteo de grupos funcionales y GETAWAY

^cModelo combinado entre conteo de grupos funcionales, 3DMoRSE y GETAWAY

^dModelo combinado entre conteo de grupos funcionales, 3DMoRSE y GETAWAY con predicciones para el 76 % de los positivos y el 70 % de los negativos

El modelo de recuperación positiva tuvo una precisión del 86 % y clasificó correc-

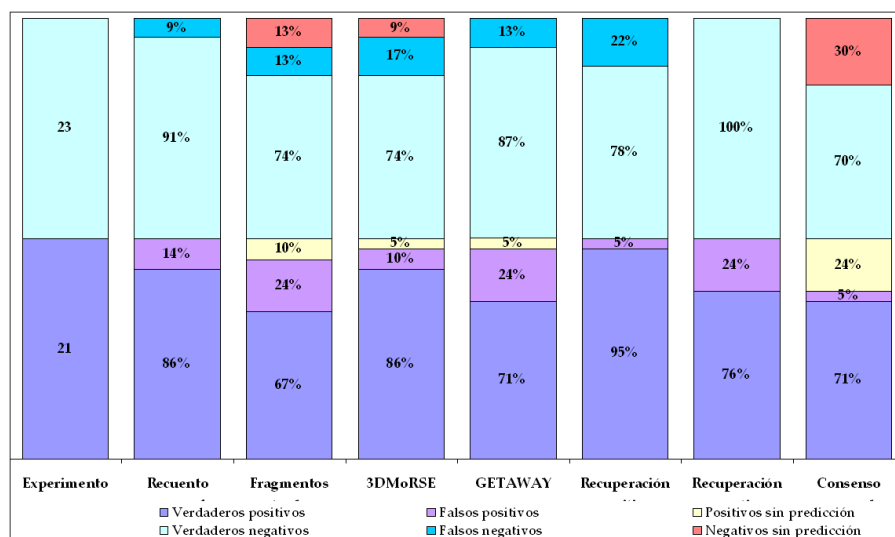


Figura 3.6: Clasificaciones realizadas por los modelos individuales y por los consensos para los carbonilos α , β -insaturados en el ensayo de Ames.

tamente el 95 % de las sustancias mutagénicas. Estos resultados que superaron a todos los modelos individuales, hacen de este modelo una herramienta de máxima utilidad si uno está interesado en la identificación de todas las sustancias positivas posibles o en encontrar no-mutágenos definitivos. El modelo de recuperación negativa clasificó correctamente todas las sustancias no-mutagénicas con una precisión igual que el mejor de los modelos individuales. Este modelo es útil cuando se desea identificar compuestos mutagénicos definitivos. El modelo consenso general es el menos preciso (70 %) de todos los modelos presentados. A pesar de que el número de moléculas no clasificadas (con predicciones contradictorias realizadas por los tres modelos individuales) es elevado (27 %), de entre las predichas (70 %), lo son con precisión el 100 % de los negativos y el 95 % de los positivos (consenso teórico); datos que confieren una gran confiabilidad en las predicciones realizadas con este modelo (Tabla 3.6). Al mismo tiempo, la tasa de falsos positivos (5 %) y falsos negativos (0 %) se redujo con este modelo de consenso respecto a los modelos individuales (Figura 3.6).

Los modelos de recuperación negativa y consenso general mejoraron al incluir el QSAR4 (resultados no publicados), obteniéndose unos modelos (Tabla 3.7) con mayor precisión que los anteriores (91 % y 84 %, respectivamente). Cuando se realiza una predicción con el modelo consenso general, el 86 % de los compuestos positivos y el 95 % de los compuestos negativos son clasificados con precisión (consenso teórico), con una tasa del 14 % de falsos positivos y 5 % de falsos negativos (Figura 3.7).

Tabla 3.7: Resultados de los modelos seleccionados para realizar los consensos basados en el ensayo de mutagenicidad de Ames incluyendo el QSAR4.

Familia	Sensibilidad %	Especificidad %	Precisión %	Kappa
Conteo de grupos funcionales (QSAR2)	86	91	89	0.77
Fragmentos centrados en átomos	67	74	70	0.51
TOPS-MODE(QSAR4)	85	86	85	0.707
3DMoRSE	86	74	80	0.64
GETAWAY	71	87	80	0.60
Recuperación positiva ^a	95	78	86	0.73
Recuperación negativa ^b	86	96	91	0.82
Modelo consenso general ^c	86	83	84	0.72
Modelo consenso teórico ^d (QSAR6)	86	95	90	0.81

^aModelo combinado entre conteo de grupos funcionales y 3DMoRSE

^bModelo combinado entre conteo de grupos funcionales y GETAWAY

^cModelo combinado entre conteo de grupos funcionales y TOPS-MODE

^dModelo combinado entre conteo de grupos funcionales y TOPS-MODE con predicciones para el 100 % de los positivos y el 87 % de los negativos

Se demuestra que tanto los modelos individuales como los consenso son herramientas útiles para la identificación de sustancias mutagénicas según el ensayo de Ames. Además, el modelo QSAR4 está incluido en la base de datos de la Unión Europea para la evaluación regulatoria de productos químicos con arreglo al Reglamento REACH (QMRF Q14-26-8-158).

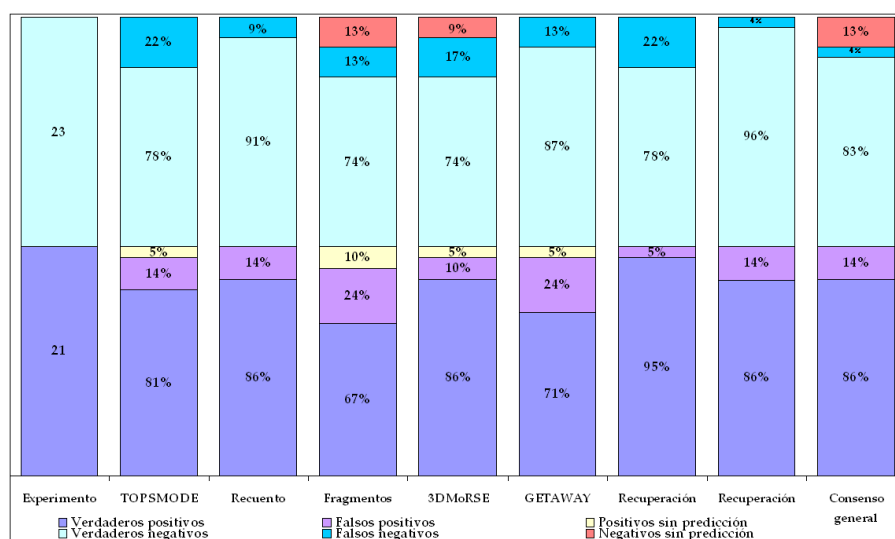


Figura 3.7: Clasificaciones realizadas por los modelos individuales y por los consensos incluyendo el QSAR4 para los carbonilos α , β -insaturados en el ensayo de Ames.

3.2.2. SAS Y COMPARACIÓN CON EL SISTEMA DE EXPERTOS TOXTREE

Como hemos podido comprobar en el apartado anterior, el modelo QSAR2 posee una mayor precisión que el sistema de expertos TOXTREE. Esta mejora se hace más palpable al observar los valores de las contribuciones de enlace obtenidas por el QSAR2 para las SAS detectadas por el software TOXTREE usando las mismas sustancias. Estos valores de contribución están influenciados por el resto de grupos funcionales o subestructuras presentes en la molécula, cualidad que puede ser cuantificada por los descriptores TOPS-MODE, convirtiéndolos en una herramienta más útil que otros para modular las actividades de ciertos grupos funcionales y así obtener unas SAS más precisas. En la Figura 4 de la referencia 162 podemos ver una serie de sustancias correctamente clasificadas por el QSAR2 junto con los valores obtenidos de contribución de enlace por este modelo. Para estas sustancias, el TOXTREE no reconoce alerta estructural alguna, dando lugar a falsos negativos. En esta figura podemos destacar subes-

estructuras presentes en conocidas sustancias mutagénicas (MX y CMCF) como el anillo 2-furanona(5H) y los dobles enlaces clorados, subestructuras que contribuyen positivamente a la mutagenicidad de la sustancia y que, como ya hemos dicho, el TOXTREE no posee implementadas. Además, el modelo QSAR2 es capaz de modular la mutagenicidad de las SAs detectadas por el TOXTREE (Figura 5 referencia 162), como por ejemplo la SA10 correspondiente a los carbonilos α, β -insaturados. En esta figura se puede observar que la mutagenicidad no es debida a la SA10 sino a la presencia en la molécula de halógenos en el doble enlace o grupos epóxido (SA7). En la tabla 8 de la referencia 162, se aprecia que, en líneas generales, las SAs detectadas por el TOXTREE no presentan valores de contribución positiva a la mutagenicidad determinados por el QSAR2.

A la vista de estos resultados, el modelo QSAR2 realizado con los descriptores TOPS-MODE podría ayudar a mejorar aquellos sistemas de expertos (ej. TOXTREE) donde las SAs estén implementadas. Además se vuelve a poner de manifiesto la calidad que poseen estos descriptores como *generadores de conocimiento*.

3.2.3. MUTAGENICIDAD EN CÉLULAS DE MAMÍFERO

Al igual que hicimos en el ensayo de mutagenicidad de Ames, obtuvimos una serie de modelos para el ensayo en células de mamífero (MCM) (Tabla 3.5). Para este ensayo el mejor modelo se obtuvo con los fragmentos centrados en átomos (QSAR5) con una mejora destacable en los parámetros estadísticos con respecto de las otras familias. Al mismo tiempo, el modelo correspondiente a los descriptores TOPS-MODE (QSAR3), a pesar de obtener con él unos parámetros estadísticos de menor calidad, ofrece unas clasificaciones para el grupo de entrenamiento más balanceadas; así mismo posee la característica (comentada en el apartado anterior) de obtener una distribución de la mutagenicidad a escala local, por lo que el número de subestructuras que se pueden extraer es superior. Por lo tanto se emplearon ambos modelos para la extracción de las posibles SAs.

No existen en la literatura unas SAs definidas para este punto final y gracias a estos

modelos se han podido extraer una serie de subestructuras que afectan notablemente a los valores de la mutagenicidad. Un ejemplo de estas SAs obtenidas por los modelos QSAR3 y QSAR5 se observa en la Figura 3.8. Las variables más significativas (H-046, C-015 y C-016) del modelo QSAR5 influyen en la mutagenicidad; así, podemos ver que para el grupo de entrenamiento un aumento del número de hidrógenos unidos a carbono sp^3 sin heteroátomos en el siguiente (H-046) conduce a un aumento del tamaño de la molécula y, por lo tanto, una reducción de la mutagenicidad. Este aumento de tamaño dificulta el paso de la sustancias al interior de la célula y, por lo tanto, su acceso al material genético.

Otras conclusiones que se pueden extraer de estos dos modelos son, por un lado, que la presencia de un doble enlace en la posición terminal de la cadena (C-015) favorece la mutagenicidad, mientras que, por otro lado, sustituyentes alquílicos en el doble enlace la dificultan (C-016, o comparado los fragmentos F_1 , F_2 , F_4 y F_5 con F_7 , F_8 , F_9 y F_{10} , respectivamente). Esto es debido a una reducción de la carga positiva en el carbono terminal del doble enlace, sitio preferido para el ataque nucleofílico^{111,186}, producido a través de un mecanismo de acción tipo adición de Michael con el sulfhidrilo del glutatión (GSH) o bien mediante una reacción enzimática catalizada por la glutatión transferasa^{187,188}. El déficit de GSH hasta niveles inferiores al 20%¹⁸⁹ provocado por los carbonilos α , β -insaturados, favorece la generación de radicales libres de oxígeno (ROS). Estos ROS pueden iniciar peroxidación lipídica y otros procesos, dando lugar a un aumento del daño celular citotóxico/genotóxico¹⁹⁰. Por lo tanto, la presencia de un doble enlace terminal sin substituyentes electrodonadores hace a estos compuestos más mutagénicos para estos ensayos. Al mismo tiempo y, siguiendo el mecanismo de adición de Michael, la presencia de substituyentes electroaceptores en el doble enlace incrementa la mutagenicidad de la molécula^{116,191}, p. ej. fragmento F_7 (Figura 3.8).

En comparación con los resultados del ensayo de Ames, vemos que esta reactividad como receptores de Michael es más pronunciada en el ensayo con células de mamífero, observando los valores más altos en la variación de las contribuciones, posiblemente

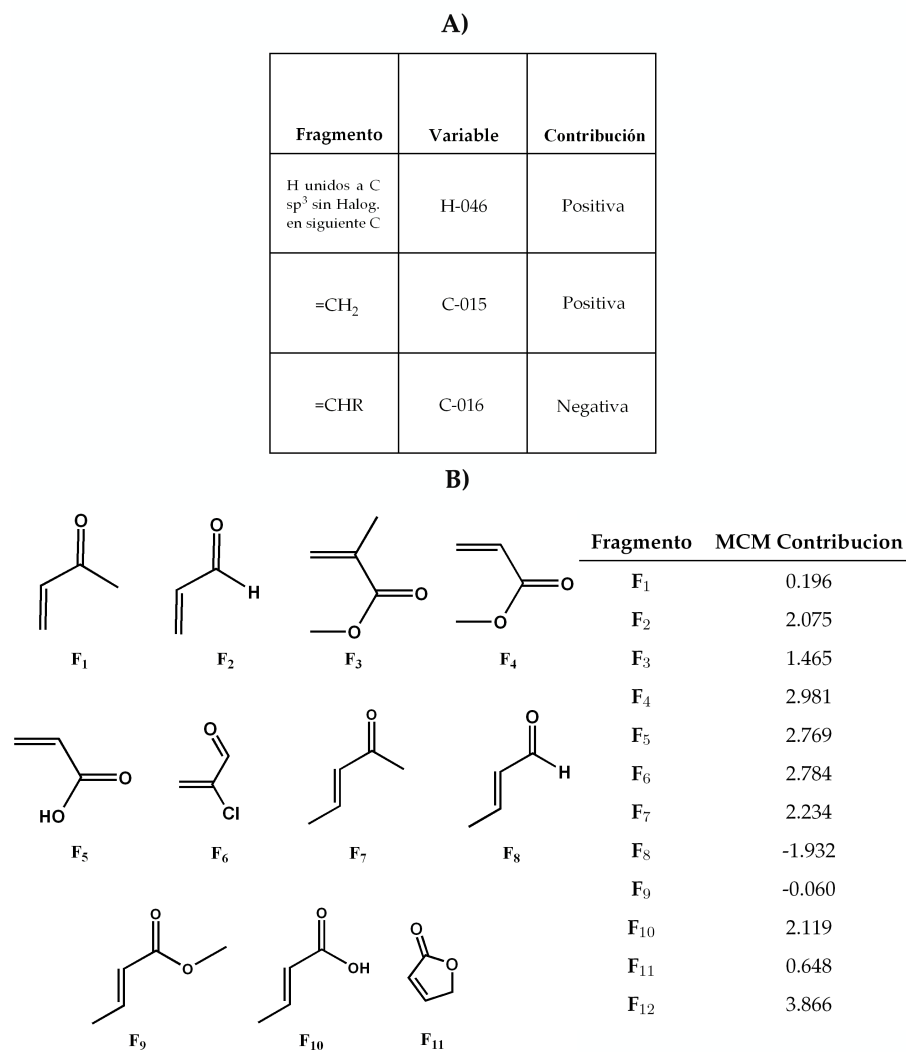


Figura 3.8: Fragmentos moleculares y contribuciones a la mutagenicidad del ensayo en células de mamífero de los modelos. A) QSAR5, B) QSAR3.

porque, como hemos dicho, los aceptores tipo Michael son electrófilos blandos y como tales su reactividad es mayor con nucleófilos blandos como el GSH, que, al parecer, es el paso principal para producir daño en el DNA en células de mamífero para estas sustancias^{189, 190}.

Además el modelo QSAR5 es otro de los modelos incluidos por la Unión Europea dentro de su base de datos para la evaluación regulatoria de productos químicos con arreglo al Reglamento REACH (QMRF Q14-26-8-160).

3.3. Complejación con la β -CD

En los estudios QSPR de complejación con la β -CD que realizamos de diversas familias de sustancias empleando para ello distintas familias de descriptores, están resumidos en la Tabla 3.8, junto con sus parámetros estadísticos.

Existen diferencias sustanciales en la varianza experimental explicada por ambos modelos QSPR1 y QSPR2, en comparación con el resto. Así, mientras que los modelos topológico y TOPS-MODE son capaces de explicar un 84 % y 86 % de la varianza experimental, respectivamente, los otros modelos, sólo pueden explicar un 79,9 %. La capacidad de predicción, expresada por Q_{CV-LOO}^2 y Q_{EXT}^2 , es superior en ambos modelos, incluso que en aquellos basados en descriptores tridimensionales (3DMoRSE, WHIM, RDF, GETAWAY y Randić).

Tanto el modelo QSPR1 como el QSPR2 determinan, para este conjunto de moléculas, que las interacciones estéricas (van der Waals) e hidrofóbicas son de primordial importancia en los procesos de inclusión con la β -CD. Como ya comentamos en la introducción, se ha observado que estas interacciones son las de mayor importancia en la complejación con la β -CD. Un ejemplo de esto lo podemos ver en los valores que adoptan las variables más significativas (${}^1\Omega Xu$ y ${}^2\Omega ZM1V$) de la ecuación 3.1 (Figura 3.9) para el QSPR1, donde se destaca que el aumento del tamaño molecular debido a la presencia de un anillo bencénico estabiliza el complejo debido a fuertes interacciones de van der Waals y la presencia de grupos hidroxilo alifáticos (más hidrofílicos) influye

negativamente.

Tabla 3.8: Mejores modelos QSPR obtenidos con las distintas familias de descriptores.

Familia	Descriptores	N ^o variables	N ²	R ²	F	s	AIC	FIT	Q ² _{CV-LOO}	Q ² _{boot}	a(R ²)	a(Q ²)	Q ² _{EXT}
TOFS-MODE ¹	$\mu_1 \mu_2^{Std}, \mu_1^{Hgd}, \mu_1^{Dip2}, \mu_3^{cW}, \mu_1$	8	185	0.868	145.61	0.329	0.12	4.65	0.851	0.845	0.007	-0.1	0.834
QSPR2	$\mu_1 \mu_4^{Dip2}, \mu_4^{Ab-logL16}, \mu_4^{Ab-\sum \beta_2^O}, \mu_4^{Pols}$	9	185	0.841	103.05	0.363	0.147	3.487	0.821	0.812	0.014	-0.105	0.764
Topológicos ¹ QSPR1	ZM1, ZM1V, SMTIV, LPRS, PHI, J, Xu, T(N..S), T(O..O)	10	185	0.799	69.34	0.414	0.193	2.425	0.776	0.763	0.025	-0.111	0.691
GETAWAY	HGM, H3m, H0v, HATSp, R6v, R4v+, R7e, R4p, R6p, R8p+	10	185	0.809	76.17	0.399	0.176	2.663	0.775	0.760	0.025	-0.114	0.741
Basados en autovalores	AEige, SEige, VRA1, VRx2, VRm2, SEigm, VRA2, VEA1, SEigv, VRp	10	185	0.797	68.63	0.416	0.195	2.399	0.771	0.731	0.025	-0.256	0.663
Conectividad	X0, X1, X1A, X2A, X0v, XMOD, X3, X3A, X3sol, RDCHI	10	185	0.792	66.80	0.420	0.199	2.336	0.765	0.754	0.023	-0.106	0.725
Autovalores de Burden	BEHm1, BELm4, BELm5, BELm7, BEHv1, BEHv8, BELv8, BEHe1, BELe7, BELe	10	185	0.760	55.48	0.452	0.230	1.940	0.727	0.712	0.026	-0.103	0.689
Propiedades moleculares	Ui, Hy, AMR, MLOGP2, GVWA18, Inflam-50, Hypert-50, Intect-80, Intect-50, BLTP96	10	185	0.764	56.77	0.448	0.226	1.985	0.726	0.705	0.028	-0.11	0.664
3DMolSE	Mor01u, Mor02m, Mor03m, Mor04m, Mor01v, Mor07v, Mor31e, Mor01p, Mor04p, Mor05p	10	185	0.734	48.40	0.476	0.255	1.692	0.689	0.667	0.027	-0.105	0.681
WHIM	L2m, L1p, L3s, Els, Tp, Au, Ae, As, Du, De	10	185	0.697	40.18	0.508	0.291	1.405	0.654	0.632	0.027	-0.108	0.480
RDF	RDF010u, RDF015u, RDF020u, RDF085u, RDF020m, RDF040m, RDF015v, RDF030e, RDF050p, RDF060p	10	185	0.513	18.46	0.644	0.459	0.645	0.453	0.532	0.031	-0.11	0.393
Perfiles moleculares de Randić	DR01, DR02, DR08, DFI7, SF01m, SP04, SP05, SP07, SPT7, SHP	10	185	0.513	18.46	0.644	0.459	0.645	0.453	0.532	0.031	-0.11	0.393

¹ Modelo ortogonalizado

² Número de sustancias

$$\begin{aligned}
 \log K = & 0.488(\pm 0.027)^1 \Omega Xu - 0.392(\pm 0.026)^2 \Omega ZM1V - 0.239(\pm 0.027)^3 \Omega LPRS + \\
 & + 0.076(\pm 0.026)^4 \Omega SMTIV + 0.337(\pm 0.026)^6 \Omega ZM1 - 0.134(\pm 0.027)^7 \Omega T(N..S) - \\
 & - 0.161(\pm 0.030)^8 \Omega PHI - 0.160(\pm 0.027)^9 \Omega J - 0.127(\pm 0.026)^{10} \Omega T(O..O) + \\
 & + 2.535(\pm 0.027)
 \end{aligned}
 \tag{3.1}$$

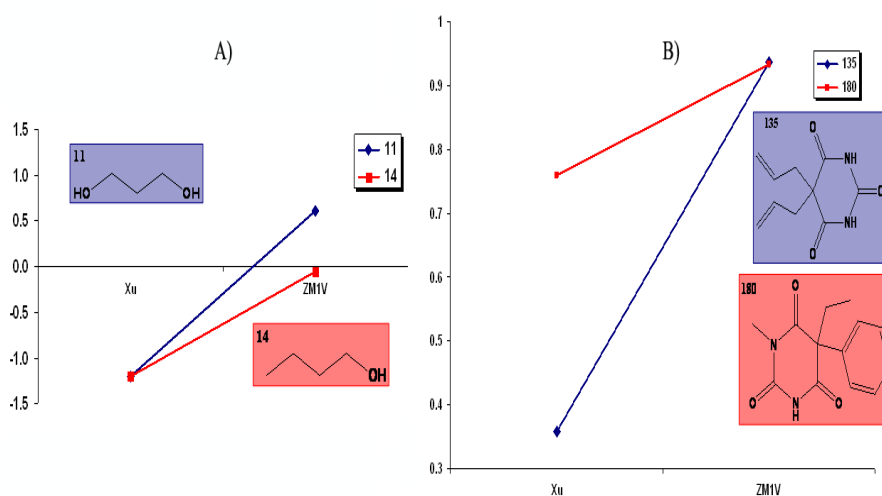


Figura 3.9: Contribuciones de cada variable al valor final de log K para A) 1,3-propanodiol (sustancia 11) y 1-butanol (sustancia 14) B) alobarbital (sustancia 135) y mefobarbital (sustancia 180)

Para el QSPR2 la influencia de las interacciones estéricas (van der Waals) e hidrofóbicas viene determinada por: (i) las variables ponderadas con la hidrofobicidad y el radio de van der Waals explican un 32.3 % y 28.5 %, respectivamente, de la varianza (ver Tabla 3 referencia 165); (ii) el carácter hidrofóbico de los fragmentos F_9 a F_{15} y F_{19} a F_{21} , los cuáles presentan unas elevadas contribuciones al fenómeno de la complejación (ver Figura 3.10) y (iii) el incremento de la ramificación (fragmentos F_9 a F_{11}) ya que esta es necesaria para lograr el óptimo desarrollo de los contactos de van der Waals con el interior de la β -CD.











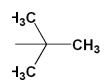
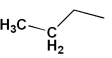
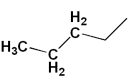
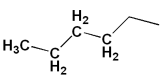





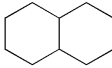
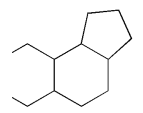





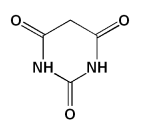
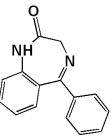


					Fragmento	Contribución
					F ₁	-0.361
F ₁	F ₂	F ₃	F ₄	F ₅	F ₂	-0.081
					F ₃	-0.062
					F ₄	0.048
					F ₅	-0.208
F ₆	F ₇	F ₈	F ₉	F ₁₀	F ₆	0.066
					F ₇	0.126
					F ₈	0.627
					F ₉	0.276
					F ₁₀	0.611
F ₁₁	F ₁₂	F ₁₃	F ₁₄	F ₁₅	F ₁₁	0.912
					F ₁₂	0.363
					F ₁₃	0.521
					F ₁₄	0.685
					F ₁₅	0.464
F ₁₆	F ₁₇	F ₁₈	F ₁₉	F ₂₀	F ₁₆	-0.410
					F ₁₇	-0.421
					F ₁₈	-1.078
					F ₁₉	0.598
					F ₂₀	0.912
					F ₂₁	1.454
F ₂₁	F ₂₂	F ₂₃	F ₂₄	F ₂₅	F ₂₁	-0.156
					F ₂₂	-0.116
					F ₂₃	-0.587
					F ₂₄	-0.178
					F ₂₅	0.042
					F ₂₆	-1.162
F ₂₆	F ₂₇	F ₂₈	F ₂₉	F ₃₀	F ₂₆	-0.558
					F ₂₇	0.152
					F ₂₈	0.064

Figura 3.10: Fragmentos moleculares seleccionados y contribuciones a la constante de complejación con la β -CD.

Los descriptores TOPS-MODE (QSPR2) permiten el desarrollo de QSPR robustos y con carácter predictivo de similar a los que usan descriptores globales¹³⁶. Por otro lado, permiten la interpretación de los resultados en términos de contribuciones de fragmento identificando aquellos grupos o fragmentos moleculares que pueden ser responsables de la propiedad estudiada de la misma manera que los estudios subestructurales. Para ésto, los estudios subestructurales necesitan recoger una cantidad importante de datos para cada tipo de compuestos, mientras TOPS-MODE, es capaz de reconocer este patrón estructural para sólo un compuesto presente en el set de datos¹⁶⁰. Ésto es posible debido a que estos descriptores, los cuales describen la estructura molecular como un todo, en términos de características hidrofóbicas, estéricas y electrónicas de las moléculas y al mismo tiempo pueden ser transformados en contribuciones locales. Además, permiten obtener nuevas hipótesis, que pueden constituir la base para nuevas interpretaciones estructurales después de confirmaciones experimentales.

3.4. Predicciones de complejación con la β -CD y mutagenicidad para los ácidos haloacéticos y los monómeros dentales

Si la utilidad de los modelos QSAR radica en su habilidad para predecir con precisión la actividad para las nuevas sustancias, esta precisión debe quedar asegurada de algún modo. Así, además de la validación del modelo, se debe incluir la determinación de un dominio de aplicación definido en el espacio de los descriptores moleculares empleados para obtener el mismo. Existen varios métodos para evaluar el dominio de aplicación de los modelos QSAR^{192,193} pero de ellos el más común es la determinación de los valores de *leverage* (h) de cada compuesto¹⁹⁴. Una sustancia estaría dentro del dominio de aplicación del modelo cuando presente un valor de *leverage* (h) inferior al umbral h^* (h^* se fija generalmente $3p/n$, donde n es el número de los compuestos y p el número de parámetros del modelo). Sólomente las predicciones para las sustancias cuyo valor de h esté dentro del dominio de aplicación en el cual se ha construido el modelo pueden considerarse de confianza¹⁹⁵. En la Tabla 3.9 tenemos los valores de

mutagenicidad y complejación con la β -CD para los HAA y en la Tabla 3.10 para los carbonilos α - β -insaturados.

Tabla 3.9: Valores de potencia mutagénica y complejación con la β -CD reales y predichos por los distintos modelos para la familia completa de ác. haloacéticos.

NOMBRE	QSAR1	$h^* = 0.60$	QSPR1	$h^* = 0.14$	QSPR2	$h^* = 0.129$
	logTA100	h	log K Pred.	h	log K Pred.	h
fluoroiodoacético	1.698	0.447	0.168	0.103	0.929	0.065
difluoroiodoacético	1.008	0.39	-0.424	0.246	1.503	0.096
iodoacético	0.776	-	0.759	0.044	0.795	0.042
bromoacético	0.364	-	0.823	0.038	0.325	0.069
cloroiodoacético	-0.557	0.281	0.735	0.079	1.142	0.085
clorofluoroiodoacético	-0.832	0.235	0.171	0.214	1.979	0.152
dibromoacético	-1.203	-	0.753	0.078	1.018	0.099
bromocloroacético	-1.207	-	0.792	0.077	0.784	0.085
bromoiodoacético	-1.326	0.371	0.693	0.083	1.378	0.108
bromofluoroiodoacético	-1.342	0.303	0.147	0.212	2.288	0.212
bromofluoroacético	-1.527	0.164	0.216	0.101	0.578	0.078
tricloroacético	-1.806	0.385	0.822	0.206	2.088	0.116
dicloroacético	-1.83	-	0.827	0.078	0.555	0.079
fluoroacético	-1.856	0.158	0.287	0.065	0.021	0.078
cloroacético	-1.943	-	0.864	0.036	0.124	0.074
clorofluoroacético	-1.951	0.186	0.246	0.101	0.363	0.078
difluoroacético	-1.972	0.191	-0.346	0.144	0.201	0.084
dibromofluoroacético	-2.401	0.336	0.184	0.215	2.190	0.147
bromodicloroacético	-2.41	0.368	0.798	0.202	2.408	0.168
diclorofluoroacético	-2.489	0.396	0.225	0.221	1.596	0.072
bromodifluoroacético	-2.713	0.395	-0.395	0.248	1.427	0.071
dicloroiodoacético	-2.714	0.555	0.758	0.197	2.518	0.237
tribromoacético	-2.756	0.341	0.743	0.195	3.085	0.300
bromoclorofluoroacético	-2.844	0.392	0.206	0.218	1.887	0.102
dibromocloroacético	-2.949	0.364	0.772	0.198	2.740	0.231
clorodifluoroacético	-3.254	0.445	-0.381	0.252	1.163	0.065
trifluoroacético	-3.274	0.441	-1.215	0.310	0.786	0.086
bromocloroiodoacético	-4.402	0.695	0.729	0.195	2.857	0.308
dibromoiodoacético	-5.554	0.798	0.697	0.193	3.210	0.379
fluorodiodoacético	-8.45	0.907	0.105	0.210	2.429	0.300
diodoacético	-9.656	0.926	0.627	0.091	1.756	0.138
clordiiodoacético	-15.668	0.967	0.682	0.193	3.020	0.397
bromodiodoacético	-18.84	0.977	0.646	0.193	3.381	0.465
triiodoacético	-45.679	0.996	0.592	0.195	3.590	0.546

Respecto a la mutagenicidad de los HAA vemos que los ácidos fluoroiodoacético y difluoroiodoacético presentan valores altos de potencia, incluso mayores que los más mutagénicos conocidos para esta familia como son, el ác. iodoacético o bromoacético. Además es posible de que estas dos sustancias estén presentes en aguas fluoradas ri-

Tabla 3.10: Valores de actividad mutagénica y complejación con la β -CD reales y predichos por los distintos modelos para los monómeros dentales más comunes.

NOMBRE	AMES QSAR2 ^a	AMES QSAR4 ^a	MCGM QSAR3 ^a	MCGM QSAR5 ^a	log K Pred. QSPR1 ^a	log K Pred. QSPR2 ^a
trietilenglicol dimetacrilato (TEGDMA)	-1 ^b	-1 ^b	1 ^b	1 ^b	-0.827	2.912
bisfenol-A- glicidil metacrilato (Bis-GMA)	-1 ^b	-1 ^b	1	1	3.467	1.351
uretano dimetacrilato (UDMA)	-1	-1	1	-1	-1.443	4.314
hidroxietilmetacrilato (HEMA)	-1 ^b	-1 ^b	1	1	0.632	1.310
metilmetacrilato (MMA)	-1	-1	1	1	0.765	0.985
butilmetacrilato (BMA)	-1 ^b	-1 ^b	1	1	1.114	2.377
etilmetacrilato (EMA)	-1 ^b	-1 ^b	1	1	0.895	1.574
acrilato de metilo	-1 ^b	-1 ^b	1 ^b	1 ^b	0.765	0.563
Etil acrilato	-1	-1	1 ^b	1 ^b	0.861	1.180
2-hidroxietil acrilato	-1 ^b	-1 ^b	1 ^b	1 ^b	0.559	0.929
2-etoxietil acrilato	-1	-1	1	1	0.557	1.777
2-etoxipropil acrilato	-1	-1	1	1	0.572	2.273
tetrahidrofurfuril metacrilato (THFMA)	-1	-1	1	1	2.448	2.348
hexane-1,6-diyl bis(2-methylacrylate) (1,6-ADMA)	-1	-1	1	1	0.418	3.543
octane-1,8-diyl bis(2-methylacrylate) (1,8-ADMA)	-1	-1	1	1	0.423	3.972
3-(acryloyloxy)-2-hydroxypropyl methacrylate (GAM)	-1	-1	1	1	-0.286	2.299
2-hydroxypropane-1,3-diyl bis(2-methylacrylate) (GMR)	-1	-1	1	1	-0.190	2.538
2-(phenylcarbamoyloxy)ethyl methacrylate (MEPC)	-1	-1	1	1	2.024	2.707
6-hydroxyhexyl methacrylate (6-HHMA)	-1	-1	1	1	0.775	2.767
4-(2,3-diamino-3-oxopropyl)phenyl methacrylate (MTYA)	-1	-1	1	1	2.392	2.890

^a Valores en negrita corresponden a predicciones fuera del dominio de aplicación del modelo correspondiente.

^b Valores experimentales.

cas en bromuro/ioduro. En futuras determinaciones experimentales de mutagenicidad sería interesante priorizar estas dos sustancias frente a la familia completa de HAA.

Respecto a la complejación con la β -CD de los HAAs el modelo QSPR2 obtenido con los descriptores basados en la aproximación TOPS-MODE, consigue obtener predicciones para un mayor número de sustancias que el QSPR1, que, como se ha dicho, es obtenido con los descriptores topológicos. En general, los valores obtenidos del log K no superan los valores más comunes de complejación para fármacos o ingredientes alimentarios (log $K=2,3$ con valores medios de 2,69)^{138, 196}, por lo que es de esperar que la interacción entre los ácidos haloacéticos y la β -CD sea de escasa importancia.

Por otro lado, para los monómeros dentales (ver Tabla 3.10), entre los que cabe destacar el sistema Bis-GMA/TEGDMA¹⁹⁷ y el UDMA, (el cual mejora en resistencia a los anteriores¹⁹⁸) podemos observar como los modelos QSAR desarrollados presentan predicciones negativas para el ensayo de Ames y un carácter mutagénico para el ensayo con células de mamífero; si bien, destacar que el monómero UDMA presenta ambigüedad para este último punto final. La mutagenicidad predicha por estos modelos QSAR para el ensayo con células de mamífero genera una alerta respecto a estas sustancias, las cuáles se deben tener en cuenta para futuras determinaciones experimentales.

En cuanto a la complejación con la β -CD de estos monómeros observamos, al igual que sucede con los HAA, que el QSPR2 obtiene predicciones válidas para un mayor número de sustancias que el QSPR1. Hay que resaltar que monómeros como el TEGDMA, 1,6-ADMA, 1,8-ADMA, GMR, MEPC y 6-HHMA, anteriormente predichos como mutagénicos, presentan valores de complejación. Por lo tanto, estas sustancias podrían desplazar de sus complejos a aquellos fármacos o ingredientes alimentarios cuyo valor de log K sea inferior a sus valores correspondientes, pudiéndose llegar a algún tipo de interacción. Para futuras investigaciones, sería interesante determinar experimentalmente la mutagenicidad o carcinogenicidad de estas sustancias predichas como mutagénicas, además de determinar su complejación con la β -CD y el efecto derivado de ésta interacción (antagonismo o sinergismo) mediante ensayos *in vivo*.

CONCLUSIONES

Las siguientes conclusiones fueron obtenidas en relación a los objetivos de este trabajo.

- Fue posible modelar la potencia mutagénica en *S. Typhimurium* cepa TA100 sin activación metabólica para una serie de derivados halogenados empleando para ello distintas familias de descriptores. El modelo QSAR1 basado en los descriptores derivados de TOPS-MODE obtuvo los mejores resultados estadísticos, prediciendo mutagenicidad para los ácidos fluoroiodoacético y difluoroiodoacético, los cuáles podrían encontrarse en aguas fluoradas ricas en bromuro (o ioduro). Por tanto, deberían confirmarse estos datos experimentalmente, además de controlarse y regularse los valores máximo admisibles de estas sustancias en el agua potable para evitar efectos mutagénicos.
- Fue posible modelar la actividad mutagénica en *S. Typhimurium* y en células de mamífero para una serie de carbonilos α, β -insaturados, empleando para ello distintas familias de descriptores. Se obtuvieron varios modelos con capacidad suficiente para predecir la actividad mutagénica (QSAR2 y QSAR4 para el ensayo de Ames y QSAR3 y QSAR5 para el ensayo en células de mamífero). Además, los modelos QSAR desarrollados en este trabajo han proporcionado nuevas pruebas que avalan su utilidad desde el punto de vista del modelado molecular (bajos recursos computacionales); a su vez pueden ser empleados como instrumentos para la evaluación del riesgo toxicológico, incluyéndose en aquellos sistemas de

expertos donde las SAs estén implementadas, como ocurre en el TOXTREE.

- Fue posible modelar la complejación con la β -CD para una serie de moléculas no-congénicas empleando para ello distintas familias de descriptores. Con los modelos obtenidos (QSPR1 y QSPR2) se hicieron predicciones para las dos familias de compuestos, observándose que para los HAA esta interacción no sería termodinámicamente significativa y, en cambio, sí lo sería para los carbonilos α , β -insaturados mutagénicos como el TEDGMA, 1,6-ADMA, 1,8-ADMA, GMR, MEPC y 6-HHMA, los cuáles podrían llegar a desplazar al huésped de sus complejos.
- Los mecanismos de actuación mutagénica tanto para los HAA como para los carbonilos α , β -insaturados, así como los mecanismos de complejación con la β -CD deducidos de los modelos obtenidos corroboran las hipótesis realizadas hasta el momento por otros autores.
- Las alertas estructurales identificadas de mutagenicidad tanto para los derivados halogenados como para los carbonilos α , β -insaturados, así como las contribuciones de determinados fragmentos a la constante de complejación con la β -CD, fueron consistentes con las evidencias experimentales. Además, las SAs encontradas podrían ayudar a la elucidación del posible mecanismo de acción mutagénica para cada uno de los ensayos estudiados. Al mismo tiempo, tanto para el estudio de actividades (mutagenicidad) como para el de propiedades (complejación), se vuelve a demostrar la utilidad de los descriptores basados en la aproximación TOPS-MODE como *generadores de conocimiento*.

4.1. Recomendaciones

Como recomendación final a este trabajo cabe señalar la necesidad de estudiar experimentalmente la mutagenicidad de sustancias como los ácidos fluoriodoacéti-

co, difluoriodoacético, TEDGMA, 1,6-ADMA, 1,8-ADMA, GMR, MEPC y 6-HHMA, así como los efectos sobre la salud debidos a su interacción con la β -CD.

Parte I

ANEXO



ELSEVIER

Available online at www.sciencedirect.com

Bioorganic & Medicinal Chemistry 16 (2008) 5720–5732

Bioorganic &
Medicinal
Chemistry

Halogenated derivatives QSAR model using spectral moments to predict haloacetic acids (HAA) mutagenicity

Alfonso Pérez-Garrido,^{a,*} Maykel Pérez González^{b,c,✱} and Amalio Garrido Escudero^{a,*}

^aEnvironmental Engineering and Toxicology Department, Catholic University of San Antonio, Guadalupe, Murcia, C.P. 30107, Spain

^bMolecular Simulation and Drug Design Group, Chemical Bioactive Center, Central University of Las Villas, Santa Clara, Villa Clara, C.P. 54830, Cuba

^cDepartment of Organic Chemistry, Vigo University, C.P. 36200, Vigo, Spain

Received 19 November 2007; revised 29 February 2008; accepted 25 March 2008

Available online 30 March 2008

Abstract—The risk of the presence of haloacetic acids in drinking water as chlorination by-products and the shortage of experimental mutagenicity data for most of them requires a research work. This paper describes a QSAR model to predict direct mutagenicity for these chemicals. The model, able to describe more than 90% of the variance in the experimental activity, was developed with the use of the spectral moment descriptors. The model, using these descriptors with multiplicative effects provides better results than other linear descriptors models based on Geometrical, RDF, WHIM, eigenvalue-based indices, 2D-autocorrelation ones, and information descriptors, taking into account the statistical parameters of the model and the cross-validation results. The structural alerts and the mutagenicity-predicted values from the model output are in agreement with references from other authors. The mutagenicity predicted values for the three haloacetic acids, which have available experimental data (TCAA—Trichloroacetic acid, BDCAA—Bromodichloroacetic acid, and TBAA—Tribromoacetic acid), are reasonably close to their experimental values, specially for the latest two.

© 2008 Elsevier Ltd. All rights reserved.

1. Introduction

The introduction of water disinfection processes was a significant success in the control of waterborne diseases.¹ Drinking water disinfection is required to remove harmful pollutants, including pathogenic microorganisms. Some studies demonstrated that concentrated extracts of disinfected drinking water were toxic in many in vivo and in vitro bioassays.² Most drinking water disinfection by-products form as a result of the reaction between organic matter in raw water and chemical disinfectants like chlorine. These organic compounds come from two major sources: (1) breakdown products of naturally occurring materials (NOM), which include humic acids, microorganisms and their metabolites, and some petroleum-based high molecular weight aliphatic and aromatic hydrocarbons; and (2) products

from domestic and commercial activities, including agricultural and urban runoff and wastewater discharges. Drinking water disinfection by-products (DBPs) represent an important class of environmentally hazardous chemicals. They can increase the risk for human health in long-term basis. Epidemiological studies demonstrate that individuals who consume chlorinated drinking water are exposed to a higher risk of developing a cancer of stomach, pancreas, kidney, bladder, and rectum as well as Hodgkins and non-Hodgkins lymphoma.^{3–5} The issues of human exposure to DBPs for epidemiological and health risk assessment were recently reviewed.⁶

The haloacetic acids are the second greater group of drinking water disinfection by-products and some of them are rodent liver carcinogens^{7–9} and mutagenic in *Salmonella typhimurium*.^{10–15} There is a shortage of information about carcinogenic and mutagenic potency for these chemicals. The use of methods which are able to predict such values are important for toxicological risk assessment. This is the reason for our having developed this QSAR model in order to predict the mutagenicity in *S. typhimurium* strain TA100. The mutagenic-

Keywords: QSAR; Spectral moments; Haloacetic acids; Mutagenicity.

* Corresponding authors. Tel.: +34 968 278 755 (A.P.-G.); +34 968 278 544 (A.G.E.); e-mail addresses: Aperez@pdi.ucam.edu; AGescudero@pdi.ucam.edu

✱ Deceased.

ity in *S. typhimurium* as determined by the Ames test is used world-wide for initial screening to determine the mutagenic potential of new chemicals and drugs. It is known that there is a high predictive value for rodent carcinogenicity when a mutagenic response is obtained.^{16–18}

A Ref. 19 has been found describing a QSAR model to infer the mechanism of action of mono-, di-, and tri-halogenated acids by relating the 1- and 2-atom fragment contributions towards the total toxicity as predicted by TOPKAT to certain types of descriptor. The theory being that there is a good correlation between the two is a reasonable indication of a particular mechanism of action.

There are more than 3200 molecular descriptors that can be used to solve the problem outlined above.²⁰ A useful kind of descriptors in Medicinal Chemistry have been introduced some years ago.^{21–27} The descriptors based on spectral moments are a good example of these ones. Recently, the MARCH-INSIDE (MARKovian CHEMicals IN Silico DESIGN) descriptors were introduced.^{28–30} This kind of descriptors have proved to be very useful in studies with proteins,^{31,32} anticancer compounds³³, and antimicrobials.³⁴ The other type of descriptors based on spectral moments are named TOPS-MODE (TOPological Sub-structural MOlecular DESIGN) descriptors^{35–37} and they are the spectral moments of the bond matrix weighted in the main diagonal with different physicochemical parameters. These descriptors are easy to calculate and they can be used when there are a heterogeneous series of compounds.^{22,38}

The successful application of this theoretical approach to the modeling of toxicological^{39–41} and ecotoxicological^{42,43} properties has also inspired us to perform a more exhaustive study. The intention was to test and validate TOPS-MODE applicability in assessing discovery of leads and mutagenic impact of chemicals. The selection of a data set on mutagenic toxicity is not random.

Descriptors based on spectral moments show a clear interest in new QSAR model research in bioorganic and medicinal chemistry fields. We focus our research work to develop a QSAR model based on spectral moments because of the advantages of easy use and understanding.

2. Materials and methods

2.1. Data set

A data set of 42 halogenated derivatives was collected from literature and U.S. National Toxicological Program (NTP) (Table 1). The updated NTP database is also available at the following web site: http://ntp-apps.niehs.nih.gov/ntp_tox/index.cfm. In this data set are included nitrohaloalkanes, haloacids, haloaldehydes, halocetones, haloalcohols, haloepoxides, and haloalkanes and the most of them are well-known alkylating agents. A big group of such compounds is present in drinking water as disinfection by-products.

The activity is defined as the logarithm of TA100 strain *S. typhimurium* Ames test⁴⁴ without activation and with preincubation. It is calculated as the slope of the linear portion of the dose–response curve.⁴⁵ The data set includes the values from Plewa et al.^{13,15} The preincubation protocol used by these authors (1 h instead of 20 min) is different from that of the NTP protocol. The values from the previous by mentioned authors were adjusted using a correction factor calculated as the ratio between the NTP values (tested following a preincubation period of 20 min) and the Plewa et al. values.

2.2. The TOPS-MODE descriptors

The TOPS-MODE descriptors are based on the calculation of the spectral moments of the so-called bond matrix.⁴⁶ The theoretical basis has been described in previous reports.^{35,36} Nevertheless, an overview of this descriptor family is going to be given below. The bond matrix is defined as a square and symmetric matrix whose entries are ones or zeros if the corresponding bonds are adjacent or not. The order of this matrix (m) is the number of bonds in the molecular graph, being two bonds adjacent if they are incident to a common atom. The spectral moments of the edge adjacency matrix are defined as the traces. That is the sum of the main diagonal of the different powers of such matrix. Several bond weights such as standard bond distance (Std), standard bond dipole moments (Dip, Dip2), hydrophobicity (H), polar surface area (Pols), polarizability (Pol), molar refractivity (Mol), van der Waals radii (vdW), and Gasteiger–Marsilli charges (Gas) were used for computing the spectral moments of the bond matrix. Since most of the approaches for computing physicochemical properties from fragment are based on atom-additive methods, several transform from atomic to bond contributions were carried out. The way in which these atomic contributions were transformed into bond contributions has been described by Estrada et al.⁴⁷:

$$w(i, j) = \frac{w_i}{\delta_i} + \frac{w_j}{\delta_j} \quad (1)$$

where w_i and δ_i are the atomic weight and vertex degree of the atom i . The calculation of the TOPS-MODE descriptors was carried out with the software MODE-SLAB 1.0.⁴⁸ The input of the software consists of SMILES codes for each compound.⁴⁹ We calculated the first 15 spectral moments ($\mu_1–\mu_{15}$) for each bond weight and the number of bonds in the molecules (μ_0). Also, we multiplied μ_0 and μ_1 for the first 15 spectral moments obtaining 30 new variables. These variables include very valuable information due to the nonlinear behaviour of the biological process.²⁶ To apply the current approach to the structure–toxicity relationship, the following steps should be followed: first, to select an adequate training set according to the aim and scope of the research. Second, to draw the molecular graphs for each molecule included in the training set. The third step is to differentiate the molecular bonds with appropriate weights. The fourth one is to compute the spectral moments of the bond matrix for each molecule of the

Table 1. Names, CAS number, mutagenic potency, and reference to the compounds used in this study

Compounds	Name	CAS	Log TA100	Reference
1	2,4,4-Trichloro-3-(dichloromethyl)-2-butenoic acid	97055-37-3	2.821	95
2	(Z)-2,4,4-Trichloro-3-formyl-2-butenoic acid	117823-31-1	4.070	95
3	(Z)-2,4-Dichloro-3-formyl-2-butenoic acid	—	2.847	95
4	(Z)-2-Chloro-3-methyl-4-oxo-2-butenoic acid	—	−0.319	95
5	(E)-2,4,4-Trichloro-3-(chloromethyl)-2-butenoic acid	—	3.466	95
6	(S)-2,3-Dibromopropanal	5221-17-0	−0.522	96
7	Dichloroacetic acid	79-43-6	−1.829	13
8	Chloroacetic acid	79-11-8	−1.943	13
9	Bromoacetic acid	79-08-3	0.363	13
10	2-Bromopropane	75-26-3	−1.716	NTP
11	2,3-Dichloro-1-propene	78-88-6	−0.0065	NTP
12	(R)-1,2-Dichloropropane	78-87-5	−2.473	NTP
13	2,2-Dichloroacetyl Chloride	79-36-7	−1.212	NTP
14	1,3-Dichloro-2-propanol	96-23-1	−1.837	NTP
15	(S)-2,3-Dibromo-1-propanol	96-13-9	−0.334	NTP
16	1,2-Dibromoethane	106-93-4	−0.607	NTP
17	(R)-2-(Chloromethyl)oxirane	106-89-8	−0.692	NTP
18	2-Chloroacetaldehyde	107-20-0	−0.789	NTP
19	2-Chloroethanol	107-07-3	−2.074	NTP
20	3-Chloro-1-propene	107-05-1	−2.326	NTP
21	(E)-1,4-Dichloro-2-butene	110-57-6	0.179	NTP
22	(2S,3S,4S,5S)-1,6-Dibromohexane-2,3,4,5-tetraol	488-41-5	−1.607	NTP
23	(R)-2-(Fluoromethyl)oxirane	503-09-3	−1.092	NTP
24	2-Bromoethanol	540-51-2	−2.075	NTP
25	(E)-1,3-Dichloro-1-propene	542-75-6	0.010	NTP
26	(S)-3-Iodo-1,2-propanediol	554-10-9	−1.260	NTP
27	2-Chloro-2-nitropropane	594-71-8	−1.874	NTP
28	(R)-1-Chloro-1-nitropropane	600-25-9	−0.679	NTP
29	(S)-2,3-Dichloro-1-propanol	616-23-9	−0.746	NTP
30	3-Bromo-1-propanol	627-18-9	−1.694	NTP
31	Dibromoacetic acid	631-64-1	−1.203	NTP
32	2-Chloroethyl acrylate	2206-89-5	−1.214	NTP
33	(S)-2-(2,2,2-Trichloroethyl)oxirane	3083-25-8	−1.193	NTP
34	(S)-2-(Trichloromethyl)oxirane	3083-23-6	0.455	NTP
35	(R)-2-(Bromomethyl)oxirane	3132-64-7	−0.371	NTP
36	3-Chloro-N,N-dimethyl-1-propanamine	5407-04-5	−2.400	NTP
37	Bromochloroacetic acid	5589-96-8	−1.207	NTP
38	(S)-2,3-Dibromopropyl acrylate	19660-16-3	−1.550	NTP
39	(R)-1-Bromo-2-propanol	19686-73-8	−1.793	NTP
40	2-Bromoacrylaldehyde	14925-39-4	−0.207	97
41	3-Bromo-3-buten-2-one	61203-01-8	−0.886	97
42	Iodoacetic acid	64-69-7	0.776	15

data set. The fifth step is to find a quantitative structure–toxicity relationship by using a regression analysis:

$$P = a_0\mu_0 + a_1\mu_1 + a_2\mu_2 + \dots + a_k\mu_k + b \quad (2)$$

where P is the studied activity, in our case, the log TA100 partitioning, μ_k is the k -th spectral moment, and the a_k are the coefficients obtained by linear regression. The sixth step is to test the predictive capability of the regression model by cross-validation procedures and an external prediction set. And finally, to compute the contribution of the different substructures to determine their quantitative contribution to the mutagenicity of the studied molecules.

2.3. Structural alerts identification

The identification of structural alerts (fragment contribution) to the toxicity is based on bond contributions. This procedure, implemented in MODESLAB software, consists in transforming a QSAR model into a bond additive scheme. As a result we calculate for each mole-

cule the toxicological property as a sum of bond contributions. Bond contributions are derived from the local spectral moments. They are defined as the diagonal entries of the different powers of the weighted E matrix.

$$\mu_k^T(i) = b_{ii}(T)^k \quad (3)$$

where $\mu_k^T(i)$ is the k -th local spectral moment of the bond i , $b_{ii}(T)$ are the diagonal entries of the weighted E matrix and T is the type of bond weight.

For a given molecule, we can substitute the values of the local spectral moments computed by Eq. 3 into Eq. 4 and thus gather the total contribution to the toxicity of its different bonds.

$$P = b_0 + \sum_k a_k \cdot \mu_k^T \quad (4)$$

Since the activity modeled is expressed as log TA100, positive bond contributions increase the TA100 value and increase the mutagenic activity and vice versa. The structural information highlighted by the bond contri-

butions may allow, together with other theoretical and experimental data, a better understanding of the mechanisms of mutagenic action of the involved chemicals. Also, it can be useful for the proposal of new metabolic routes associated with the mutagenesis phenomenon.

2.4. Computational strategies

Calculation of spectral moments was carried out using Modeslab 1.0 software⁴⁸ taking the Simplified Molecular Input Line Entry Specification (SMILES) format⁴⁹ of the geometrically optimized compound structure by Cosmic module of Tsar 3.3 software (Accelrys Inc., <http://www.accelrys.com>). The other family of descriptors like Geometrical (74 descriptors), RDF (150 descriptors), WHIM (99 descriptors), eigenvalue-based indices (44 descriptors), and 2D-autocorrelation (96 descriptors) was calculated used Dragon Web.⁵⁰ The variables with constants or close to constants values were deleted. The mathematical models were obtained by means of Multiple Regression Analysis (MRA) as implemented in the Tsar 3.3 software. The variables to be included in the equation were selected using forward stepwise procedure as variable selection strategy.⁵¹

2.5. Model selection and validation

The statistical significance of the models was determined by examining the squared regression coefficient (R^2), the standard deviation (s), the Fisher ratio (F) and the ratio between the number of cases and the number of adjustable parameters in the model ρ statistic which we assume as the criterion $\rho \geq 4$.⁵²

In addition, further criteria exist to compare the quality of the models obtained. One of them uses the correlation coefficient R which is barely meaningful because it tends to select as many variables as possible as well as the standard deviation s . The other criterion is the Kubinyi function (FIT), being closely related to the F value, which was created and proved to be useful.^{53,54} The best model will be the one that exhibits the high value of this function. The other of these criteria was formulated by Akaike sometime ago.^{55,56} Akaike's information criteria (AIC) take into account the statistical goodness of fit and the number of parameters that have to be estimated to achieve that degree of fit. The model that produces the minimum value of these statistics should be considered potentially the most useful. The outliers detection was carried out for the compounds that have large residual.

The robustness of the models and their predictivity were evaluated by Q^2 leave-one-out (LOO) cross-validation, an equivalent statistic to R^2 , and bootstrapping test (Q_{boot}^2). The stability when a heavy perturbation in the training set is applied was checked by response randomization (Y-scrambling) ($a(R^2)$ and $a(Q^2)$) procedures. These calculations were carried out with the software Mobydigs Computer Software 1.0.⁵⁷ To sum up, a good quality of the models was indicated by high values in F , FIT, and ρ , lower values in AIC and s , as well as close to one values in R^2 , Q^2 , and Q_{boot}^2 .

2.6. Orthogonalization of descriptors

The main drawback of collinearity from the point of view of a QSAR model is about the stability of the coefficients in the linear regression model. In the case of the TOPS-MODE descriptors this can be translated into false interpretation of bond contributions. The magnitude and sign of them can be falsified by the effect produced by the existence of collinear variables in the model. We employed the Randić' method of orthogonalization which has been described in detail in several papers.^{58–62} Thus, we will give only a general overview here.

The first step for orthogonalizing the molecular descriptors is to select the appropriate order of orthogonalization, which, in this case, is the order of significance of the variables in the model. The first variable (v_1) is taken as the first orthogonal descriptors Ωv_1 and the second one is orthogonalized respect to it by taking the residual of its correlation with Ωv_1 . The process is repeated until all variables are completely orthogonalized and the orthogonal variables are then used to obtain the new model. For the extraction of the information contained in the orthogonalized descriptors we followed the procedure reported by Estrada et al.²²

2.7. Applicability domain of the models

Once we obtained the model we defined its applicability domain to make predictions for the rest of haloacetic acids. There are several methods for assessing the applicability domain (AD) of QSAR models⁶³ but the most common one encompasses determining the leverage values for each compound.⁶⁴ To visualize the AD of a QSAR model, the plot of standardized residuals versus leverage values (h) (the Williams plot) can be used for an immediate and simple graphical detection of both the response outliers (i.e., compounds with standardized residuals greater than two standard deviation units) and structurally influential chemicals in a model ($h > h^*$). These calculations were carried out with the software Mobydigs Computer Software 1.0.⁵⁷ Figure 8 shows the Williams plot, i.e., for each compound of the training set. From this plot, the applicability domain is established inside a squared area within ± 2 standard deviations and a leverage threshold h^* of 0.68 ($h^* = 3\kappa/n$, being κ the number of model parameters and n the number of objects). For making predictions, predicted mutagenicity data must be considered reliable only for those chemicals that fall within the applicability domain on which the model was constructed.⁶⁵

3. Results and discussion

3.1. QSAR model

The best QSAR model obtained with the spectral moments is given as follows together with the statistical parameters of the regression.

$$\begin{aligned} \log TA100 = & 8.052 \cdot 10^{-11} (\pm 1.54 \cdot 10^{-11}) \mu_{15}^{\text{Std}} \\ & - 5.302 \cdot 10^{-2} (\pm 1.12 \cdot 10^{-2}) \mu_1^{\text{Pols}} \\ & + 5.723 \cdot 10^{-8} (\pm 1.09 \cdot 10^{-8}) \mu_7^{\text{Mol}} \\ & - 7.991 \cdot 10^{-10} (\pm 1.26 \cdot 10^{-10}) \mu_0 \mu_{13}^{\text{Dip}} \\ & + 2.152 \cdot 10^{-8} (\pm 4.39 \cdot 10^{-9}) \mu_0 \mu_5^{\text{Pols}} \\ & + 1.599 \cdot 10^{-2} (\pm 1.59 \cdot 10^{-3}) \mu_1 \mu_2^{\text{Dip}^2} \\ & - 3.331 \cdot 10^{-10} (\pm 1.01 \cdot 10^{-10}) \mu_1 \mu_{14}^{\text{Dip}^2} \\ & - 2.208 \cdot 10^{-8} (\pm 4.66 \cdot 10^{-9}) \mu_1 \mu_9^{\text{Pol}} \\ & - 1.163 (\pm 0.28) \end{aligned}$$

$N = 42$; $R^2 = 0.842$; $Q_{(\text{CV-LOO})}^2 = 0.684$;
 $s = 0.679$; $F = 21.98$; $\text{AIC} = 0.659$; $\text{FIT} = 1.659$
 $Q_{\text{boot}}^2 = 0.639$; $a(r^2) = 0.114$; $a(Q^2) = -0.49$

(5)

where N is the number of compounds included in the model, R the correlation coefficient, s standard deviation of the regression, F the Fisher ratio, Q^2 the correlation coefficient of the cross-validation, AIC the Akaike Information Criterion and FIT the Kubinyi Function. The values of the descriptors are presented in Table 2.

Although this theoretical model has eight variables and acceptable statistical parameters, a step-by-step outlier extraction procedure led to different models with a better statistical profile. In this study, two outliers extracted represented a 4.76% of the whole data. Compounds **4** ((*Z*)-2-chloro-3-methyl-4-oxo-2-butenic acid) and **9** (Bromoacetic acid) present large residuals and should be considered as outliers. Compound **4** is structurally similar to compounds **1** (2,4,4-trichloro-3-(dichloromethyl)-2-butenic acid), **2** ((*Z*)-2,4,4-trichloro-3-formyl-2-butenic acid), **3** ((*Z*)-2,4-dichloro-3-formyl-2-butenic acid), and **5** ((*E*)-2,4,4-trichloro-3-(chloromethyl)-2-butenic acid) of the training set. Compound **4** has a main action mechanism throughout adenine adduct⁶⁶ and for that reason it has a low value of the activity respect its peer butenoic acids which have a mechanism of action through guanosine adduct. Even though it is known that AT sites are the primary targets in studied strains TA98 or TA100, only compounds with mechanism of action throughout guanosine adducts (GC sites) are detected by a TA100 strain mutagenicity Ames test. This fact explains the observed differences between both. The other outlier, compound **9**, and, like Iodoacetic acid but in minor amount, could induce its genotoxic damage via an oxidative stress mechanism⁶⁷ unlike the rest of brominated derivatives present in the training set. For the same reason the Iodoacetic acid must be an outlier but together with (*S*)-3-iodo-1,2-propanediol is the unique Iodine derivatives presents in this training set. On removal of these compounds from the training set, the next equation is obtained:

$$\begin{aligned} \log TA100 = & 8.95 \cdot 10^{-11} (\pm 1.26 \cdot 10^{-11}) \mu_{15}^{\text{Std}} \\ & - 5.77 \cdot 10^{-2} (\pm 9.42 \cdot 10^{-3}) \mu_1^{\text{Pols}} \\ & + 5.87 \cdot 10^{-8} (\pm 8.79 \cdot 10^{-9}) \mu_7^{\text{Mol}} \\ & - 8.57 \cdot 10^{-10} (\pm 1.03 \cdot 10^{-10}) \mu_0 \mu_{13}^{\text{Dip}} \\ & + 2.30 \cdot 10^{-8} (\pm 3.61 \cdot 10^{-9}) \mu_0 \mu_5^{\text{Pols}} \\ & + 1.75 \cdot 10^{-2} (\pm 1.33 \cdot 10^{-3}) \mu_1 \mu_2^{\text{Dip}^2} \\ & - 4.07 \cdot 10^{-10} (\pm 8.39 \cdot 10^{-11}) \mu_1 \mu_{14}^{\text{Dip}^2} \\ & - 2.26 \cdot 10^{-8} (\pm 3.77 \cdot 10^{-9}) \mu_1 \mu_9^{\text{Pol}} \\ & - 1.24 (\pm 0.23) \end{aligned}$$

$N = 40$; $R^2 = 0.902$; $Q_{(\text{CV-LOO})}^2 = 0.842$;
 $s = 0.548$; $F = 35.730$; $\text{AIC} = 0.727$; $\text{FIT} = 1.529$
 $Q_{\text{boot}}^2 = 0.718$; $a(r^2) = 0.125$; $a(Q^2) = -0.392$

(6)

We detected high correlation coefficients among the descriptor values of model (Eq. 6). Table 3 shows that some of regression coefficients were higher than 0.70, showing that they were closely correlated. Therefore, orthogonalization of the molecular descriptors was conducted.

Orthogonalization of molecular descriptors was undertaken to avoid collinearity among variables and model overfitting. Collinearity of variables should be as low as possible because interrelatedness among different descriptors can result in highly unstable models. The QSAR model obtained with the spectral moments (Eq. 7) after orthogonalization is given below, together with the statistical parameters of regression analysis.

$$\begin{aligned} \log TA100 = & 6.52 \cdot 10^{-3} (\pm 5.64 \cdot 10^{-4}) \Omega \mu_1 \mu_2^{\text{Dip}^2} \\ & - 1.80 \cdot 10^{-10} (\pm 2.42 \cdot 10^{-11})^2 \Omega \mu_0 \mu_{13}^{\text{Dip}} \\ & + 1.34 \cdot 10^{-11} (\pm 6.35 \cdot 10^{-12})^3 \Omega \mu_{15}^{\text{Std}} \\ & + 7.09 \cdot 10^{-9} (\pm 2.22 \cdot 10^{-9})^4 \Omega \mu_7^{\text{Mol}} \\ & + 4.81 \cdot 10^{-9} (\pm 1.62 \cdot 10^{-9})^5 \Omega \mu_0 \mu_5^{\text{Pols}} \\ & - 4.50 \cdot 10^{-2} (\pm 9.05 \cdot 10^{-3})^6 \Omega \mu_1^{\text{Pols}} \\ & - 1.85 \cdot 10^{-8} (\pm 3.67 \cdot 10^{-9})^7 \Omega \mu_1 \mu_9^{\text{Pol}} \\ & - 4.07 \cdot 10^{-10} (\pm 8.39 \cdot 10^{-11})^8 \Omega \mu_1 \mu_{14}^{\text{Dip}^2} \\ & - 1.91 (\pm 0.13) \end{aligned}$$

$N = 40$; $R^2 = 0.902$; $Q_{(\text{CV-LOO})}^2 = 0.842$;
 $s = 0.548$; $F = 35.730$; $\text{AIC} = 0.727$; $\text{FIT} = 1.529$
 $Q_{\text{boot}}^2 = 0.718$; $a(r^2) = 0.125$; $a(Q^2) = -0.392$

(7)

Once the non-desirable collinearity problems among the descriptors were eliminated the model obtained with the TOPS-MODE descriptors was compared with other families.

3.2. Comparison with other descriptors

The spectral moments were compared with other methodologies such as Geometrical,²⁰ RDF,⁶⁸ WHIM,²⁰

Table 2. Values of the spectral moments used in the model

Compounds	Name	Log TA100	μ_{15}^{Std}	μ_1^{Pois}	μ_7^{Mol}	$\mu_{10}\mu_{13}^{\text{Dip}}$	$\mu_0\mu_5^{\text{Pois}}$	$\mu_1\mu_2^{\text{Dip}^2}$	$\mu_1\mu_{14}^{\text{Dip}^2}$	$\mu_1\mu_5^{\text{Pol}}$
1	2,4,4-Trichloro-3-(dichloromethyl)-2-butenic acid	2.821	32399259648	57.53	5319063.5	3938757888	120429368	638.02	8035869184	101569274
2	(Z)-2,4,4-Trichloro-3-formyl-2-butenic acid	4.071	17575616512	74.60	3180590.25	2981445888	131358560	559.84	4532435968	49024272
3	(Z)-2,4-Dichloro-3-formyl-2-butenic acid	2.848	10623868928	74.60	2456188	2577081600	131358560	458.56	2109218685	31900258
4	(Z)-2-Chloro-3-methyl-4-oxo-2-butenic acid	-0.320	7693142016	74.60	1102993.125	2400710656	131358560	363.63	1201073250	29644918
5	(E)-2,4,4-Trichloro-3-(chloromethyl)-2-butenic acid	3.466	24558442496	57.53	4594256	3502516480	120429368	528.86	4903678976	79299192
6	(S)-2,3-Dibromopropanal	-0.522	29399392256	17.07	19273382	3796207104	13531792	208.51	3007512320	71086728
7	Dichloroacetic acid	-1.830	9651770368	57.53	2302502	1081926919	60212696	184.65	3086136064	21476786
8	Chloroacetic acid	-1.943	3874263808	57.53	1577735.875	832599616	60212696	129.84	1112014735	11595926
9	Bromoacetic acid	0.364	4498661888	57.53	10937820	795104960	60212696	127.76	1063220478	22664120
10	2-Bromopropane	-1.717	28033912832	0.00	8497154	2682537216	10800	53.98	609960128	95866400
11	2,3-Dichloro-1-propene	-0.007	5745217024	0.00	2748616.5	870004416	3440	83.27	362619648	30664026
12	(R)-1,2-Dichloropropane	-2.473	33391790080	0.00	2751213.25	3755125504	10800	118.55	1896276782	68256624
13	2,2-Dichloroacetyl Chloride	-1.212	10644528128	17.07	2722970.75	1619424644	9018675	205.95	4067141888	25971036
14	1,3-Dichloro-2-propanol	-1.838	39579848704	40.46	3183758.75	4337495552	77919208	155.83	2450607360	49380968
15	(S)-2,3-Dibromo-1-propanol	-0.334	45910818816	40.46	19330196	4466391040	77915872	150.53	2520851200	89414640
16	1,2-Dibromoethane	-0.608	12928412672	0.00	21928546	1119365902	4620	79.61	941540928	64058428
17	(R)-2-(Chloromethyl)oxirane	-0.693	61269241856	9.23	1899091.5	5696471552	143557.9219	118.07	2917894912	31245614
18	2-Chloroacetaldehyde	-0.789	2879185920	17.07	1702408.375	847295168	9018675	100.31	706531968	11804425
19	2-Chloroethanol	-2.074	10087707648	40.46	1808295.375	1123018729	56662730	62.17	578029184	20640296
20	3-Chloro-1-propene	-2.326	3641498880	0.00	2375211.25	625260096	3440	38.46	126195640	24974387
21	(E)-1,4-Dichloro-2-butene	0.179	7802283520	0.00	3720415.5	1441410646	8360	118.55	519270304	46837468
22	(2S,3S,4S,5S)-1,6-Dibromohexane-2,3,4,5-tetraol	-1.607	1.49E+11	161.84	22287698	24518873088	651642432	485.00	9049610240	198085360
23	(R)-2-(Fluoromethyl)oxirane	-1.092	58288287744	9.23	601365.25	5640154112	143557.9219	114.14	2778343680	16835970
24	2-Bromoethanol	-2.076	11155117056	40.46	11206682	1071490027	56662730	60.43	544268672	34284384
25	(E)-1,3-Dichloro-1-propene	0.011	4120854016	0.00	2833947.5	663904256	3440	83.27	277559616	24379362
26	(S)-3-Iodo-1,2-propanediol	-1.261	42241433600	80.92	187263327	3959013376	169988944	57.37	561084736	512510048
27	2-Chloro-2-nitropropane	-1.875	56509394944	43.14	1141175.375	5407563264	123703072	85.18	1384615900	76588288
28	(R)-1-Chloro-1-nitropropane	-0.679	33910702080	43.14	1291310.5	3943764992	123702896	85.18	1031692288	53157732
29	(S)-2,3-Dichloro-1-propanol	-0.747	40742600704	40.46	2827282.5	4724036096	77915872	155.83	2710674432	50294844
30	3-Bromo-1-propanol	-1.694	25292779520	40.46	11172864	2804492544	77915872	85.87	879171264	55146760
31	Dibromoacetic acid	-1.203	12629614592	57.53	18618924	1466588211	60212696	179.88	2826977536	55349956
32	2-Chloroethyl acrylate	-1.215	12221314048	26.30	2575745.75	2902270976	21329620	278.79	1935080897	47216572
33	(S)-2-(2,2,2-Trichloroethyl)oxirane	-1.194	1.11E+11	9.23	3150522.5	9492755456	192085.2969	333.86	12412784640	99825096
34	(S)-2-(Trichloromethyl)oxirane	0.455	1.03E+11	9.23	3065597	6200264704	143557.9219	264.14	12681416704	69385176
35	(R)-2-(Bromomethyl)oxirane	-0.371	63128502272	9.23	11247736	5607643648	143557.9219	115.61	2829678592	45953544
36	3-Chloro-N,N-dimethyl-1-propanamine	-2.400	39723761664	3.24	3492382.5	5760996864	52503.07813	193.93	1979739511	120393544
37	Bromochloroacetic acid	-1.207	11056564224	57.53	10061060	1537919671	60212696	182.26	2953888512	35974208
38	(S)-2,3-Dibromopropyl acrylate	-1.551	48278032384	26.30	20117806	8053475840	25907394	443.37	5407938560	149360656
39	(R)-1-Bromo-2-propanol	-1.793	33050613760	40.46	11188269	3278631680	77919208	85.87	1019503675	69423184
40	2-Bromoacrylaldehyde	-0.208	462832512	17.07	9807250	305336992	10519860	98.47	200182816	18189858
41	3-Bromo-3-buten-2-one	-0.886	3105279488	17.07	9815323	1785719556	15034232	144.31	553114304	41683616
42	Iodoacetic acid	0.776	5343832576	57.53	187019040	718894052	60212696	81.38	395034432	352284288

Table 3. Correlation matrix of the eight variables of the model

	μ_{15}^{Std}	μ_1^{Pols}	μ_7^{Mol}	$\mu_0\mu_{13}^{\text{Dip}}$	$\mu_0\mu_5^{\text{Pols}}$	$\mu_1\mu_2^{\text{Dip2}}$	$\mu_1\mu_{14}^{\text{Dip2}}$	$\mu_1\mu_9^{\text{Pol}}$
μ_{15}^{Std}	1.00	0.28	-0.02	0.90	0.51	0.31	0.75	0.25
μ_1^{Pols}		1.00	0.29	0.50	0.89	0.47	0.23	0.39
μ_7^{Mol}			1.00	-0.02	0.16	-0.16	-0.16	0.91
$\mu_0\mu_{13}^{\text{Dip}}$				1.00	0.75	0.43	0.62	0.27
$\mu_0\mu_5^{\text{Pols}}$					1.00	0.43	0.29	0.36
$\mu_1\mu_2^{\text{Dip2}}$						1.00	0.66	0.04
$\mu_1\mu_{14}^{\text{Dip2}}$							1.00	0.06
$\mu_1\mu_9^{\text{Pol}}$								1.00

eigenvalue-based indices,⁶⁹ information,^{70–73} and 2D-autocorrelation descriptors^{74–76} (Table 4).

The models employing the above mentioned descriptors were developed using the same data set excluding outliers (40 compounds in total) and a maximum of eight variables for keeping the ratio between cases and variables greater to 5.^{24,39–41,63,77} The comparisons were done based on regression analysis results, the predictive capability of the generated models.

There were substantial differences in the explanations of the experimental variance given by these models, compared with the TOPS-MODE model system. Thus, while the TOPS-MODE model was able to explain 90% of mutagenic potency, the other models could only explain 82.8% of such variance, at best. This implies that the predictive capability of the TOPS-MODE model is better than the other five models, not only in the statistical parameters of regression but also, and more importantly, in terms of stability for inclusion/exclusion of chemicals, as measured by the determination coefficient (Q^2). Thus, the value of the determination coefficient for leave-one-out cross-validation for the model obtained with the spectral moment ($Q^2 = 0.842$) was the highest of all. Moreover the spectral moments possess the best goodness-of-fit if we observe the values of F , s , AIC, and FIT.

3.3. Structural alerts identification (fragments contribution)

As it is shown in Table 5, the variables weighted with bond dipole moment explain the 66.9% of the descriptor

values for specific data set of chemicals used in the analysis. The variables weighted with polar surface, polarizability, molar refractivity, and standard bond distance accounted for 10.6%, 8.0%, 3.3%, and 1.4% of the variance respectively. Thus, bond dipole moment was a key molecular driver of this mutagenicity. The lack of hydrophobicity features agrees with some author regarding the mutagenic direct-action mechanism.⁷⁸

One advantage of the present approach for QSTR and QSAR studies is that it can provide information explaining how structural features of molecules can account for their biological activities. Thus it is possible to detect fragments that contribute positively or negatively to a particular biological activity and their effects can be interpreted in terms of physicochemical properties.⁷⁹ The structural fragments identified in the present study are shown in Figure 1, and their contributions to mutagenic potency are in Table 6.

Table 5. Contribution of spectral moments to the model

Variables	Global R^2 (step by step)	Contribution to global R^2 for every variable
$\Omega\mu_1\mu_2^{\text{Dip2}}$	0.421	0.421
$\Omega^2\mu_0\mu_{13}^{\text{Dip}}$	0.595	0.174
$\Omega^7\mu_1\mu_9^{\text{Pol}}$	0.675	0.08
$\Omega^6\mu_1^{\text{Pols}}$	0.753	0.078
$\Omega^8\mu_1\mu_{14}^{\text{Dip2}}$	0.827	0.074
$\Omega^4\mu_7^{\text{Mol}}$	0.86	0.033
$\Omega^5\mu_0\mu_5^{\text{Pols}}$	0.888	0.028
$\Omega^3\mu_{15}^{\text{Std}}$	0.902	0.014

Table 4. The statistical parameters of the linear regression models with 40 compounds (except outliers) obtained for the six kinds of descriptors involved in the comparison

Descriptors	Variables	R^2	F	p	s	Q^2	AIC	FIT	Q_{boot}^2	$a(R^2)$	$a(Q^2)$
Spectral moments	$\mu_{15}^{\text{Std}}, \mu_1^{\text{Pols}}, \mu_7^{\text{Mol}}, \mu_0\mu_{13}^{\text{Dip}}, \mu_0\mu_5^{\text{Pols}}, \mu_1\mu_2^{\text{Dip2}}, \mu_1\mu_{14}^{\text{Dip2}}, \mu_1\mu_9^{\text{Pol}}$	0.902	35.730	$<10^{-5}$	0.548	0.842	0.727	1.529	0.718	0.125	-0.392
RDF	RDF055v, RDF010u, RDF030v, RDF055p, RDF025u, RDF025p, RDF020e, RDF035u	0.812	16.790	$<10^{-5}$	0.754	0.682	0.838	1.287	0.175	0.155	-0.489
Geometrical	G(O..Cl), FDI, HOMT, G(O..I), SPH, SPAN, G2, ASP	0.828	18.731	$<10^{-5}$	0.726	0.659	0.766	1.435	0.442	0.16	-0.478
Eigenvalue-based indices	SEigZ, SEige, AEigp, VRA1, LPI, SEigv, AEigm, Eigle	0.785	14.211	$<10^{-5}$	0.811	0.647	0.958	1.088	0.505	0.143	-0.555
WHIM	L2s, L3u, E3m, G2e, G3u, E3e, Dp, G3s	0.778	13.649	$<10^{-5}$	0.824	0.574	0.989	1.045	0.293	0.158	-0.522
Information	IC1, IDDE, IC2, CIC4, HVcpx, Yindex, Uindex, TIC5	0.810	16.617	$<10^{-5}$	0.762	0.562	0.845	1.271	0.430	0.167	-0.553
2D-autocorrelation	GATS5v, ATS6e, MATS5e, ATS1m, GATS2e, ATS2m, GATS4m, MATS5p	0.798	15.334	$<10^{-5}$	0.787	0.390	0.902	1.178	0.222	0.15	-0.648

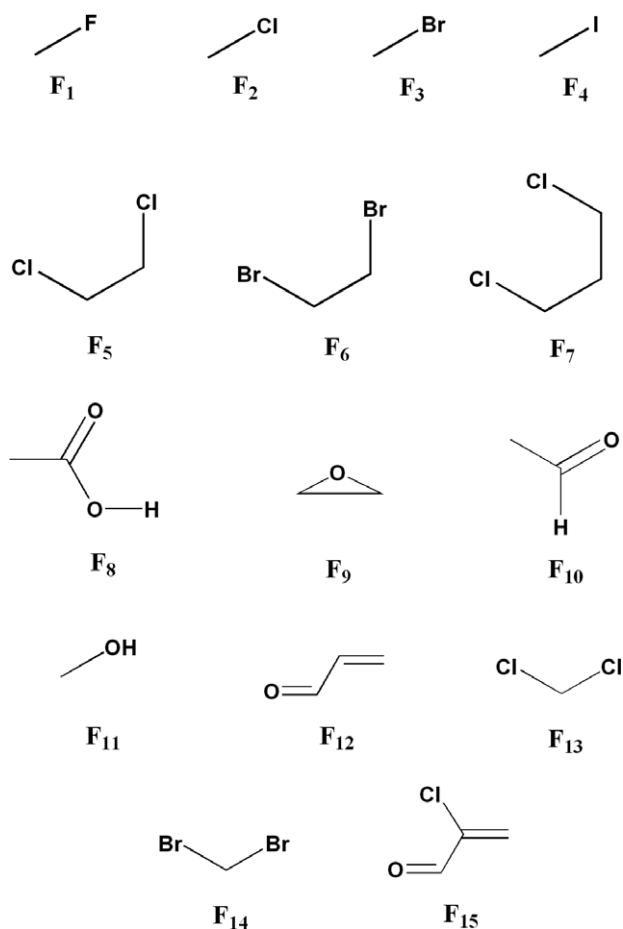


Figure 1. Structural of selected fragment for which their contribution to the mutagenic potency was calculated.

Table 6. The contributions of different structural fragments to the mutagenic activity

Fragments	Contributions
F_1	0.14
F_2	0.45
F_3	0.69
F_4	2.20
F_5	0.90
F_6	1.60
F_7	2.15
F_8	-0.63
F_9	0.67
F_{10}	0.50
F_{11}	-1.09
F_{12}	1.07
F_{13}	0.95
F_{14}	1.47
F_{15}	2.20

The results obtained for fragments F_1 , F_2 , F_3 , and F_4 shows that the mutagenicity order follows the following rule: $I > Br > Cl > F$. This sequence order is in agreement with the halide order of reactivity and its potential as a leaving group. Some examples of these compounds are in Figure 2. The chemical reactivity of monohaloacetic acids is expected to be similar to that of alkyl halides. The reactivity of methyl halides is primarily dependent

on the carbon–halogen bond dissociation energy which is related to the bond strength. Since the bond dissociation energy of the halogen follows the order $I < Br < Cl$, the strength of carbon–halogen bond increases accordingly. Polarizability and delocalization of the electron could also contribute to making iodine a better leaving group than bromine and much better leaving group than chlorine.¹⁵ This positive contribution of the presence of the halogen substituent in the mutagenicity was confirmed previously by González et al.⁸⁰

The dihalogenated (Fig. 3), if it is vicinal, increase the mutagenicity since it can act, either directly or after Glutathione conjugation to form the episulfonium ion (powerful electrophile) as cross-linking agents^{81,82} (Fig. 4), which we can see through the values of the fragments F_{13} , F_{14} , F_6 , and F_7 .

Moreover, various other fragments, present in the molecules under study, like epoxide and carbonyl group, agree with the values reported by other authors confirming the known mechanism. Epoxides, aldehydes and carbonyl groups α,β -unsaturated increase this kind of mutagenicity, if we look at the positive values of fragments F_9 , F_{10} , and F_{12} (Fig. 5). The epoxides and aldehydes are potential alkylating agents and specially short-chain aldehydes.^{83–85} Carbonyl groups α,β -unsaturated have an initial mechanism-type Michael addition⁸⁶ (Fig. 6). The chlorine atom presence in position 2 increases the mutagenicity⁸⁷ like we can see in fragments F_{12} and F_{15} (Fig. 5), due to the cross-linking potential with another DNA or protein nucleophilic center.⁸⁸

Carboxylic acid and alcohol groups reduce the mutagenicity in the light of the negative values for the fragments F_8 and F_{11} (Fig. 7), since both acids and aliphatic alcohols, the latest one with low numbers of carbon atoms, negative results were in this type of assay.^{89–93}

3.4. Applicability domain

The prime overall goal of QSAR research is to develop models that provide accurate predictions for as many chemicals as possible in the universe, particularly for those that have not been tested or for which reliable experimental data are still not available, as well as the properly assessed safety of new chemicals. We defined the applicability domain determining the leverage values for each compound. As seen in Figure 8, the majority of compounds of the training set are inside of this area, however, three halogenated compounds have a leverage greater than h^* , but show standard deviation values within the limit, which implies that they are not to be considered outliers but influential chemicals.⁶³

4. Prediction for haloacetic acids

The predictions for the rest of haloacetic acids obtaining with this model are shown in Table 7.

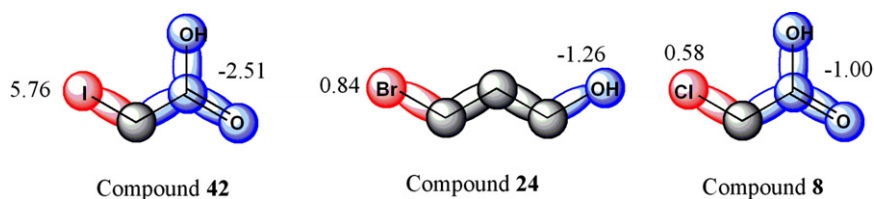


Figure 2. Contribution of the halogens (the fragments in blue have a negative contribution and the fragments in red have a positive contribution).

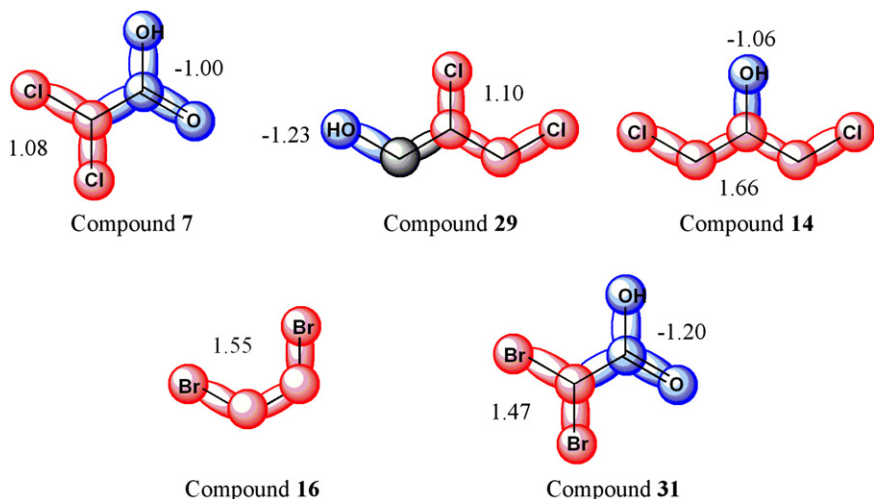


Figure 3. Fragment contribution of some dihalogenated derivatives.

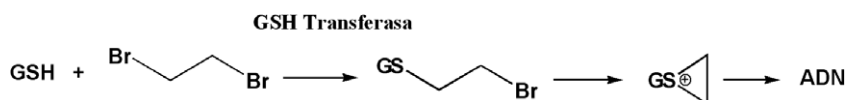


Figure 4. Dihalogenated activation mechanism through glutathione-S-transferase.

The compounds marked are the substances which are outside of the applicability domain and then its predictions are not valid. This model predicts a mutagenic activity for most of the dihaloacetic acids greater than

trihaloacetic acids. This fact is in agreement with the halogen atom-leaving tendency. This tendency tends to decrease when the halogenation degree increases. The electron-withdrawing effect of the second or third halogen diminishes the leaving potential of the first halogen.⁸² The values of the three haloacetic acids: TCAA (TriChloroacetic acid), BDCAA (Bromodichloroacetic acid), and TBAA (Tribromoacetic acid) have an experimental negative value. The model predicts very low values. These results are closer to the experimental results, especially for BDCAA and TBAA. These results can be explained because the model is aimed primarily by the Dipole moments which decrease with increasing the degree of halogenation (Table 8).

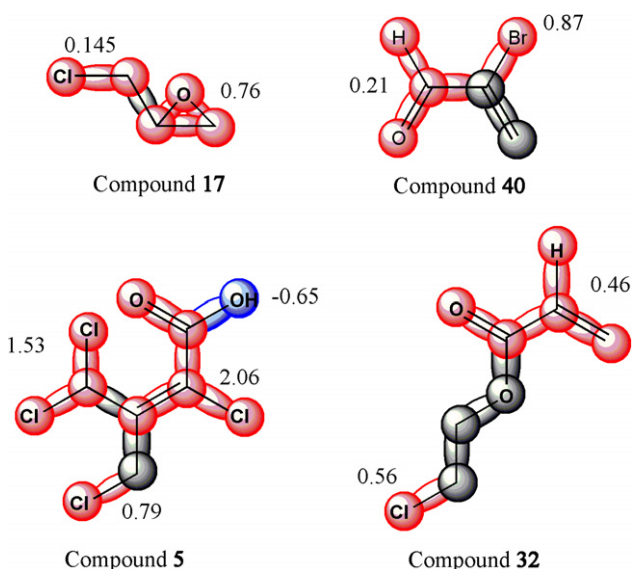


Figure 5. Fragment contribution of some epoxide, aldehyde and carbonyl α,β -unsaturated.

According to the results, we can say that the fluoroiodoacetic (FIAA) and difluoroiodoacetic (DFIAA) acid could show mutagenicity and these substances could show mutagenicity. These compounds can be found most probably in fluorinated waters with a high content of bromide (and iodide).⁹⁴

5. Conclusions

In this paper, we modeled the mutagenicity activity in *S. typhimurium* Ames test TA100 strain without metabolic

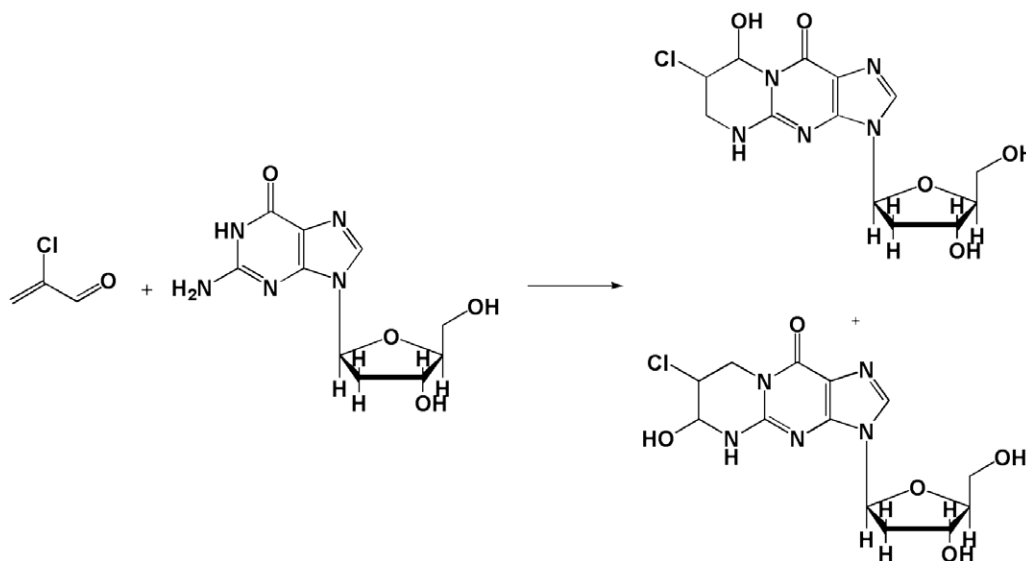


Figure 6. Addition Michael type mechanism for the carbonyl group α,β -unsaturated with chlorine in position 2 with deoxyguanosine.

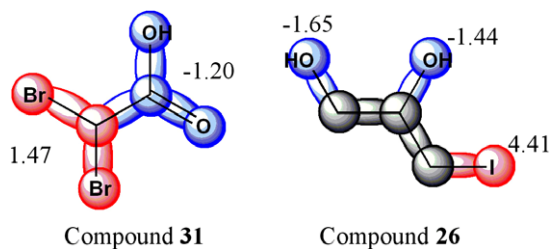


Figure 7. Fragment contribution of some hydroxyl and acid moieties.

activation to predict such property for haloacetic acids (HAA). For this purpose, we employed the spectral moments and the variables that were found to be most sig-

nificant to build the model. These variables were basically dipole moments and polar surface, polarizability, molar refractivity, and standard bond distance.

The spectral moments with multiplicative effects had better results than other linear descriptors such as Geometrical, RDF, WHIM, eigenvalue-based indices, 2D-autocorrelation, and information indices, taking into account the statistical parameters of the model and the cross-validation results.

The structural alerts and the mutagenicity predicted values from the model output are in agreement with references from other authors. This model predicts mutagenicity for fluoroiodoacetic and difluoroiodoacetic

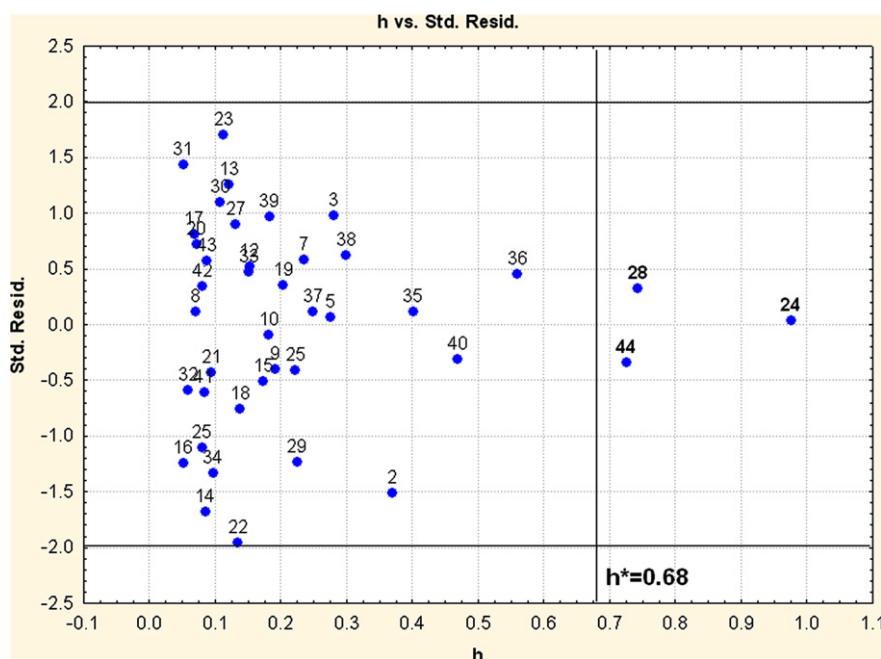


Figure 8. Applicability domain of the model of Eq. 7.

Table 7. Prediction for the haloacetic acids

Compounds	Log TA100 _{Pred.}	<i>h</i> , Leverage	Mutagenicity ^a
Fluoroiodoacetic acid (FIAA)	1.698	0.447	ND
Difluoroiodoacetic acid (DFIAA)	1.008	0.390	ND
Chloroiodoacetic acid (CIAA)	−0.557	0.281	ND
Chlorofluoroiodoacetic acid (CFIAA)	−0.832	0.235	ND
Bromoiodoacetic acid (BIAA)	−1.326	0.371	ND
Bromofluoroiodoacetic acid (BFIAA)	−1.342	0.303	ND
Bromofluoroacetic acid (BFAA)	−1.527	0.164	ND
Trichloroacetic acid (TCAA)	−1.806	0.385	− ^b
Fluoroacetic acid (FAA)	−1.856	0.158	ND
Chlorofluoroacetic acid (CFAA)	−1.951	0.186	ND
Difluoroacetic acid (DFAA)	−1.972	0.191	ND
Difluoroacetic acid (DBFAA)	−2.401	0.336	ND
Bromodichloroacetic acid (BDCAA)	−2.410	0.368	− ^b
Dichlorofluoroacetic acid (DCFCAA)	−2.489	0.396	ND
Bromodifluoroacetic acid (BDFCAA)	−2.713	0.395	ND
Dichloroiodoacetic acid (DCIAA)	−2.714	0.555	ND
Tribromoacetic acid (TBAA)	−2.756	0.341	− ^b
Bromochlorofluoroacetic acid (BCFCAA)	−2.844	0.392	ND
Dibromochloroacetic acid (DBCAA)	−2.949	0.364	ND
Chlorodifluoroacetic acid (CDFCAA)	−3.254	0.445	ND
Trifluoroacetic acid (TFAA)	−3.274	0.441	ND
Bromochloroiodoacetic acid (BCIAA)	−4.402	0.695	ND
Dibromoiodoacetic acid (DBIAA)	−5.554	0.798	ND
Fluorodiiodoacetic acid (FDIAA)	−8.450	0.907	ND
Diiodoacetic acid (DIAA)	−9.656	0.926	ND
Chlorodiiodoacetic acid (CDIAA)	−15.668	0.967	ND
Bromodiiodoacetic acid (BDIAA)	−18.840	0.977	ND
Triiodoacetic acid (TIAA)	−45.679	0.996	ND

^a Mutagenicity values in *S. typhimurium* TA100 strain without metabolic activation extracts of the MDL Toxicity Database. ND no data.

^b −, Negative assay.

Table 8. Dipole moments of the haloacetic acids

Compounds	Dipole moment ^a
FAA	3.230
DFAA	0.144
TFAA	1.745
CAA	2.716
DCAA	1.284
TCAA	1.564
BAA	2.722
DBAA	1.048
TBAA	1.552
IAA	1.794
DIAA	2.163
TIAA	1.445

^a Dipole moment calculated with the Vamp module of Tsar 3.3 software.

tic acids and these compounds can be found most probably in fluorinated waters high in bromide (and iodide).

Acknowledgment

The authors acknowledge to MODESLAB 1.0 software owners for delivering a free copy of such program.

References and notes

- Akin, E. W.; Hoff, J. C.; Lippy, E. C. *Environ. Health Perspect.* **1982**, *46*, 7.
- Wilcox, P.; Williamson, S. *Environ. Health Perspect.* **1986**, *69*, 141.
- Morris, R. D.; Audet, A. M.; Angelillo, I. F.; Chalmers, T. C.; Mosteller, F. *Am. J. Public Health* **1992**, *82*, 955.
- Koivusalo, M.; Jaakkola, J. J.; Vartiainen, T.; Hakulinen, T.; Karjalainen, S.; Pukkala, E.; Tuomisto, J. *Am. J. Public Health* **1994**, *84*, 1223.
- Bull, R. J.; Birnbaum, L. S.; Cantor, K. P.; Rose, J. B.; Butterworth, B. E.; Pegram, R.; Tuomisto, J. *Fundam. Appl. Toxicol.* **1995**, *28*, 155.
- Nieuwenhuijsen, M. J.; Toledano, M. B.; Elliott, P. *J. Exposure Anal. Environ. Epidemiol.* **2000**, *10*, 586.
- Bull, R. J.; Sánchez, I. M.; Nelson, M. A.; Larson, J. L.; Lansing, A. J. *Toxicology* **1990**, *63*, 341–359.
- DeAngelo, A. B.; Daniel, F. B.; Stober, J. A.; Olson, G. R. *Fundam. Appl. Toxicol.* **1991**, *16*, 337–347.
- DeAngelo, A. B.; Daniel, F. B.; Most, B. M.; Olson, G. R. *Toxicology* **1996**, *114*, 207–221.
- Herbert, V.; Gardner, A.; Colman, N. *Am. J. Clin. Nutr.* **1980**, *33*, 1179–1182.
- Nestman, E. R.; Chu, I.; Kowbel, D. J.; Matula, T. I. *Can. J. Genet. Cytol.* **1980**, *22*, 35–40.
- DeMarini, D. M.; Perry, E.; Shelton, M. L. *Mutagenesis* **1994**, *9*, 429–437.
- Kargalioglu, Y.; McMillan, B. J.; Minear, R. A.; Plewa, M. J. *Teratog. Carcinog. Mutagen.* **2002**, *22*, 113–128.
- Kundu, B.; Richardson, S. D.; Swartz, P. D.; Matthews, P. P.; Richard, A. M.; DeMarini, D. M. *Mutat. Res.* **2004**, *562*, 39–65.
- Plewa, M. J.; Wagner, E. D.; Richardson, S. D.; Thruston, A. D. J.; Woo, Y. T.; McKague, A. B. *Environ. Sci. Technol.* **2004**, *38*, 4713–4722.
- McCann, J.; Choi, E.; Yamasaki, E.; Ames, B. N. *Proc. Natl. Acad. Sci. U.S.A.* **1975**, *72*, 5135–5139.

17. Sugimura, T.; Sato, S.; Nagao, M.; Yahagi, T.; Matsushima, T.; Seino, Y.; Takeuchi, M.; Kawachi, T. *Fundamental of Cancer Prevention*. In Magee, P., Takayama, S., Sugimura, T., Matsushima, T., Eds.; University Park Press: Baltimore, 1976.
18. Zeiger, E.; Haseman, J. K.; Shelby, M. D.; Margolin, B. H.; Tennant, R. W. *Environ. Mol. Mutagen.* **1990**, *16*, 1–14.
19. Venkatapathy, R.; Bruce, R.; Moudgal, C. Presented at the EPA Science Forum, Mandarin Oriental Hotel, Washington, DC; Available from: <<http://www.epa.gov/ord/scienceforum/2004/poster-ord-NtoZ.htm>>, 2004.
20. Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*; Wiley-VCH: Mannheim, 2000.
21. Saiz-Urra, L.; González, M. P.; Collado, I. G.; Hernandez-Galan, R. *J. Mol. Graphics Modell.* **2007**, *25*, 680–690.
22. Estrada, E.; Molina, E. *J. Mol. Graphics Modell.* **2006**, *25*, 275–288.
23. González, M. P.; Teran, C.; Teijeira, M. *Bioorg. Med. Chem. Lett.* **2006**, *16*, 1291–1296.
24. Morales, A. H.; González, M. P.; Rieumont, J. *Polymer* **2004**, *45*, 2045–2050.
25. Morales, A. H.; Perez, M. A. C.; González, M. P.; Ruiz, R. M.; Díaz, H. G. *Bioorg. Med. Chem. Lett.* **2005**, *13*, 2477–2488.
26. Morales, A. H.; Cabrera, M. A.; Combes, R. D.; González, M. P. *Toxicology* **2006**, *220*, 51.
27. Morales, A. H.; González, M. P.; Cordeiro, M. N. D. S.; Perez, M. A. C. *Toxicol. Appl. Pharmacol.* **2007**, *221*, 189–202.
28. González-Díaz, H.; Gia, O.; Uriarte, E.; Hernadez, I.; Ramos, R.; Chaviano, M.; Seijo, S.; Castillo, J. A.; Morales, L.; Santana, L.; Akpaloo, D.; Molina, E.; Cruz, M.; Torres, L. A.; Cabrera, M. A. *J. Mol. Model.* **2003**, *9*, 395–407.
29. González-Díaz, H.; Olazabal, E.; Castanedo, N.; Sánchez, I. H.; Morales, A.; Serrano, H. S.; González, J.; de Armas, R. R. *J. Mol. Model.* **2002**, *8*, 237–245.
30. González-Díaz, H.; Torres-Gomez, L. A.; Guevara, Y.; Almeida, M. S.; Molina, R.; Castanedo, N.; Santana, L.; Uriarte, E. *J. Mol. Model.* **2005**, *11*, 116–123.
31. González-Díaz, H.; Pérez-Castillo, Y.; Podda, G.; Uriarte, E. *J. Comput. Chem.* **2007**, *28*, 1990–1995.
32. González-Díaz, H.; Saiz-Urra, L.; Molina, R.; González-Díaz, Y.; Sánchez-González, A. *J. Comput. Chem.* **2007**, *28*, 1042–1048.
33. González-Díaz, H.; Bonet, I.; Terán, C.; De Clercq, E.; Bello, R.; Garcia, M. M.; Santana, L.; Uriarte, E. *Eur. J. Med. Chem.* **2007**, *42*, 580–585.
34. Prado-Prado, F. J.; González-Díaz, H.; Santana, L.; Uriarte, E. *Bioorgan. Med. Chem.* **2007**, *15*, 897–902.
35. Estrada, E. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 844–849.
36. Estrada, E. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 320–328.
37. Estrada, E. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 23–27.
38. González, M. P.; Díaz, H. G.; Ruiz, R. M.; Cabrera, M. A.; de Armas, R. R. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1192–1199.
39. Gonzalez, M. P.; Teran, C.; Fall, Y.; Diaz, L. C.; Helguera, A. M. *Polymer* **2005**, *46*, 2783.
40. González, M. P.; Helguera, A. M.; Molina, R.; García, J. R. *Polymer* **2004**, *45*, 2773.
41. González, M. P.; Dias, L.; Helguera, A. M. *Polymer* **2004**, *45*, 5353.
42. González, M. P.; Helguera, A. M.; Cabrera, M. A.; González, H. *Bioorg. Med. Chem.* **2005**, *13*, 1775.
43. González, M. P.; Díaz, H. G.; Cabrera, M. A.; Ruiz, R. M. *Bioorg. Med. Chem.* **2004**, *12*, 735.
44. Mortelmans, K.; Zeiger, E. *Mutat. Res.* **2000**, *455*, 29–60.
45. Bernstein, L.; Kaldor, J.; McCaan, J.; Pike, M. C. *Mutat. Res.* **1982**, *97*, 267–281.
46. Estrada, E. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 31–33.
47. Estrada, E.; Uriarte, E.; Gutierrez, Y.; González, H. *SAR QSAR Environ. Res.* **2003**, *14*, 145.
48. Gutierrez, Y.; Estrada, E. 'Modes Lab, version 1.0', 2002.
49. Weininger, D. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31–36.
50. Tetko, I. V.; Gasteiger, J.; Todeschini, R.; Mauri, A.; Livingstone, D.; Ertl, P.; Palyulin, V. A.; Radchenko, E. V.; Zefirov, N. S.; Makarenko, A. S.; Tanchuk, V. Y.; Prokopenko, V. V. *J. Comput. Aided Mol. Des.* **2005**, *19*, 453–463.
51. Draper, N. R.; Smith, H. *Applied Regression Analysis*, Second ed.; John Wiley and Sons: New York, 1981.
52. Garcia-Domenech, R.; Julian-Ortiz, J. V. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 445–449.
53. Kubinyi, H. *Quant. Struct.-Act. Relat.* **1994**, *13*, 285.
54. Kubinyi, H. *Quant. Struct.-Act. Relat.* **1994**, *13*, 393.
55. Akaike, H. Information theory and an extension of the maximum likelihood principle. In *Proceedings of the Second International Symposium on Information Theory*; Akademiai Kiado: Budapest, 1973.
56. Akaike, H. *IEEE Trans. Automat. Control* **1974**, *AC-19*, 716.
57. Todeschini, R.; Ballabio, D.; Consonni, V.; Mauri, A.; Pavan, M. 'Mobydigs Computer Software', 2004.
58. Lucic, B.; Nikolic, S.; Trinajstic, N.; Juric, D. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 532.
59. Klein, D.; Randi, M.; Babic, D.; Lucic, B.; Nikolic, S.; Trinajstic, N. *Int. J. Quantum Chem.* **1997**, *63*, 215.
60. Randic, M. *New J. Chem.* **1991**, *15*, 517–525.
61. Randic, M. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 311–320.
62. Randic, M. *J. Mol. Struct. (Theochem.)* **1991**, *233*, 45–59.
63. Eriksson, L.; Jaworska, J.; Worth, A. P.; Cronin, M. T. D.; McDowell, R. M.; Gramatica, P. *Environ. Health Perspect.* **2003**, *111*, 1361.
64. Gramatica, P. *QSAR Comb. Sci.* **2007**, *00*, 1–9.
65. Vighi, M.; Gramatica, P.; Consolaro, F.; Todeschini, R. *Ecotoxicol. Environ. Saf.* **2001**, *49*, 206–220.
66. Munter, T.; Kronberg, L.; Sjöholm, R. *Chem. Res. Toxicol.* **1996**, *9*, 703708.
67. Cemelli, E.; Wagner, E. D.; Anderson, D.; Richardson, S. D.; Plewa, M. J. *Environ. Sci. Technol.* **2006**, *40*, 1878–1883.
68. Hemmer, M. C.; Steinhauer, V.; Gasteiger, J. *J. Vib. Spectrosc.* **1999**, *19*, 151–164.
69. Balaban, A. T.; Ciubotariu, D.; Medeleanu, M. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 517–523.
70. Bonchev, D. *Information Theoretic Indices for Characterization of Chemical Structures*; RSP-Wiley: Chichester (UK), 1983.
71. Raychaudhury, C.; Ray, S. K.; Ghosh, J. J.; Roy, A. B.; Basak, S. C. *J. Comput. Chem.* **1984**, *5*, 581–588.
72. Balaban, A. T.; Balaban, T.-S. *J. Math. Chem.* **1991**, *8*, 383–397.
73. Magnuson, V. R.; Harriss, D. K.; Basak, S. C. Chemical applications of topology and graph theory. In King, R. B., Ed.; Elsevier: Amsterdam (The Netherlands), 1983.
74. Moran, P. A. P. *Biometrika* **1950**, *37*, 1723.
75. Moreau, G.; Broto, P. *Nouv. J. Chim.* **1980**, *4*, 359–360.
76. Moreau, G.; Broto, P. *Nouv. J. Chim.* **1980**, *4*, 757–764.
77. González, M. P.; Helguera, A. M.; González-Díaz, H. *Polymer* **2004**, *45*, 2073.
78. LaLonde, R. T.; Leo, H.; Perakyla, H.; Dence, C. W.; Farrell, R. P. *Chem. Res. Toxicol.* **1992**, *5*, 392.
79. Benigni, R.; Giuliani, A. *Bioinformatics* **2003**, *19*, 1194–1200.
80. González, M. P.; Dias, L. C.; Helguera, A. M. *Polymer* **2004**, *45*, 5353–5359.

81. Guengerich, F. P. *Jpn. J. Toxicol. Environ. Health* **1997**, *43*, 69–82.
82. Woo, Y.-T.; Lai, D.; Arcos, J. C.; Argus, M. F. *Chemical Induction of Cancer, Structural Bases and Biological Mechanism*. In *Aliphatic and Polyhalogenated Carcinogens*; Academic Press: Orlando, Florida, 1985; vol. IIIB.
83. Woo, Y. T.; Lai, D. Y.; McLain, J. L.; Ko Manibusan, M.; Dellarco, V. *Environ. Health Perspect.* **2002**, *110*, 75.
84. Simon, P.; Epe, B.; Mtzel, P.; Schiffmann, D.; Wild, D.; Ottenwilder, H.; Fedtke, N.; Bolt, H. M.; Henschler, D. *J. Biochem. Toxicol.* **1997**, *1*, 43–55.
85. Castelain, P. H.; Criado, B.; Cornet, M.; Laib, R.; Rogiers, V.; Kirsch-Volders, M. *Mutagenesis* **1993**, *8*, 387–393.
86. Eder, E.; Henschler, D.; Neudecker, T. *Xenobiotica* **1982**, *12*, 831–848.
87. Eder, E.; Weinfurtner, E. *Chemosphere* **1994**, *29*, 2455–2466.
88. Van Beerendonk, G. J. M.; Nivard, M. J. M.; Vogel, E. W.; Nelson, S. D.; Meerman, J. H. N. *Mutagenesis* **1992**, *7*, 19–24.
89. McGregor, D. B.; Cruzan, G.; Callander, R. D.; May, K.; Banton, M. *Mutat. Res.* **2005**, *565*, 181.
90. Stolzenberg, S. J.; Hine, C. H. *J. Toxicol. Environ. Health* **1979**, *5*, 1149–1158.
91. Simmon, V. F.; Kauhanen, K.; Tardiff, R. G. Progress in Genetic Toxicology. In *Chapter Mutagenic Activity of Chemicals Identified in Drinking Water*; Elsevier: North Holland Press, Amsterdam, 1977; pp 249–268.
92. Heck, J. D.; Vollmuth, T. A.; Cifone, M. A.; Jagannath, D. R.; Myhr, B.; Curren, R. D. *The Toxicologist* **1989**, *9*, 257.
93. Philipose, B.; Singh, R.; Khan, K. A.; Giri, A. K. *Mutat. Res.* **1997**, *393*, 123–131.
94. Krasner, S. W.; Weinberg, H. S.; Richardson, S.; Pastor, S. J.; Chinn, R.; Scimmenti, M. J.; Onstad, G. D.; Thruston, A. D. *Environ. Sci. Technol.* **2006**, *40*, 7175–7185.
95. Franzen, R.; Goto, S.; Tanabe, K.; Morita, M. *Mutat. Res.* **1998**, *417*, 31–37.
96. Gordon, W. P.; Soederlund, E. J.; Holme, J. A.; Nelson, S. D.; Iyer, L.; Rivedal, E.; Dybing, E. *Carcinogenesis* **1985**, *6*, 705–709.
97. Omichinski, J. P.; Soederlund, E. J.; Bausano, J. A.; Dybing, E.; Nelson, S. D. *Mutagenesis* **1987**, *2*, 287–292.



Contents lists available at ScienceDirect

Bioorganic & Medicinal Chemistry

journal homepage: www.elsevier.com/locate/bmc

Convenient QSAR model for predicting the complexation of structurally diverse compounds with β -cyclodextrins

Alfonso Pérez-Garrido^{a,b,*}, Aliuska Morales Helguera^{c,d,e}, Adela Abellán Guillén^b
M. Natália D. S. Cordeiro^e, Amalio Garrido Escudero^a

^aEnvironmental Engineering and Toxicology Dpt., Catholic University of San Antonio, Murcia, C.P., Guadalupe 30107, Spain

^bDepartment of Food and Nutrition Technology, Catholic University of San Antonio, Murcia, C.P., Guadalupe 30107, Spain

^cDepartment of Chemistry, Faculty of Chemistry and Pharmacy, Central University of Las Villas, Santa Clara, 54830 Villa Clara, Cuba

^dMolecular Simulation and Drug Design Group, Chemical Bioactive Center, Central University of Las Villas, Santa Clara, 54830 Villa Clara, Cuba

^eREQUIMTE, Chemistry Department, Faculty of Sciences, University of Porto, 4169-007 Porto, Portugal

ARTICLE INFO

Article history:

Received 29 July 2008

Revised 4 November 2008

Accepted 12 November 2008

Available online 24 November 2008

Keywords:

QSAR

Topological descriptors

 β -Cyclodextrins

Complex stability constant

ABSTRACT

This paper reports a QSAR study for predicting the complexation of a large and heterogeneous variety of substances (233 organic compounds) with β -cyclodextrins (β -CDs). Several different theoretical molecular descriptors, calculated solely from the molecular structure of the compounds under investigation, and an efficient variable selection procedure, like the Genetic Algorithm, led to models with satisfactory global accuracy and predictivity. But the best-final QSAR model is based on Topological descriptors meanwhile offering a reasonable interpretation. This QSAR model was able to explain ca. 84% of the variance in the experimental activity, and displayed very good internal cross-validation statistics and predictivity on external data. It shows that the driving forces for CD complexation are mainly hydrophobic and steric (van der Waals) interactions. Thus, the results of our study provide a valuable tool for future screening and priority testing of β -CDs guest molecules.

© 2008 Elsevier Ltd. All rights reserved.

1. Introduction

Cyclodextrins (CDs) are cyclic oligomers composed of either six (α -cyclodextrin), seven (β -cyclodextrin), eight (γ -cyclodextrin), or more α -D-glucopyranose units linked in a toroidal structure by α -(1-4)glycosidic bonds (Fig. 1). Their overall molecular shape is normally portrayed in terms of a truncated cone with primary and secondary hydroxyl groups crowning the narrower rim and wider rim, respectively¹ (Fig. 1). CDs are among the most frequently used host molecules in a wide range of applications in industrial, pharmaceutical, agricultural, and other fields, including improving the solubility and stability of drugs and biopharmaceuticals, and selectively binding materials that fit into the central hole in affinity purification and chromatography methods.^{2,3} Experimental determination of the complex binding constant is often difficult and time consuming because of the low solubility of the guest molecules in aqueous solution. For instance, 10 days were required for gathering data related to the equilibrium system of digitoxin and β -cyclodextrin.⁴ The employment of QSAR/QSPR methodology allows cost savings by reducing the laboratory resources needed, and the time required to create and investigate new compounds with better complexing pro-

file. For this reason, QSAR/QSPR is a useful alternative tool in the research of novel compounds.

Computer-based methods have already been applied as tools for predicting CD binding constants and for studying the driving forces involved in the encapsulation phenomena. These applications have been excellently reviewed by Lipkowitz in the late nineties.⁵ Molecular modeling using quantum mechanics calculations, Monte Carlo or molecular dynamics simulations, etc., group-contribution models; quantitative-structure-activity/property relationship (QSAR/QSPR) techniques based on 2D, 3D molecular descriptors and on comparative molecular field analysis; statistical analysis tools; and artificial neural networks have whole been used to predict the thermodynamic stability of CDs inclusion complexes and to elucidate the most important factors influencing the host-guest interactions.^{6–15}

The general picture that emerges from the joint analysis of the large body of available experimental and theoretical work reveals that there are five major interactions between CD-hosts and guest molecules, namely (i) hydrophobic interactions, (ii) van der Waals interactions, (iii) hydrogen-bonding between polar groups of the guest and the hydroxyl groups of the host, (iv) relaxation by release of high-energy water from the cyclodextrin cavity upon substrate inclusion, and (v) relief of the conformational strain in a cyclodextrin-water adduct. CD complex formation usually results from different combinations of these forces.

* Corresponding author. Tel.: +34 968 278 755.

E-mail address: Aperez@pdi.ucam.edu (A. Pérez-Garrido).

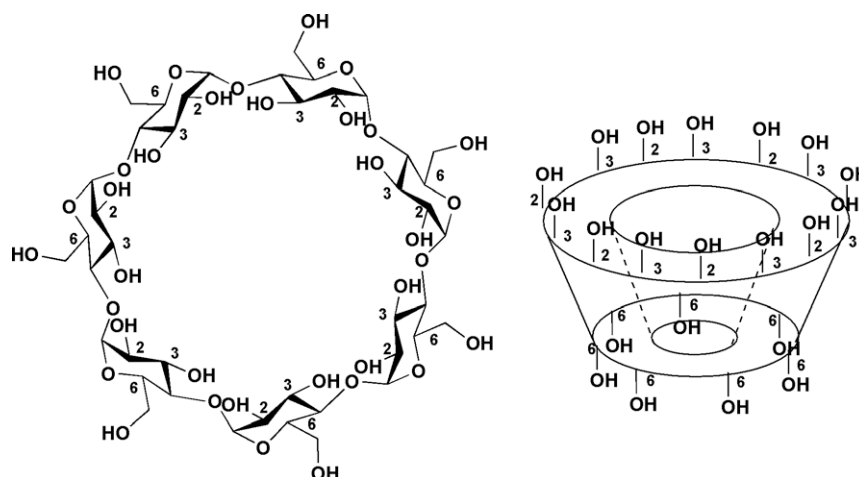


Fig. 1. Chemical structure of β -cyclodextrin.

The aim of the present study is to build a multiple linear regression QSAR model, able to correlate and predict the complex stability constant between diverse guest molecules and β -CDs, since these are the most commonly used. Special emphasis will be given to elucidate the driving forces leading to the complexation of the set of molecules under study. For this purpose, we resorted to the free-software package DRAGON, available at the internet site: www.vcclab.org/lab/pclient/.¹⁶ DRAGON contains more than 1600 molecular descriptors divided into several families: 0D (constitutional descriptors), 1D (e.g., functional group counts), 2D (e.g., topological descriptors and connectivity indices), and 3D (e.g., GET-AWAY, WHIM, RDF and 3D-MORSE descriptors). These descriptors have proved to be particularly useful in QSAR/QSPR modeling studies, and to provide satisfactory correlation between the modeled target and molecular parameters.^{17–32}

2. Materials and methods

2.1. Data set

The overall data set of 233 substances comprised a large number of classes of organic compounds: aromatic hydrocarbons, alcohols, phenols, ethers, aldehydes, ketones, acids, esters, nitriles, anilines, halogenated compounds, heterocycles, nitro, sulfur and steroids and barbital compounds. This set of guest molecules was extracted from the work of Suzuki,⁶ and the experimental endpoint to be predicted is the β -CD complex stability constant (K) in water at 298 K taken from references therein. Two of such guest molecules are stereoisomers chemicals 214 and 215, which could not be distinguished by the present 2D descriptors but had nevertheless different K values. Thus, one of the isomers was discarded (chemical 215), being only the other one (chemical 214) considered in our study with an averaged value of K . Moreover, all K values were log-transformed ($\log K$) for being of practical use in the following QSAR modeling. Table 1 displays a complete list of the chemicals along with the reported experimental data.

2.2. Model selection and validation

The structures of the compounds were first drawn with the aid of ISIS/Draw software ver. 2.5.³³ Molecular structures were then fully optimized with the molecular mechanics method (MM2)³⁴ followed by the PM3 semi-empirical Hamiltonian^{35,36} implemented in MOPAC ver. 6.0.³⁷ Subsequently, different families of descriptors were calculated using the web-DRAGON.¹⁶ Input variables with constants or closed to constants values were immedi-

ately eliminated. To validate the models, k -means cluster analysis was used to split the original dataset of chemicals into training and test sets. Mathematical models were obtained afterwards by means of multiple linear regression analysis along with a variable subset selection procedure.

2.3. k -Means cluster analysis

Developing rational approaches for the selection of training and test set compounds is an active area of research. These approaches range, for instance, from straightforward random selection³⁸ to various clustering techniques.³⁹ The main goal of k -means cluster analysis (k -MCA) is to partition the original series of compounds into several statistically representative classes of chemicals, among which one might then select the training and test set compounds. Here, we have decided that the training set should contain 80% (186/233) of the original data and the test set the remaining 20%, to guarantee that any kind of substance as determined by the clusters derived from k -MCA was represented in each set.

Starting from all descriptors of all 0–3D family types, those that produce the greatest separation of clusters meanwhile ensuring a statistically acceptable data partition were selected. In so doing, we took into account the number of members in each cluster and the standard deviation of the variables in the cluster (as low as possible). The k -MCA split the compounds into four clusters comprising 76, 53, 69 and 35 members with standard deviations of 0.08, 0.14, 0.2 and 0.2, respectively. Selection of the training and test sets was carried out by taking compounds belonging to each cluster, proportionally to the size of the cluster. We also made an inspection of the standard deviation between and within clusters, the respective Fisher ratio and p level of significance (ought to be lower than 0.05)^{40,41} (Table 2).

2.4. Variable selection

Presently, there is a vast amount and wide range of molecular descriptors with which one can model the activity of interest. This makes the search for gathering the most suitable subset quite complicated and time consuming because of the many possible combinations, especially if one tries to define an accurate, robust, and (above all) interpretable model. For this reason, we applied the genetic algorithm (GA)⁴² procedure for selecting the variables, as implemented in Mobydigs software ver. 1.0.⁴³ The particular GA simulation applied here resorted to the generation of 100 regression models, ordered according to their increased internal predictive performance (verified by cross-validation). First of all, models

Table 1
Names, observed and predicted activity^a, and leverage values for the compounds used in this study

No	Name	Log <i>K</i> _{obs}	Log <i>K</i> _{pred}	Partition	Leverage	Ref.	No	Name	log <i>K</i> _{obs}	Log <i>K</i> _{pred}	Partition	Leverage	Ref.
1	Carbon tetrachloride	2.20	2.25	Test	0.058	10	37	Iodobenzene	2.93	2.44	Training	—	12
2	Chloroform	1.43	1.39	Training	—	10	38	3-Fluorophenol	1.70	1.60	Training	—	8
3	Methanol	-0.49	-0.74	Training	—	8	39	4-Fluorophenol	1.73	1.66	Training	—	8
4	Acetonitrile	-0.27	-0.38	Training	—	10	40	3-Chlorophenol	2.28	2.41	Training	—	8
5	Acetaldehyde	-0.64	-0.38	Training	—	10	41	4-Chlorophenol	2.61	2.47	Training	—	12
6	Ethanol	-0.03	-0.08	Test	0.114	8	42	3-Bromophenol	2.51	2.52	Test	0.013	8
7	1,2-Ethanediol	-0.19	0.17	Training	—	8	43	4-Bromophenol	2.65	2.58	Test	0.015	12
8	Acetone	0.42	0.40	Training	—	10	44	3-Iodobenzene	2.93	2.65	Training	—	8
9	1-Propanol	0.57	0.54	Test	0.062	8	45	4-Iodobenzene	2.98	2.71	Training	—	12
10	2-Propanol	0.63	0.77	Training	—	10	46	Nitrobenzene	2.04	1.75	Training	—	10
11	1,3-Propanediol	0.67	0.70	Training	—	8	47	4-Nitrophenol	2.39	1.80	Training	—	12
12	Tetrahydrofuran	1.47	0.90	Test	0.034	10	48	Benzene	2.23	1.40	Test	0.025	64
13	Cyclobutanol	1.18	1.25	Training	—	8	49	Phenol	1.98	1.71	Training	—	12
14	1-Butanol	1.22	1.07	Training	—	64	50	Hydroquinone	2.05	1.93	Test	0.015	12
15	2-Butanol	1.19	1.34	Training	—	8	51	4-Nitroaniline	2.48	2.14	Test	0.020	12
16	2-Methyl-1-propanol	1.62	1.35	Training	—	8	52	Aniline	1.60	1.92	Training	—	10
17	2-Methyl-2-propanol	1.68	1.33	Training	—	8	53	Sulfanilamide	2.76	2.72	Training	—	11
18	1,4-Butanediol	0.64	1.12	Training	—	8	54	Cyclohexanol	2.67	2.55	Training	—	64
19	Diethylamine	1.36	1.22	Training	—	10	55	1-Hexanol	2.33	1.79	Training	—	64
20	Cyclopentanol	2.08	1.87	Training	—	8	56	2-Hexanol	1.98	2.28	Training	—	8
21	1-Pentanol	1.80	1.49	Training	—	64	57	2-Methyl-2-pentanol	1.99	2.55	Training	—	8
22	2-Pentanol	1.49	1.87	Training	—	8	58	3-Methyl-3-pentanol	2.15	2.27	Training	—	8
23	3-Pentanol	1.35	1.70	Training	—	8	59	4-Methyl-2-pentanol	2.04	2.40	Training	—	8
24	2-Methyl-1-butanol	2.08	1.71	Training	—	8	60	3,3-Dimethyl-2-butanol	2.75	2.28	Training	—	8
25	2-Methyl-2-butanol	1.91	1.93	Training	—	8	61	1,6-Hexanediol	1.69	1.65	Training	—	8
26	3-Methyl-1-butanol	2.25	1.88	Test	0.021	8	62	Benzonitrile	2.23	1.84	Training	—	12
27	3-Methyl-2-butanol	1.92	1.82	Training	—	8	63	Benzothiazole	2.38	2.59	Training	—	65
28	2,2-Dimethyl-1-propanol	2.71	1.94	Test	0.074	64	64	4-Nitrobenzoic acid	2.34	1.71	Training	—	12
29	1,5-Pentanediol	1.22	1.43	Test	0.120	8	65	Benzaldehyde	1.78	1.89	Training	—	10
30	1,4-Dibromobenzene	2.97	3.14	Training	—	12	66	Benzoic acid	2.12	2.05	Training	—	64
31	1,4-Diiodobenzene	3.17	3.38	Training	—	12	67	4-Hydroxybenzaldehyde	1.75	2.04	Training	—	8
32	3,5-Dibromophenol	2.56	3.20	Training	—	8	68	4-Hydroxybenzoic acid	2.20	2.09	Training	—	12
33	3,5-Dichlorophenol	2.07	2.99	Test	0.020	8	69	Benzyl chloride	2.45	2.70	Training	—	12
34	1-Chloro-4-nitrobenzene	2.15	2.39	Training	—	12	70	Toluene	2.09	2.03	Training	—	10
35	Fluorobenzene	1.96	1.37	Test	0.020	12	71	benzyl alcohol	1.71	2.25	Training	—	10
36	Bromobenzene	2.50	2.31	Training	—	12	72	Anisole	2.32	2.11	Training	—	12
73	<i>m</i> -Cresol	1.98	2.24	Training	—	8	109	<i>N,N</i> -Dimethylaniline	2.36	2.80	Training	—	12
74	<i>p</i> -Cresol	2.40	2.30	Training	—	12	110	Barbital	1.78	2.39	Training	—	11
75	4-Methoxyphenol	2.21	2.27	Training	—	12	111	cyclooctanol	3.30	3.31	Training	—	8
76	3-Methoxyphenol	2.11	2.19	Training	—	8	112	1-Octanol	3.17	2.13	Test	0.212 ^c	8
77	4-Hydroxybenzyl alcohol	2.16	2.39	Training	—	12	113	2-Octanol	3.13	2.74	Training	—	8
78	Hydrochlorothiazide	1.76	1.94	Training	—	11	114	Quinoline	2.12	2.47	Training	—	65
79	<i>N</i> -Methylaniline	2.12	2.38	Training	—	12	115	3-Cyanophenyl acetate	1.49	2.24	Training	—	8
80	1-Butylimidazole	2.19	2.96	Training	—	66	116	4-Hydroxycinnamic acid	2.83	2.61	Training	—	11
81	1-Heptanol	2.85	2.00	Test	0.164 ³	8	117	Ethyl benzoate	2.73	2.53	Training	—	12
82	Phenylacetylene	2.36	2.07	Training	—	12	118	4'-Hydroxypropiofenone	2.63	2.70	Training	—	8
83	Thianaphthene	3.23	2.83	Training	—	65	119	3'-Hydroxypropiofenone	2.61	2.61	Training	—	8
84	4-Fluorophenyl acetate	2.11	2.24	Test	0.031	8	120	<i>p</i> -Tolyl acetate	2.49	2.78	Training	—	8
85	3-Fluorophenyl acetate	1.91	2.13	Training	—	8	121	3-Methylphenyl acetate	2.21	2.69	Training	—	8
86	4-Chlorophenyl acetate	2.50	2.93	Training	—	8	122	4-Methoxyphenyl acetate	2.45	2.45	Training	—	8
87	3-Chlorophenyl acetate	2.44	2.84	Training	—	8	123	4-Propylphenol	3.55	3.14	Training	—	8
88	4-Bromophenyl acetate	2.68	3.02	Training	—	8	124	3-Propylphenol	3.28	3.05	Training	—	8
89	3-Bromophenyl acetate	2.67	2.94	Test	0.018	8	125	4-Isopropylphenol	3.58	3.18	Training	—	8
90	4-Iodobenzene	3.00	3.15	Training	—	8	126	3-Isopropylphenol	3.44	3.08	Training	—	8
91	3-Iodobenzene	3.07	3.06	Training	—	8	127	4-Isopropoxyphenol	2.86	3.08	Training	—	8
92	4-Nitrophenyl acetate	2.13	1.91	Training	—	8	128	2-Norbornaneacetate	3.59	3.42	Test	0.040	64
93	Acetophenone	2.27	2.34	Training	—	12	129	1-Benzylimidazole	2.61	3.12	Training	—	66
94	Phenyl acetate	2.10	2.39	Training	—	8	130	<i>m</i> -Methylcinnamic acid	2.93	2.95	Training	—	11
95	Methyl benzoate	2.50	2.24	Training	—	12	131	4-Ethylphenyl acetate	2.83	2.97	Test	0.014	8
96	3-Hydroxyacetophenone	2.06	2.35	Training	—	8	132	3-Ethylphenyl acetate	2.68	2.82	Training	—	8
97	4-Hydroxyacetophenone	2.18	2.44	Training	—	12	133	4-Ethoxyphenyl acetate	2.54	2.63	Training	—	8
98	Acetoanilide	2.20	2.65	Test	0.011	12	134	3-Ethoxyphenyl acetate	2.49	2.47	Test	0.019	8
99	<i>p</i> -Xylene	2.38	2.61	Training	—	12	135	Allobarbital	1.98	2.28	Training	—	11
100	Ethylbenzene	2.59	2.55	Training	—	12	136	4- <i>n</i> -Butylphenol	3.97	3.44	Test	0.027	8
101	Phenetole	2.49	2.57	Training	—	12	137	3- <i>n</i> -Butylphenol	3.76	3.35	Training	—	8
102	2-Phenylethanol	2.15	2.72	Training	—	8	138	3-Isobutylphenol	4.21	3.45	Training	—	8
103	3-Ethylphenol	2.60	2.66	Training	—	8	139	4- <i>sec</i> -Butylphenol	4.18	3.41	Training	—	8
104	4-Ethylphenol	2.69	2.75	Training	—	12	140	3- <i>sec</i> -Butylphenol	4.06	3.31	Training	—	8
105	4-Ethoxyphenol	2.33	2.66	Test	0.013	8	141	4- <i>tert</i> -Butylphenol	4.56	3.69	Test	0.032	8
106	3-Ethoxyphenol	2.35	2.58	Test	0.010	8	142	3- <i>tert</i> -Butylphenol	4.41	3.58	Training	—	8
107	3,5-Dimethoxyphenol	2.34	2.25	Training	—	8	143	Menadion	2.27	2.42	Test	0.023	11
108	<i>N</i> -Ethylaniline	2.34	2.83	Test	0.016	12	144	Sulfapyridine	2.70	2.68	Training	—	11
145	Sulfamonomethoxine	2.48	1.87	Training	—	11	181	4- <i>n</i> -Amylphenyl acetate	3.80	3.35	Training	—	8
146	Sulfisoxazole	2.32	2.58	Training	—	11	182	Flufenamic acid	3.10	2.75	Training	—	64
147	4- <i>n</i> -Propylphenyl acetate	3.15	3.13	Training	—	8	183	Meclofenamic acid	2.67	3.38	Training	—	64
148	3- <i>n</i> -Propylphenyl acetate	3.28	2.96	Training	—	8	184	Nitrazepam	1.97	1.97	Training	—	11

Table 1 (continued)

No	Name	Log K_{obs}	Log K_{pred}	Partition	Leverage	Ref.	No	Name	log K_{obs}	Log K_{pred}	Partition	Leverage	Ref.
149	4-Isopropylphenyl acetate	2.88	3.26	Training	—	8	185	Flurbiprofen	3.69	3.02	Training	—	11
150	3-Isopropylphenyl acetate	3.36	3.09	Training	—	8	186	Sulfaphenazole	2.35	2.17	Training	—	11
151	4- <i>n</i> -Amylphenol	4.19	3.65	Test	0.039	8	187	Bendroflumethiazide	1.90	2.40	Training	—	11
152	4- <i>tert</i> -Amylphenol	4.70	3.84	Training	—	8	188	Mefenamic acid	2.49	2.40	Training	—	11
153	Carbutamide	2.29	2.82	Training	—	11	189	Acetohexamide	2.94	3.18	Test	0.047	11
154	Pentobarbital	3.01	2.79	Test	0.042	11	190	Fludiazepam	2.33	2.45	Training	—	11
155	Amobarbital	3.07	3.01	Training	—	64	191	Nimetazepam	1.73	1.99	Training	—	11
156	Thiopental	3.28	3.40	Training	—	11	192	Fenbufen	2.63	3.19	Training	—	11
157	Dibenzofuran	2.97	2.77	Training	—	65	193	Ketoprofen	2.85	2.77	Training	—	11
158	Dibenzothiophene	3.48	3.39	Training	—	65	194	Medazepam	2.40	3.09	Training	—	11
159	Phenazine	2.41	2.69	Training	—	65	195	Progabide	2.53	2.98	Test	0.080	11
160	Thianthrene	3.57	3.82	Test	0.039	65	196	Griseofulvin	1.47	1.56	Training	—	11
161	Carbazole	2.44	3.01	Training	—	65	197	Tolnaftate	3.83	3.38	Training	—	11
162	Phenoxazine	2.69	2.85	Test	0.021	65	198	Prostacyclin	2.94	3.70	Training	—	11
163	Phenothiazine	2.73	3.20	Training	—	65	199	Triamcinolone	3.37	3.32	Test	0.087	67
164	furosemide	1.78	2.47	Test	0.071	11	200	cortisone	3.35	3.49	Training	—	11
165	Phenobarbital	3.22	2.50	Test	0.033	64	201	Prednisolone	3.56	3.65	Test	0.065	67
166	Sulfisomidine	2.10	2.32	Test	0.108	11	202	Hydrocortisone	3.60	3.77	Training	—	11
167	Sulfamethomidine	2.33	1.94	Test	0.106	11	203	Corticosterone	3.85	3.89	Test	0.073	67
168	Sulfadimethoxine	2.26	1.50	Training	—	11	204	Dexamethasone	3.65	3.63	Training	—	11
169	4- <i>n</i> -Butylphenyl acetate	3.62	3.26	Training	—	8	205	Betamethasone	3.73	3.82	Training	—	67
170	3- <i>n</i> -Butylphenyl acetate	3.66	3.08	Training	—	8	206	Paramethasone	3.40	3.59	Training	—	67
171	3-Isobutylphenyl acetate	3.83	3.24	Training	—	8	207	Cortisone-21-acetate	3.62	3.45	Training	—	67
172	4- <i>tert</i> -Butylphenyl acetate	3.85	3.72	Training	—	8	208	Prednisolone-21-acetate	3.76	3.63	Training	—	67
173	Cyclobarbital	2.71	2.90	Training	—	11	209	Hydrocortisone-21-acetate	3.51	3.69	Training	—	67
174	Hexobarbital	3.08	3.02	Training	—	11	210	Fluocinolone acetone	3.48	2.97	Training	—	67
175	1-Adamantaneacetate	4.32	4.04	Training	—	64	211	Triamcinolone acetone	3.51	3.39	Training	—	67
176	Acridine	2.33	2.91	Training	—	65	212	Spirolactone	4.44	3.79	Training	—	67
177	Phenanthridine	2.57	2.82	Training	—	65	213	Dehydrocholic acid	3.38	3.39	Training	—	67
178	Xanthene	2.71	2.99	Training	—	65	214	Chenodeoxycholic acid	4.36 ^b	4.74	Training	—	67
179	<i>N</i> -Phenylanthranilic acid	2.89	2.85	Training	—	64	215	Ursodeoxycholic acid	4.51 ^b	—	Training	—	67
180	Mephobarbital	3.16	2.53	Training	—	11	216	Cholic acid	3.50	4.38	Test	0.121	67
217	Hydrocortisone-17-butyrate	3.23	3.25	Training	—	67	226	1- α - <i>O</i> -benzylglycerol	2.11	3.22	Test	0.019	66
218	Cinnarizine	3.64	3.71	Training	—	11	227	Sulfamerazine	1.97	2.37	Training	—	11
219	Cycloheptanol	3.23	2.94	Training	—	8	228	Butyl 4-hydroxybenzoate	3.39	2.86	Training	—	11
220	2-Methoxyethanol	0.22	0.58	Test	0.068	8	229	Butyl 4-aminobenzoate	3.19	3.14	Training	—	11
221	3-Hydroxycinnamic acid	2.56	2.54	Training	—	11	230	Benzidine	3.35	3.54	Test	0.021	11
222	Ethyl 4-hydroxybenzoate	3.01	2.49	Training	—	11	231	Triflumizole	2.66	2.60	Training	—	11
223	Ethyl 4-aminobenzoate	2.69	2.81	Test	0.012	11	232	Diazepam	2.33	2.75	Training	—	11
224	4-Methylcinnamic acid	2.65	3.04	Training	—	11	233	Prostaglandine E2	3.09	2.91	Training	—	11
225	Sulfadiazine	2.52	2.25	Test	0.077	11							

^a β -CD complex stability constant (K), then log-transformed ($\log K$).

^b Chemicals 214 and 215 were replaced by only one compound (chemical 214) with an averaged $\log K$ value (=4.44).

^c Chemicals 81 and 112 have leverage values above the threshold (0.14) and, for that reason, its predictions were not taken into account when calculating Q_{EXT}^2 .

Table 2

Standard deviation between and within clusters, degrees of freedom (df), Fisher ratio (F) and level of significance (p) of the variables in the k -means cluster analysis

	Between SS	df	Within SS	df	F	p
VEZ1	208.9675	3	23.03249	229	692.5517	<10 ⁻⁵
VEm1	208.9593	3	23.04073	229	692.2767	<10 ⁻⁵
VEv1	209.6369	3	22.36308	229	715.5670	<10 ⁻⁵
VEe1	209.1965	3	22.80353	229	700.2717	<10 ⁻⁵
VEp1	209.6248	3	22.37521	229	715.1377	<10 ⁻⁵
Xu	211.0286	3	20.97139	229	768.1187	<10 ⁻⁵

with one to two variables were developed by the variable subset selection procedure in order to explore all low combinations. The number of descriptors was subsequently increased one by one, and new models formed. The GA was stopped when further increments in the size of the model did not increase internal predictivity in any significant degree. Furthermore, the following conditions were used on our GA simulation: the maximum number of variables in a model was 10, the number of best retained models for each size was 5, the trade off between crossovers and mutation parameter (T) was from 0.3 to 0.7, and selection bias ($B\%$) was from 30 to 90.

Table 3
Best models derived using from 2 to 10 variables for each family of descriptors

	Variables	N ^a	F	s	Q _{CV-100} ²	In domain ^b (%)	Q _{boot} ²
Topologic	ZM1, S1K, ZM1V, SMTIV, LPRS PHI, J, Xu, T(N..S), T(O..O)	10	93.72	0.362	0.821	95.74	0.814
GETAWAY	HGM, H3m, H0v, HATSp, R6v R4v+, R7e, R4p, R6p, R8p+	10	69.34	0.414	0.776	95.74	0.763
Eigenvalues-based	AEige, SEige, VRA1, VRv2, VRm2 SEigm, VRA2, VEA1, SEigv, VRp1	10	74.07	0.398	0.769	97.87	0.760
Conectivity	χ ₀ , χ ₁ , χ _{1A} , χ _{2A} , χ _{0v} χ _{MOD} , χ ₃ , χ _{3A} , χ _{3sol} , RDCHI	10	68.63	0.416	0.771	93.62	0.731
Burden eigenvalues	BEHm1, BELm4, BELm5, BELm7, BEHv1 BEHv8, BELv8, BEHe1, BELe7, BELe8	10	66.80	0.420	0.765	97.87	0.754
Molecular propieties	Ui, Hy, AMR, MLOGP2, GVWAI-80 Inflam-50, Hypert-50, Infect-80, Infect-50, BLTF96	10	55.48	0.452	0.727	95.74	0.712
3DMoRSE	Mor01u, Mor02m, Mor03m, Mor04m, Mor01v Mor07v, Mor31e, Mor01p, Mor04p, Mor05p	10	56.77	0.448	0.726	95.74	0.705
WHIM	L2m, L1p, L3s, E1s, Tp Au, Ae, As, Du, De	10	48.40	0.476	0.689	93.62	0.667
RDF	RDF010u, RDF015u, RDF020u, RDF085u, RDF020m RDF040m, RDF015v, RDF030e, RDF050p, RDF060p	10	40.18	0.508	0.654	93.62	0.632
Randić molecular profile	DP01, DP02, DP08, DP17, SP01 SP04, SP05, SP07, SP17, SHP2	10	26.16	0.584	0.557	95.74	0.532
Galvez topological charge	JG11, GG16, GGI1, JG16, GGI4 JG17, GGI2, GGI5, JGI5, JGI4	10	17.65	0.644	0.441	97.87	0.415

^a Number of variables.^b Percentage of chemicals from the training set within the applicability domain.

2.5. Model validation

Goodness of fit of the models was assessed by examining the determination coefficient (R^2), the standard deviation (s), Fisher ratio (F) and the ratio between the number of cases and the number of adjustable parameters in the model (known as the ρ statistics; notice that ρ should be ≥ 4).⁴⁴ Other important statistics, namely the Kubinyi function (FIT)^{45,46} and Akaike's information criteria (AIC)^{47,48} were taken into account, as they give enough criteria for comparing models with different parameters, numbers of variables and numbers of chemicals. As to the robustness and predictivity of the models, these were evaluated by means of cross-validation, basically leave-one-out (CV-LOO) and bootstrapping testing techniques calculated with the training set, by looking to the outcome statistics of both techniques (i.e., Q_{CV-LOO}^2 and Q_{boot}^2) as well as to the Q_{EXT}^2 values obtained with the test set substances that fall within the applicability do-

Table 4
Symbols and description of the topological descriptors involved in the QSAR model (Eq. 1)

Symbols	Descriptor definition
ZM1	First Zagreb index M1
ZM1V	First Zagreb index by valence vertex degrees
SMTIV	Schultz MTI by valence vertex degrees
Xu	Xu index
J	Balaban distance connectivity index
S1K	1-Path Kier alpha-modified shape index
T(N..S)	Sum of topological distances between N..S
T(O..O)	Sum of topological distances between O..O
PHI	Kier flexibility index
LPRS	Log of product of row sums (PRS)
VEZ1	Eigenvector coefficient sum from Z weighted distance matrix (Barysz matrix)
VEm1	Eigenvector coefficient sum from mass weighted distance matrix
VEv1	Eigenvector coefficient sum from van der Waals weighted distance matrix
VEe1	Eigenvector coefficient sum from electronegativity weighted distance matrix
VEp1	Eigenvector coefficient sum from polarizability weighted distance matrix

main of the model. Further, the stability under heavy perturbations in the training set was checked by examining the outcome statistics of a response randomization procedure (Y scrambling) for the training and test sets ($a(R^2)$ and $a(Q^2)$ values). All these calculations were carried out with software Mobydigs ver. 1.0.⁴³

To sum up, good quality of the models is indicated by high F , FIT and ρ values, by low s and AIC values, as well as by values closed to one for R^2 , Q_{CV-LOO}^2 , Q_{boot}^2 and Q_{EXT}^2 (save for $a(R^2)$ and $a(Q^2)$ values, which check random correlations).

2.6. Model orthogonalization

The main drawback of collinearity from the point of view of QSAR/QSPR modeling is that it increases the standard errors associated with the individual regression coefficients, thereby decreasing their value for purposes of interpretability. To overcome this problem, we employed here the Randić method of orthogonalization.^{49–53} The first step for orthogonalizing the molecular descriptors is to select the appropriate order of orthogonalization, which, in this case, is the order of significance of the variables in the model. The first variable (v_1) is taken as the first orthogonal descriptor and the second one is orthogonalized with respect to it by taking the residual of its correlation with v_1 . The process is repeated until all variables are completely orthogonalized, after which they are further standardized. Orthogonal standardized variables are then used to obtain a new model.

2.7. Applicability domain of the model

Given that the real utility of a QSAR model relies on its ability to accurately predict the modeled activity for new chemicals, careful assessment of the model's true predictive power is a must. This includes the model validation but also the definition of the applicability domain of the model in the space of molecular descriptors used for deriving the model. There are several methods for assessing the applicability domain of QSAR/QSPR models^{54,55} but the most common one encompasses determining the leverage values for each compound.⁵⁶ A Williams plot, that is, the plot of standardized residuals versus leverage values (h), can then be used for an

Table 5
Correlation matrix for intercorrelations among the ten variables of the QSAR model (Eq. 1)

	<i>Xu</i>	<i>ZM1V</i>	<i>LPRS</i>	<i>SMTIV</i>	<i>S1K</i>	<i>ZM1</i>	<i>T(N..S)</i>	PHI	<i>J</i>	<i>T(O..O)</i>
<i>Xu</i>	1.00	—	—	—	—	—	—	—	—	—
<i>ZM1V</i>	0.95	1.00	—	—	—	—	—	—	—	—
<i>LPRS</i>	0.99	0.94	1.00	—	—	—	—	—	—	—
<i>SMTIV</i>	0.92	0.90	0.97	1.00	—	—	—	—	—	—
<i>S1K</i>	0.97	0.92	0.98	0.95	1.00	—	—	—	—	—
<i>ZM1</i>	0.98	0.94	0.99	0.95	0.96	1.00	—	—	—	—
<i>T(N..S)</i>	0.28	0.36	0.26	0.26	0.27	0.26	1.00	—	—	—
<i>PHI</i>	0.68	0.57	0.68	0.67	0.77	0.58	0.15	1.00	—	—
<i>J</i>	−0.54	−0.53	−0.56	−0.54	−0.44	−0.57	−0.17	−0.18	1.00	—
<i>T(O..O)</i>	0.73	0.72	0.79	0.86	0.80	0.79	0.04	0.56	−0.37	1.00

Table 6
Step-by-step analysis of the forward stepwise orthogonalization process

Step	¹ ΩXu	² $\Omega ZM1V$	⁴ $\Omega ZM1$	³ $\Omega LPRS$	⁹ ΩJ	⁸ ΩPHI	⁷ $\Omega T(N..S)$	¹⁰ $\Omega T(O..O)$	⁶ $\Omega SMTIV$	⁵ $\Omega S1K$	Intercept	$R^2(Adj.)$	$\Delta R^2(Adj.)$	<i>p</i> -Level
1	0.0763	—	—	—	—	—	—	—	—	—	1.595	0.303	0.303	<10 ^{−5}
2	0.0763	−0.0109	—	—	—	—	—	—	—	—	1.595	0.502	0.199	<10 ^{−5}
3	0.0763	−0.0109	0.0703	—	—	—	—	—	—	—	1.595	0.650	0.148	<10 ^{−5}
4	0.0763	−0.0109	0.0703	−0.0389	—	—	—	—	—	—	1.595	0.724	0.074	<10 ^{−5}
5	0.0763	−0.0109	0.0703	−0.0389	−1.0263	—	—	—	—	—	1.595	0.756	0.032	<10 ^{−5}
6	0.0763	−0.0109	0.0703	−0.0389	−1.0263	−0.2921	—	—	—	—	1.595	0.781	0.025	<10 ^{−5}
7	0.0763	−0.0109	0.0703	−0.0389	−1.0263	−0.2921	−0.0492	—	—	—	1.595	0.803	0.022	<10 ^{−5}
8	0.0763	−0.0109	0.0703	−0.0389	−1.0263	−0.2921	−0.0492	−0.0132	—	—	1.595	0.825	0.022	<10 ^{−5}
9	0.0763	−0.0109	0.0703	−0.0389	−1.0263	−0.2921	−0.0492	−0.0132	0.0001	—	1.595	0.833	0.008	0.003450
10	0.0763	−0.0109	0.0703	−0.0389	−1.0263	−0.2921	−0.0492	−0.0132	0.0001	0.0465	1.595	0.834	0.001	0.123876

immediate and simple graphical detection of both the response outliers and structurally influential chemicals in the model. In this plot, the applicability domain is established inside a squared area within $\pm \chi$ standard deviations and a leverage threshold h^* (h^* is generally fixed at $3\kappa/n$, where n is the number of training compounds and κ the number of model parameters, whereas $\chi = 2$ or 3), lying outside this area (vertical lines) the outliers and (horizontal lines) influential chemicals. For future predictions, only predicted complex stability constant data for chemicals belonging to the chemical domain of the training set should be proposed and used.⁵⁷ So, calculations of Q_{EXT}^2 were performed only for those substances that had a leverage value below the threshold h^* .

3. Results and discussion

3.1. QSAR models

Several QSAR models for predicting β -cyclodextrins complex stability constants were developed, using the same training set and routine for variable selection. This was accomplished by finding regression models for the k -MCA chosen training set (185 compounds) based on GA selection (between 2 and 10 variables), in conjunction with the following eleven sets of molecular descriptors: topological, GETAWAY, eigenvalue-based indices, connectivity indices, Burden eigenvalues, molecular properties, 3D-MorSE, WHIM, RDF, Randić molecular profiles and Galvez topological charges indices. The best QSAR-models are given in Table 3.

There are substantial differences in the explanation of the experimental variance given by the topological model, when compared with the rest of the models. Thus, while the topological model is able to explain more than 84% of experimental variance, the other models, at best, can only explain 79.9% of such variance. The predictive ability-expressed as Q_{CV-100}^2 and Q_{EXT}^2 - of the topological model is also higher than the other descriptors' models, even for those based on 3D descriptors (3D-MorSE, WHIM, RDF, GETAWAY and Randić) that showed lower scores.

Topological descriptors, unlike three-dimensional descriptors, do not consider information on conformational aspects, such as bond lengths, bond angles and torsion angles, but do encode important information on adjacency, branching and relative distance among different functionalities in a numerical form. Thus, these molecular descriptors determine a wide range of physico-chemical properties of molecules. In addition, they can be derived from molecular structures using low computational resources, making them remarkably attractive in molecular modeling.

Successful correlations between β -CDs-complex stability constants and topological indices have also been found in the literature^{58,11} but with lesser number of ligands. Our resulting best-fit topological-QSAR model (a 10-variable equation) is given below together with the statistical parameters of the regression. As can be seen, this model is reasonable in both statistical significance and goodness of fit or prediction.

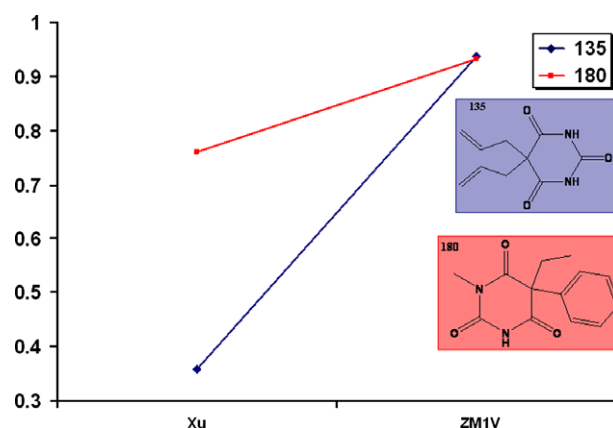


Fig. 2. Contributions from each of the variables to the final value of logK for allobarbital (chemical 135) and mephobarbital (chemical 180).

$$\begin{aligned} \log K = & 1.08(\pm 0.063)Xu - 1.51 \cdot 10^{-2}(\pm 9.23 \cdot 10^{-4})ZM1V \\ & - 0.38(\pm 2.61 \cdot 10^{-2})LPRS + 7.81 \cdot 10^{-4}(\pm 6.68 \cdot 10^{-5})SMTIV \\ & + 0.93(\pm 8.16 \cdot 10^{-2})S1K + 6.65 \cdot 10^{-2}(\pm 8.08 \cdot 10^{-3})ZM1 \\ & - 8.70 \cdot 10^{-2}(\pm 1.12 \cdot 10^{-2})T(N..S) - 0.41(\pm 5.85 \cdot 10^{-2})PHI \\ & - 1.20(\pm 0.17)J - 1.32 \cdot 10^{-2}(\pm 2.67 \cdot 10^{-3})T(O..O) \\ & - 0.77(\pm 0.28) \end{aligned} \quad (1)$$

$$\begin{aligned} N = 185; \quad R^2 = 0.843; \quad Q_{(CV-LOO)}^2 = 0.821; \quad s = 0.361; \\ F = 93.72; \quad AIC = 0.147; \quad FIT = 3.371 \quad Q_{boot}^2 = 0.813; \\ a(R^2) = 0.02; \quad a(Q^2) = -0.113; \quad Q_{EXT}^2 = 0.756 \end{aligned}$$

The meaning of each of the topological descriptor variable involved in the cluster analysis and thereby used in the model above is shown in Table 4.

An aspect deserving special attention is the degree of collinearity among the variables of the model, which can be readily diagnosed by analyzing the cross-correlation matrix. As seen in Table 5, the pairs of descriptors (Xu ; $ZM1V$), (Xu ; $LPRS$), (Xu ; $SMTIV$), (Xu ; $S1K$), (Xu ; $ZM1$), ($T(O..O)$; $ZM1V$), ($T(O..O)$; $LPRS$), ($T(O..O)$; $SMTIV$), ($T(O..O)$; $S1K$), ($T(O..O)$; $ZM1$), ($T(O..O)$; Xu), ($ZM1V$; $LPRS$), ($ZM1V$; $SMTIV$), ($ZM1V$; $S1K$), ($ZM1V$; $ZM1$), ($LPRS$; $SMTIV$), ($LPRS$; $S1K$), ($LPRS$; $ZM1$), ($SMTIV$; $S1K$), ($SMTIV$; $ZM1$), ($S1K$; $ZM1$), and ($S1K$; PHI) are correlated with each other. Rather than deleting any of these descriptors, it is of interest to examine the performance of orthogonal complements in modeling β -CD complexation.

Following the Randić technique, we have determined orthogonal complements for all variables of the non-orthogonalized model (see Table 6). As a result, variable ${}^5\Omega S1K$ was found to be statistically non-significant ($p = 0.124$; Table 6), may be because the information contained in this variable is common to the information contained in the other descriptor variables. Furthermore, the significance of adding ${}^5\Omega S1K$ to the model remains unclear as seen from the modest improvements in $R^2(Adj.)$ (adjusted determination coefficient) on going from step 9 to 10 (see in Table 6, $R^2(Adj.)$ from step 9 to 10). Thus, by removing it, we obtained the following QSAR model, which is given below after to be standardized.

$$\begin{aligned} \log K = & 0.488(\pm 0.027)^1\Omega Xu - 0.392(\pm 0.026)^2\Omega ZM1V \\ & - 0.239(\pm 0.027)^3\Omega LPRS + 0.076(\pm 0.026)^4\Omega SMTIV \\ & + 0.337(\pm 0.026)^6\Omega ZM1 - 0.134(\pm 0.027)^7\Omega T(N..S) \\ & - 0.161(\pm 0.030)^8\Omega PHI - 0.160(\pm 0.027)^9\Omega J \\ & - 0.127(\pm 0.026)^{10}\Omega T(O..O) + 2.535(\pm 0.027) \end{aligned} \quad (2)$$

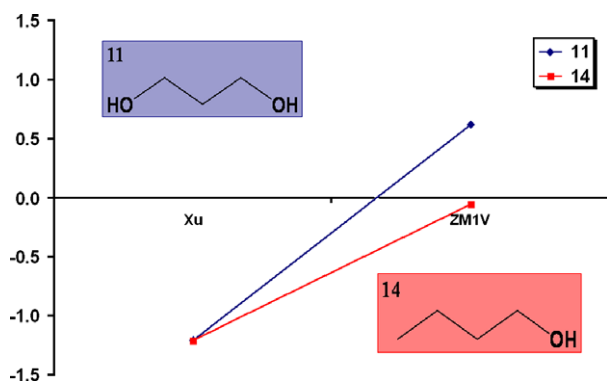


Fig. 3. Contributions from each of the variables to the final value of $\log K$ for 1,3-propanediol (chemical 11) and 1-butanol (chemical 14).

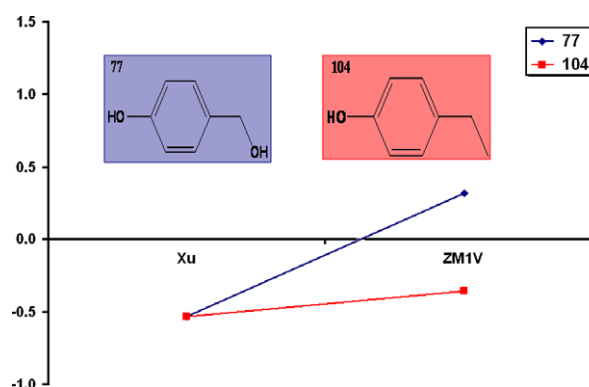


Fig. 4. Contributions from each of the variables to the final value of $\log K$ for 4-hydroxybenzyl alcohol (chemical 77) and 4-ethylphenol (chemical 104).

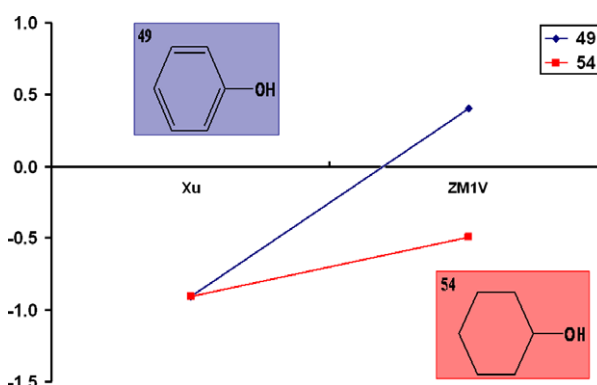


Fig. 5. Contributions from each of the variables to the final value of $\log K$ for phenol (chemical 49) and cyclohexanol (chemical 54).

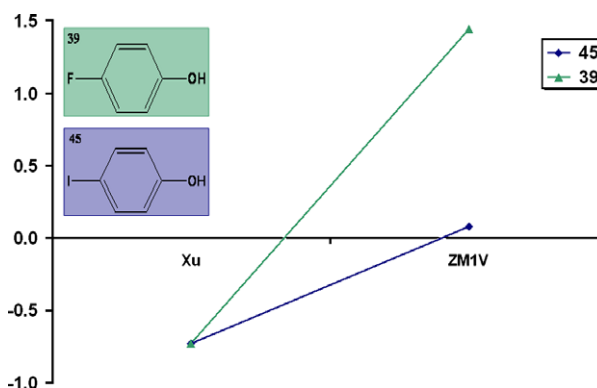


Fig. 6. Contributions from each of the variables to the final value of $\log K$ for 4-fluorophenol (chemical 39) and 4-iodophenol (chemical 45).

$$N = 185; \quad R^2 = 0.841; \quad Q_{(CV-LOO)}^2 = 0.821; \quad s = 0.363;$$

$$F = 103.05; \quad AIC = 0.147; \quad FIT = 3.487$$

$$Q_{boot}^2 = 0.812; \quad a(R^2) = 0.014; \quad a(Q^2) = -0.105; \quad Q_{EXT}^2 = 0.764$$

As can be seen in Table 6, removal of ${}^5\Omega S1K$ had little effect on the overall fitness of the model as the statistics are as robust as before, and further, by comparing Eq. 1 with Eq. 2, one can see that there are no changes in either the sign of the regression coefficients. Nevertheless, the relative contributions of the variables in the orthogonal-descriptor model are quite different to those related to the non-orthogonalized model. Therefore, for purposes of QSAR interpretability, we shall use the orthogonal-descriptor model defined in Eq. 2.

187, 188, 196, 198, 205, 210, 214, 218, 231 and 233. Even so, the latter should not be considered outliers but influential chemicals.⁵⁴

Nevertheless, all evaluations pertaining to the external set were performed by taking into account the applicability domain of our QSAR model. So, if a chemical belonging to the test set had a leverage value greater than h^* , we consider that this means that the prediction is the result of substantial extrapolation and therefore may not be reliable.⁵⁵

4. Conclusions

The forces affecting the phenomenon of complexation of chemicals with CDs are numerous and active in combination. In this study, we have examined the ability of a large and diverse set of substances to provide statistically sound and predictive QSAR models of β -CD complexation. We have thoroughly evaluated regression models in conjunction with a variety of structure representations, codifying a number of topological, physicochemical and three-dimensional aspects. For the present training set, topological descriptors provided the best model, as judged by extensive cross-validation and external-prediction. This topological-QSAR model was found to be superior to models derived using other 2D descriptors Burden eigenvalues, Galvez topological charge indices, connectivity indices, and eigenvalue-based indices- or even 3D descriptors Randić molecular profiles, RDF, 3DMORSE, GETAWAY, and WHIM- or molecular properties.

Moreover, the driving forces for CD complexation ascertained by the model are hydrophobic and steric (van der Waals) interactions mainly. Finally, this is a simple model that might be used in the prediction of β -CD complex stability constants of compounds inside the applicability domain. It may thus constitute an alternative and particular useful tool for screening large libraries of compounds.

Acknowledgments

A.M.H. acknowledges the Portuguese Fundação para a Ciência e a Tecnologia (FCT - Lisboa) (SFRH/BD/22692/2005) for financial support.

References and notes

- Saenger, W.; Jacob, J.; Gessler, K.; Steiner, T.; Daniel, S.; Sanbe, H.; Koizumi, K.; Smith, S. M.; Tanaka, T. *Chem. Rev.* **1998**, *98*, 1787–1802.
- Szejtli, J. *Chem. Rev.* **1998**, *98*, 1743–1753.
- Hedges, A. R. *Chem. Rev.* **1998**, *98*, 2035–2044.
- Yoshida, A.; Yamamoto, M.; Hirayama, F.; Uekawa, K. *Chem. Pharm. Bull.* **1988**, *36*, 4075–4080.
- Lipkowitz, K. B. *Chem. Rev.* **1998**, *98*, 1829–1873.
- Suzuki, T. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1266–1273.
- Pérez, F.; Jaime, C.; Sánchez-Ruiz, X. *J. Org. Chem.* **1995**, *60*, 3840–3845.
- Matsui, Y.; Nishioka, T.; Fujita, T. *Top. Curr. Chem.* **1985**, *128*, 61–89.
- Davis, D. M.; Savage, J. R. *J. Chem. Res. (S)* **1993**, 94–95.
- Park, J. H.; Nah, T. H. *J. Chem. Soc.* **1994**, *Perkin Trans. 2*, 1359–1362.
- Klein, C. T.; Polheim, D.; Viernstein, H.; Wolschann, P. *J. Inclusion Phenom. Macrocyclic Chem.* **2000**, *36*, 409–423.
- Liu, L.; Guo, Q.-X. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 133–138.
- Suzuki, T.; Ishida, M.; Fabian, W. M. *F. J. Comput.-Aided Mol. Des.* **2000**, *14*, 669–678.
- Cramer, R. D.; Patterson, D. E.; Bunce, J. D. *J. Am. Chem. Soc.* **1988**, *110*, 5959–5967.
- Katritzky, A. R.; Fara, D. C.; Yang, H. F.; Karelson, M.; Suzuki, T.; Solov'ev, V. P.; Varnek, A. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 529–541.
- Tetko, I. V.; Gasteiger, J.; Todeschini, R.; Mauri, A.; Livingstone, D.; Ertl, P.; Palyulin, V. A.; Radchenko, E. V.; Zefirov, N. S.; Makarenko, A. S.; Tanchuk, V. Y.; Prokopenko, V. V. *J. Comput. Aid. Mol. Des.* **2005**, *19*, 453–463.
- Saiz-Urra, L.; González, M. P.; Teijeira, M. *Bioorg. Med. Chem.* **2007**, *15*, 3565–3571.
- Saiz-Urra, L.; González, M. P.; Teijeira, M. *Bioorg. Med. Chem.* **2006**, *14*, 7347–7358.
- González, M. P.; Terán, C.; Teijeira, M.; Helguera, A. M. *Bull. Math. Bio.* **2007**, *69*, 347–359.
- Saiz-Urra, L.; González, M. P.; Fall, Y.; Gómez, G. *Eur. J. Med. Chem.* **2007**, *42*, 64–70.
- González, M. P.; Suárez, P. L.; Fall, Y.; Gómez, G. *Bioorg. Med. Chem.* **2005**, *15*, 5165–5169.
- Helguera, A. M.; Perez, M. A. C.; González, M. P. *J. Mol. Model.* **2006**, *12*, 769–780.
- Helguera, A. M.; Cordeiro, M. N. D. S.; Perez, M. A. C.; Combes, R. D.; González, M. P. *Bioorg. Med. Chem.* **2008**, *16*, 3395–3407.
- Helguera, A. M.; Rodríguez-Borges, J. E.; García-Mera, X.; Fernández, F.; Cordeiro, M. N. D. S. *J. Med. Chem.* **2007**, *50*, 1537–1545.
- Gupta, M. K.; Prabhakar, Y. S. *J. Chem. Inf. Model.* **2006**, *46*, 93–102.
- Gupta, S.; Singh, M.; Madan, A. K. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 272–277.
- Pirrung, M. C.; Tumej, L. N.; McClerren, A. L.; Raetz, C. R. H. *J. Am. Chem. Soc.* **2003**, *125*, 1575–1586.
- McElroy, N. R.; Jurs, P. C. *J. Med. Chem.* **2003**, *46*, 1066–1080.
- Hayatshahia, S. H. S.; Abdolmalekia, P.; Ghiasib, M.; Safarian, S. *FEBS Lett.* **2007**, *581*, 506–514.
- Kline, T. et al. *J. Med. Chem.* **2002**, *45*, 3112–3129.
- Sakowski, J.; Böhm, M.; Sattler, I.; Dahse, H. M.; Schlitzer, M. *J. Med. Chem.* **2001**, *44*, 2886–2899.
- Kleinman, E. F.; Campbell, E.; Giordano, L. A.; Cohan, V. L.; Jenkinson, T. H.; Cheng, J. B.; Shirley, J. T.; Pettipher, E. R.; Salter, E. D.; Hibbs, T. A.; DiCapua, F. M.; Bordner, J. *J. Med. Chem.* **1998**, *41*, 266–270.
- ISIS/Draw, Symyx MDL, San-Leandro, California, USA.
- Allinger, N. L.; Zhou, X. F.; Bergsma, J. *J. Mol. Struct. (Theochem.)* **1994**, *118*, 69–83.
- Stewart, J. J. P. *J. Comp. Chem.* **1989**, *10*, 209–220.
- Stewart, J. J. P. *J. Comp. Chem.* **1989**, *10*, 221–264.
- Frank, J. *Seiler Research Laboratory*; US Air Force Academy: Colorado, Springs Co., 1993.
- Yasri, A.; Hartsough, D. J. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1218–1227.
- Gore, P. A. J. *Handbook of applied multivariate statistics and mathematical modeling*. In *Chapter Cluster analysis*; Tinsley, H. E. A., Brown, S. D., Eds.; Academic Press: USA, 2000; pp 298–318.
- McFarland, J. W.; Gans, D. J. *Chemometric methods in molecular design*. In *Chapter Cluster Significance Analysis*; van Waterbeemd, H., Ed.; VCH: Weinheim, 1995; pp 295–307.
- Johnson, R. A.; Wichern, D. W. *Applied MultiVariate Statistical Analysis*; Prentice-Hall: New York, 1988.
- Goldberg, D. *Genetic Algorithms in Search*; Addison-Wesley: USA, 1989.
- Todeschini, R.; Ballabio, D.; Consonni, V.; Mauri, A.; Pavan, M. *Mobydigs Computer Software*, 1.0 ed.; 2004.
- García-Domenech, R.; Julian-Ortiz, J. V. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 445–449.
- Kubinyi, H. *Quant. Struct. Act. Relat.* **1994**, *13*, 285–294.
- Kubinyi, H. *Quant. Struct. Act. Relat.* **1994**, *13*, 393–401.
- Akaike, H. *Information theory and an extension of the maximum likelihood principle*. In *Proceedings of the Second International Symposium on Information Theory*; Akademiai Kiado, Budapest, 1973.
- Akaike, H. *IEEE Trans. Automat. Contr.* **1974**, *AC-19*, 716–723.
- Lucic, B.; Nikolic, S.; Trinajstic, N.; Juric, D. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 532–538.
- Klein, D.; Randić, M.; Babic, D.; Lucic, B.; Nikolic, S.; Trinajstic, N. *Int. J. Quantum Chem.* **1997**, *63*, 215–222.
- Randić, M. *N. J. Chem.* **1991**, *15*, 517–525.
- Randić, M. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 311–320.
- Randić, M. *J. Mol. Struct. (Theochem.)* **1991**, *233*, 45–59.
- Eriksson, L.; Jaworska, J.; Worth, A. P.; Cronin, M. T. D.; McDowell, R. M.; Gramatica, P. *Environmental Health Perspect.* **2003**, *111*, 1361–1375.
- Netzeva, T. I. et al. *ATLA* **2005**, *33*, 155–173.
- Gramatica, P. *QSAR Comb. Sci.* **2007**, *00*, 1–9.
- Vighi, M.; Gramatica, P.; Consolaro, F.; Todeschini, R. *Ecotoxicol. Environ. Saf.* **2001**, *49*, 206–220.
- Estrada, E.; Perdomo-López, I.; Torres-Labandeira, J. J. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1561–1568.
- Devillers, J.; Balaban, A. T. *Topological Indices and Related Descriptors in QSAR and QSPR*; Gordon and Breach Science Publishers: Australia, 1999.
- Devillers, J. *Topological indices and related descriptors in QSAR and QSPR*. In *Chapter no-free-lunch molecular descriptors in QSAR and QSPR*; Devillers, J., Balaban, A. T., Eds.; Gordon and Breach Publishers: Australia, 1999; pp 1–17.
- Ren, B. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 139–143.
- Rekharsky, M. V.; Inoue, Y. *Chem. Rev.* **1998**, *98*, 1875–1917.
- Liu, L.; Guo, Q.-X. *J. Inclusion Phenom. Macrocyclic Chem.* **2002**, *42*, 1–14.
- Inoue, Y.; Hakushi, T.; Liu, Y.; Tong, L.-H.; Shen, B.-J.; Jin, D.-S. *J. Am. Chem. Soc.* **1993**, *115*, 475–481.
- Carpignano, R.; Marzona, M.; Cattaneo, E.; Quaranta, S. *Anal. Chim. Acta* **1997**, *348*, 489–493.
- Rekharsky, M. V.; Goldberg, R. N.; Schwarz, F. P.; Tewari, Y. B.; Ross, P. D.; Yamashoji, Y.; Inoue, Y. *J. Am. Chem. Soc.* **1995**, *117*, 8830–8840.
- Wallimann, P.; Marti, T.; Fürer, A.; Diederich, F. *Chem. Rev.* **1997**, *97*, 1567–1608.

QSPR Modelling With the Topological Substructural Molecular Design Approach: β -Cyclodextrin Complexation

ALFONSO PÉREZ-GARRIDO,^{1,2} ALIUSKA MORALES HELGUERA,^{3,4,5} M. NATÁLIA D.S. CORDEIRO,⁵
AMALIO GARRIDO ESCUDERO¹

¹Environmental Engineering and Toxicology Department, Catholic University of San Antonio, Guadalupe, Murcia, C.P. 30107, Spain

²Department of Food and Nutrition Technology, Catholic University of San Antonio, Guadalupe, Murcia, C.P. 30107, Spain

³Faculty of Chemistry and Pharmacy, Department of Chemistry, Central University of Las Villas, Santa Clara, 54830 Villa Clara, Cuba

⁴Molecular Simulation and Drug Design Group, Chemical Bioactive Center, Central University of Las Villas, Santa Clara, 54830 Villa Clara, Cuba

⁵REQUIMTE, Faculty of Sciences, Chemistry Department, University of Porto, 4169-007 Porto, Portugal

Received 10 October 2008; revised 30 January 2009; accepted 11 February 2009

Published online 5 June 2009 in Wiley InterScience (www.interscience.wiley.com). DOI 10.1002/jps.21747

ABSTRACT: This study aims at developing a quantitative structure–property relationship (QSPR) model for predicting complexation with β -cyclodextrins (β -CD) based on a large variety of organic compounds. Molecular descriptors were computed following the TOPological Substructural MOlecular DEsign (TOPS-MODE) approach and correlated with β -CD complex stability constants by linear multivariate data analysis. This strategy afforded a final QSPR model that was able to explain around 86% of the variance in the experimental activity, along with showing good internal cross-validation statistics, and also good predictivity on external data. Topological substructural information influencing the complexation with β -CD was extracted from the QSPR model. This revealed that the major driving forces for complexation are hydrophobicity and van der Waals interactions. Therefore, the presence of hydrophobic groups (hydrocarbon chains, aryl groups, etc.) and voluminous species (Cl, Br, I, etc.) in the molecules renders easy their complexity with β -CDs. To our knowledge, this is the first time a correlation between TOPS-MODE descriptors and complexing abilities of β -CDs has been reported.

© 2009 Wiley-Liss, Inc. and the American Pharmacists Association *J Pharm Sci* 98:4557–4576, 2009

Keywords: QSPR; QSAR; drug design; cyclodextrins; complexation

INTRODUCTION

Cyclodextrins (CDs) are cyclic oligomers of β -D-glucose produced from starch by means of enzymatic conversion, and shaped like truncated cones

with primary and secondary hydroxyl groups crowning the narrower rim and wider rim, respectively.¹ CDs have attracted much interest in many fields, because they are able to form host–guest complexes with hydrophobic molecules and greatly modify their physical and chemical properties, mostly in terms of water solubility. For instance, upon complexation with CDs, the drugs solubility strongly increases, making them available for a wide range of pharmaceutical applications. Different drugs are currently marketed as solid or solution-based CD complex

Additional Supporting Information may be found in the online version of this article.

Correspondence to: Alfonso Pérez-Garrido (Telephone: +34-968278755; Fax: +34-968278622; E-mail: aperez@pdi.ucam.edu)

Journal of Pharmaceutical Sciences, Vol. 98, 4557–4576 (2009)

© 2009 Wiley-Liss, Inc. and the American Pharmacists Association

formulations.^{2,3} In these pharmaceutical products, CDs are mainly used as complexing agents to increase the aqueous solubility of poorly water-soluble drugs, to increase their bioavailability and stability.^{4–6} Poor solubility continues to impact the development of a large number of potential drug candidates.⁷ These factors have had a significant impact on what is required from formulators given that the number of formulation options, and by extension excipients, has to be increased to address the larger number of challenges being presented.⁸ CDs represent a true added value in this context.

In addition, CDs can also promote drug absorption across the dermal, nasal, or intestinal barrier by extracting cholesterol, phospholipids, or proteins from membranes,⁹ reduce or prevent gastrointestinal and ocular irritation, reduce or eliminate unpleasant smells or tastes,^{10,11} prevent drug–drug or drug–additive interactions, as well as to convert oils and liquid drugs into microcrystalline or amorphous powders.¹² Moreover, pharmacon–CD complexes often increase the bioavailability of the active substances and permit their controlled release.¹³ An example of the latter is the CD encapsulation of *trans*-platinum complex where it has been found that the cytotoxicity *in vitro* of the novel inclusion complex indicated a much higher activity.¹⁴

The experimental determination of CD complex binding constants is often difficult and time consuming because of the low solubility of the guest molecules in aqueous solution. Previous studies have suggested five major types of interactions: (i) hydrophobic interactions, (ii) van der Waals interactions, (iii) hydrogen-bonding between polar groups of the *guest* and the hydroxyl groups of the *host*, (iv) relaxation by release of high-energy water from the CD cavity upon substrate inclusion, and (v) relief of the conformational strain in a CD–water adduct.

In contrast, computational methods have only recently been used for predicting binding constants and to study the driving forces involved in the process. An exhaustive set of these computational applications has been excellently reviewed by Lipkowitz.¹⁵

Group-contribution models, quantitative structure–activity/property relationships (QSAR/QSPR) methods (2D-QSAR, 3D-QSAR, CoMFA), molecular modelling computations (using Quantum Mechanics, Monte Carlo/Molecular Dynamics Simulations, Molecular Mechanics, etc.), statistical

analysis tools, and artificial neural networks have all been applied to elucidate the most important factors influencing the host–guest interactions and to predict the thermodynamic stability of CDs inclusion complexes.^{16–25}

Nevertheless, it is clear that knowledge of the complexation abilities of guest molecules with CDs is deemed necessary to decide whether or not a host–guest complexation is useful in a particular application using the knowledge of what kind of bonds contribute positively to this phenomenon. In this sense, Katritzky et al.²⁵ presented a QSAR study predicting the free energies of inclusion complexation between diverse *guest* molecules and CDs using (i) CODESSA descriptors and (ii) counts of different molecular fragments. The fragmental descriptors are more easily interpretable than CODESSA descriptors. One can select the fragments whose contributions are considerable and give reasonable explanations based on physical phenomena involved in host–guest complexation. However, QSPR models based on fragments generally comprise much more variables than those using traditional descriptors, which still remain as an important problem.

The aim of the present study was to build a QSPR regression-based model, which could correlate and predict the complex stability constant between diverse guest molecules and β -CDs using the *TOP*ological *SUB*structural *MO*lecular *DES*ign (*TOPS-MODE*) descriptors.^{26–28} There is evidence that these descriptors performed well in similar QSAR/QSPR modelling studies on which they have been used because they are easy to calculate, and one can draw from the derived models useful information regarding the type of structures that contribute favourably or not to the activity or property.^{29–42} This approach is able to transform simple molecular descriptors, such as $\log P$, polar surface area, molar refraction, charges, etc., into series of descriptors that account for the distribution of these characteristics (hydrophobicity, polarity, steric effects, etc.) across the molecule. Thus, we can obtain this structural information at a local scale from the models developed using global molecular descriptors. It has been recognised that the *TOPS-MODE* approach “provides a mechanistic interpretation at a bond level and enables the generation of new hypotheses such as structural alerts.”⁴³ Such valuable information can then be used for the design of new drugs with increased bioavailability and solubility due to their complexation with β -CDs.

EXPERIMENTAL

Data Set

The overall data set of 233 substances comprised a large number of classes of organic compounds: aromatic hydrocarbons, alcohols, phenols, ethers, aldehydes, ketones, acids, esters, nitriles, anilines, halogenated compounds, heterocycles, nitro, sulphur and steroids and barbitals compounds. This set of guest molecules was extracted from the work of Suzuki,¹⁶ and the experimental endpoint to be predicted is the β -CD complex stability constants (K), which have been measured at $T = 298.15$ K using water as solvent, taken from references therein. Two of such guest molecules, chemicals 214 and 215, are stereoisomers, which could not be distinguished by the present 2D descriptors but had nevertheless different K values. Thus, one of the isomers was discarded (chemical 215), the other one (chemical 214) being only considered in our study with an averaged value of K . Moreover, all K values were log-transformed ($\log K$) for being of practical use in the following QSPR modelling. Table 1 displays a complete list of the chemicals along with the reported experimental data.

The TOPS-MODE Descriptors

The TOPS-MODE descriptors are based on the calculation of the spectral moments of the so-called bond matrix.²⁸ The theoretical foundations of the spectral moments have been reported previously,^{26,27} nevertheless an overview of this descriptor family will be given here. The spectral moments are defined as the traces of the bond adjacency matrix. That is, the sum of the main diagonal elements of different powers of such matrix. The bond adjacency matrix is a squared symmetric matrix whose entries are ones or zeros if the corresponding bonds are adjacent or not. The order of this matrix (m) is the number of bonds in the molecular graph, two bonds being adjacent if they are incident to a common atom. Furthermore, weights are introduced in the diagonal entries of this matrix to mirror fundamental physicochemical properties that might relate to the target endpoint being modelled. Here, several bond weights were used for computing the spectral moments, namely the standard bond distance (Std), standard bond dipole moments (Dip, Dip2), hydrophobicity (H), polar

surface area (Pols), polarisability (Pol), molar refractivity (Mol), van der Waals radii (Van), Gasteiger–Marsilli charges (Gas), atomic masses (Ato), solute excess molar refraction (Ab-R_2), solute dipolarity/polarisability ($\text{Ab-}\pi_2^{\text{H}}$), effective hydrogen-bond basicity ($\text{Ab-}\sum \beta_2^{\text{O}}$, $\text{Ab-}\sum \beta_2^{\text{H}}$) and solute gas-hexadecane partition coefficient ($\text{Ab-log}L$ ¹⁶) were used for computing the spectral moments of the bond matrix.

Explicitly, we have calculated the first 15 spectral moments (μ_1 – μ_{15}) for each bond weight and the number of bonds in the molecules (μ_0) without hydrogen. Also, we multiplied μ_0 and μ_1 for the first 15 spectral moments obtaining 30 new variables. Notice that in this way such variables might offset the linear approximation assumption of the model. As described previously,⁴⁴ the atomic contributions were then transformed into bond contributions as follows:

$$w_{(i,j)} = \frac{w_i}{\delta_i} + \frac{w_j}{\delta_j} \quad (1)$$

where w_i and δ_i are the atomic weight and vertex degree of the atom i . Calculation of the TOPS-MODE descriptors was carried out with the MODESLAB software (<http://www.modeslab.com>)⁴⁵ from the SMILES (Simplified Molecular Input Line Entry System) notation available for each compound.⁴⁶ To develop the structure–property relationships, the following six-step path was adopted:

1. Select a small subset of the 233 chemicals to act as a test set. The remaining chemicals form the training set for QSPR modelling.
2. Draw the molecular graphs for each molecule included in the training set.
3. Compute the spectral moment's descriptors using an appropriate set of weights.
4. Find an adequate QSPR model from the training set by a regression-based approach. The task here is to obtain a mathematical function (see Eq. 2 below) that best describes the studied property P (in our case, the $\log K$ partitioning) as a linear combination of the X -predictor variables (the spectral moments μ_k), with the coefficients a_k . Such coefficients are to be optimised by means of multiple linear regression (MLR) analysis along with a variable subset selection procedure

$$P = a_0\mu_0 + a_1\mu_1 + a_2\mu_2 + \dots + a_k\mu_k \quad (2)$$

Table 1. Names, CAS Number, Observed ($\log K_o$) and Predicted ($\log K_p$) Activity,* and Leverage (h) Values for the Compounds Used in this Study

No.	Name	CAS	$\log K_o$	$\log K_p$	Partition	h	Refs.
1	Carbon tetrachloride	56-23-5	2.20	2.29	Test	0.174 ^b	20
2	Chloroform	67-66-3	1.43	0.66	Training	—	20
3	Methanol	67-56-1	-0.49	-0.35	Training	—	18
4	Acetonitrile	75-05-8	-0.27	-0.36	Training	—	20
5	Acetaldehyde	75-07-0	-0.64	-0.23	Training	—	20
6	Ethanol	64-17-5	-0.03	0.19	Test	0.055	18
7	1,2-Ethanediol	107-21-1	-0.19	-0.02	Training	—	18
8	Acetone	67-64-1	0.42	0.40	Training	—	20
9	1-Propanol	71-23-8	0.57	0.69	Test	0.039	18
10	2-Propanol	67-63-0	0.63	0.93	Training	—	20
11	1,3-Propanediol	504-63-2	0.67	0.46	Training	—	18
12	Tetrahydrofuran	109-99-9	1.47	1.10	Test	0.034	20
13	Cyclobutanol	2919-23-5	1.18	1.51	Training	—	18
14	1-Butanol	71-36-3	1.22	1.18	Training	—	79
15	2-Butanol	78-92-2	1.19	1.37	Training	—	18
16	2-Methyl-1-propanol	78-83-1	1.62	1.37	Training	—	18
17	2-Methyl-2-propanol	75-65-0	1.68	2.01	Training	—	18
18	1,4-Butanediol	110-63-4	0.64	0.93	Training	—	18
19	Diethylamine	109-89-7	1.36	1.22	Training	—	20
20	Cyclopentanol	96-41-3	2.08	1.70	Training	—	18
21	1-Pentanol	71-41-0	1.80	1.64	Training	—	79
22	2-Pentanol	6032-29-7	1.49	1.83	Training	—	18
23	3-Pentanol	584-02-1	1.35	1.78	Training	—	18
24	2-Methyl-1-butanol	1565-80-6	2.08	1.80	Training	—	18
25	2-Methyl-2-butanol	75-85-4	1.91	2.29	Training	—	18
26	3-Methyl-1-butanol	123-51-3	2.25	1.85	Test	0.023	18
27	3-Methyl-2-butanol	1517-66-4	1.92	1.92	Training	—	18
28	2,2-Dimethyl-1-propanol	75-84-3	2.71	2.39	Test	0.027	79
29	1,5-Pentanediol	111-29-5	1.22	1.38	Test	0.026	18
30	1,4-Dibromobenzene	106-37-6	2.97	2.78	Training	—	22
31	1,4-Diiodobenzene	624-38-4	3.17	3.63	Training	—	22
32	3,5-Dibromophenol	626-41-5	2.56	2.79	Training	—	18
33	3,5-Dichlorophenol	591-35-5	2.07	2.53	Test	0.045	18
34	1-Chloro-4-nitrobenzene	100-00-5	2.15	2.52	Training	—	22
35	Fluorobenzene	462-06-6	1.96	2.02	Test	0.026	22
36	Bromobenzene	108-86-1	2.50	2.42	Training	—	22
37	Iodobenzene	591-50-4	2.93	2.87	Training	—	22
38	3-Fluorophenol	372-20-3	1.70	2.03	Training	—	18
39	4-Fluorophenol	371-41-5	1.73	2.02	Training	—	18
40	3-Chlorophenol	108-43-0	2.28	2.28	Training	—	18
41	4-Chlorophenol	106-48-9	2.61	2.27	Training	—	22
42	3-Bromophenol	591-20-8	2.51	2.42	Test	0.031	18
43	4-Bromophenol	106-41-2	2.65	2.41	Test	0.031	22
44	3-Iodophenol	626-02-8	2.93	2.85	Training	—	18
45	4-Iodophenol	540-38-5	2.98	2.84	Training	—	22
46	Nitrobenzene	98-95-3	2.04	2.32	Training	—	20
47	4-Nitrophenol	100-02-7	2.39	2.29	Training	—	22
48	Benzene	71-43-2	2.23	2.05	Test	0.033	79
49	Phenol	108-95-2	1.98	2.05	Training	—	22
50	Hydroquinone	123-31-9	2.05	2.04	Test	0.024	22
51	4-Nitroaniline	100-01-6	2.48	2.35	Test	0.082	22

Table 1. (Continued)

No.	Name	CAS	$\log K_o$	$\log K_p$	Partition	h	Refs.
52	Aniline	62-53-3	1.60	2.12	Training	—	20
53	Sulphaanilamide	63-74-1	2.76	2.16	Training	—	21
54	Cyclohexanol	108-93-0	2.67	2.24	Training	—	79
55	1-Hexanol	111-27-3	2.33	2.09	Training	—	79
56	2-Hexanol	626-93-7	1.98	2.27	Training	—	18
57	2-Methyl-2-pentanol	590-36-3	1.99	2.73	Training	—	18
58	3-Methyl-3-pentanol	77-74-7	2.15	2.56	Training	—	18
59	4-Methyl-2-pentanol	108-11-2	2.04	2.48	Training	—	18
60	3,3-Dimethyl-2-butanol	464-07-3	2.75	2.94	Training	—	18
61	1,6-Hexanediol	629-11-8	1.69	1.80	Training	—	18
62	Benzonitrile	100-47-0	2.23	1.81	Training	—	22
63	Benzothiazole	95-16-9	2.38	1.92	Training	—	80
64	4-Nitrobenzoic acid	62-23-7	2.34	2.23	Training	—	22
65	Benzaldehyde	100-52-7	1.78	1.79	Training	—	20
66	Benzoic acid	65-85-0	2.12	2.03	Training	—	79
67	4-Hydroxybenzaldehyde	123-08-0	1.75	1.78	Training	—	18
68	4-Hydroxybenzoic acid	99-96-7	2.20	2.02	Training	—	22
69	Benzyl chloride	100-44-7	2.45	2.36	Training	—	22
70	Toluene	108-88-3	2.09	2.50	Training	—	20
71	Benzyl alcohol	100-51-6	1.71	2.05	Training	—	20
72	Anisole	100-66-3	2.32	2.12	Training	—	22
73	<i>m</i> -Cresol	108-39-4	1.98	2.49	Training	—	18
74	<i>p</i> -Cresol	106-44-5	2.40	2.48	Training	—	22
75	4-Methoxyphenol	150-76-5	2.21	2.10	Training	—	22
76	3-Methoxyphenol	150-19-6	2.11	2.11	Training	—	18
77	4-Hydroxybenzyl alcohol	623-05-2	2.16	2.01	Training	—	22
78	Hydrochlorothiazide	58-93-5	1.76	1.74	Training	—	21
79	<i>N</i> -methylaniline	100-61-8	2.12	2.14	Training	—	22
80	1-Butylimidazole	4316-42-1	2.19	2.27	Training	—	81
81	1-Heptanol	111-70-6	2.85	2.51	Test	0.026	18
82	Phenylacetylene	536-74-3	2.36	2.62	Training	—	22
83	Thianaphthene	95-15-8	3.23	2.49	Training	—	80
84	4-Fluorophenyl acetate	405-51-6	2.11	2.22	Test	0.027	18
85	3-Fluorophenyl acetate	701-83-7	1.91	2.23	Training	—	18
86	4-Chlorophenyl acetate	876-27-7	2.50	2.46	Training	—	18
87	3-Chlorophenyl acetate	13031-39-5	2.44	2.47	Training	—	18
88	4-Bromophenyl acetate	1927-95-3	2.68	2.59	Training	—	18
89	3-Bromophenyl acetate	35065-86-2	2.67	2.60	Test	0.032	18
90	4-Iodophenyl acetate	33527-94-5	3.00	2.93	Training	—	18
91	3-Iodophenyl acetate	61-71-2	3.07	2.94	Training	—	18
92	4-Nitrophenyl acetate	830-03-5	2.13	2.39	Training	—	18
93	Acetophenone	98-86-2	2.27	2.20	Training	—	22
94	Phenyl acetate	122-79-2	2.10	2.22	Training	—	18
95	Methyl benzoate	93-58-3	2.50	2.12	Training	—	22
96	3-Hydroxyacetophenone	121-71-1	2.06	2.19	Training	—	18
97	4-Hydroxyacetophenone	99-93-4	2.18	2.18	Training	—	22
98	Acetoanilide	103-84-4	2.20	1.92	Test	0.018	22
99	<i>p</i> -Xylene	106-42-3	2.38	2.92	Training	—	22
100	Ethylbenzene	100-41-4	2.59	2.80	Training	—	22
101	Phenetole	103-73-1	2.49	2.67	Training	—	22
102	2-Phenylethanol	60-12-8	2.15	2.48	Training	—	18
103	3-Ethylphenol	620-17-7	2.60	2.76	Training	—	18

(Continued)

Table 1. (Continued)

No.	Name	CAS	log K_o	log K_p	Partition	h	Refs.
104	4-Ethylphenol	123-07-9	2.69	2.75	Training	—	22
105	4-Ethoxyphenol	622-62-8	2.33	2.62	Test	0.008	18
106	3-Ethoxyphenol	621-34-1	2.35	2.63	Test	0.008	18
107	3,5-Dimethoxyphenol	500-99-2	2.34	2.13	Training	—	18
108	<i>N</i> -ethylaniline	103-69-5	2.34	2.67	Test	0.013	22
109	<i>N,N</i> -dimethylaniline	121-69-7	2.36	2.49	Training	—	22
110	Barbital	57-44-3	1.78	2.05	Training	—	21
111	Cyclooctanol	696-71-9	3.30	3.00	Training	—	18
112	1-Octanol	111-87-5	3.17	2.92	Test	0.033	18
113	2-Octanol	123-96-6	3.13	3.09	Training	—	18
114	Quinoline	91-22-5	2.12	2.37	Training	—	80
115	3-Cyanophenyl acetate	55682-11-6	1.49	2.14	Training	—	18
116	4-Hydroxycinnamic acid	7400-08-0	2.83	2.51	Training	—	21
117	Ethyl benzoate	93-89-0	2.73	2.63	Training	—	22
118	4'-Hydroxypropiofenone	70-70-2	2.63	2.43	Training	—	18
119	3'-Hydroxypropiofenone	13103-80-5	2.61	2.44	Training	—	18
120	<i>p</i> -Tolyl acetate	140-39-6	2.49	2.60	Training	—	18
121	3-Methylphenyl acetate	122-46-3	2.21	2.61	Training	—	18
122	4-Methoxyphenyl acetate	1200-06-2	2.45	2.24	Training	—	18
123	4-Propylphenol	645-56-7	3.55	3.11	Training	—	18
124	3-Propylphenol	621-27-2	3.28	3.12	Training	—	18
125	4-Isopropylphenol	99-89-8	3.58	3.15	Training	—	18
126	3-Isopropylphenol	618-45-1	3.44	3.16	Training	—	18
127	4-Isopropoxyphenol	7495-77-4	2.86	3.11	Training	—	18
128	2-Norbornaneacetate	—	3.59	3.11	Test	0.052	79
129	1-Benzylimidazole	4238-71-5	2.61	2.75	Training	—	81
130	<i>m</i> -Methylcinnamic acid	3029-79-6	2.93	2.92	Training	—	21
131	4-Ethylphenyl acetate	3245-23-6	2.83	2.82	Test	0.017	18
132	3-Ethylphenyl acetate	3056-60-8	2.68	2.83	Training	—	18
133	4-Ethoxyphenyl acetate	69788-77-8	2.54	2.72	Training	—	18
134	3-Ethoxyphenyl acetate	151360-54-2	2.49	2.73	Test	0.023	18
135	Allobarbital	52-43-7	1.98	2.15	Training	—	21
136	4- <i>n</i> -butylphenol	1638-22-8	3.97	3.46	Test	0.023	18
137	3- <i>n</i> -butylphenol	4074-43-5	3.76	3.46	Training	—	18
138	3-Isobutylphenol	30749-25-8	4.21	3.65	Training	—	18
139	4- <i>sec</i> -butylphenol	99-71-8	4.18	3.46	Training	—	18
140	3- <i>sec</i> -butylphenol	3522-86-9	4.06	3.47	Training	—	18
141	4- <i>tert</i> -butylphenol	98-54-4	4.56	3.84	Test	0.045	18
142	3- <i>tert</i> -butylphenol	585-34-2	4.41	3.85	Training	—	18
143	Menadion	58-27-5	2.27	2.30	Test	0.027	21
144	Sulphapyridine	144-83-2	2.70	2.57	Training	—	21
145	Sulphamonomethoxine	1220-83-3	2.48	2.20	Training	—	21
146	Sulfisoxazole	127-69-5	2.32	2.69	Training	—	21
147	4- <i>n</i> -propylphenyl acetate	61824-46-2	3.15	3.13	Training	—	18
148	3- <i>n</i> -propylphenyl acetate	—	3.28	3.14	Training	—	18
149	4-Isopropylphenyl acetate	2664-32-6	2.88	3.16	Training	—	18
150	3-Isopropylphenyl acetate	36438-57-0	3.36	3.17	Training	—	18
151	4- <i>n</i> -amylphenol	14938-35-3	4.19	3.78	Test	0.031	18
152	4- <i>tert</i> -amylphenol	80-46-6	4.70	4.02	Training	—	18
153	Carbutamide	339-43-5	2.29	2.37	Training	—	21
154	Pentobarbital	76-74-4	3.01	3.16	Test	0.069	21
155	Amobarbital	57-43-2	3.07	3.07	Training	—	79

Table 1. (Continued)

No.	Name	CAS	$\log K_o$	$\log K_p$	Partition	h	Refs.
156	Thiopental	76-75-5	3.28	3.12	Training	—	21
157	Dibenzofuran	132-64-9	2.97	2.60	Training	—	80
158	Dibenzothiophene	132-65-0	3.48	2.86	Training	—	80
159	Phenazine	92-82-0	2.41	2.03	Training	—	80
160	Thianthrene	92-85-3	3.57	3.48	Test	0.091	80
161	Carbazole	86-74-8	2.44	3.01	Training	—	80
162	Phenoxazine	135-67-1	2.69	2.75	Test	0.060	80
163	Phenothiazine	92-84-2	2.73	3.08	Training	—	80
164	Furosemide	200-203-6	1.78	3.02	Test	0.071	21
165	Phenobarbital	50-06-6	3.22	2.70	Test	0.062	79
166	Sulfisomidine	515-64-0	2.10	2.66	Test	0.061	21
167	Sulphamethomidine	3772-76-7	2.33	2.46	Test	0.038	21
168	Sulphadimethoxine	122-11-2	2.26	2.20	Training	—	21
169	4- <i>n</i> -butylphenyl acetate	55168-27-9	3.62	3.42	Training	—	18
170	3- <i>n</i> -butylphenyl acetate	—	3.66	3.43	Training	—	18
171	3-Isobutylphenyl acetate	916728-77-3	3.83	3.62	Training	—	18
172	4- <i>tert</i> -butylphenyl acetate	3056-64-2	3.85	3.83	Training	—	18
173	Cyclobarbitol	52-31-3	2.71	2.55	Training	—	21
174	Hexobarbital	56-29-1	3.08	2.86	Training	—	21
175	1-Adamantaneacetate	875907-32-7	4.32	4.56	Training	—	79
176	Acridine	260-94-6	2.33	2.70	Training	—	80
177	Phenanthridine	229-87-8	2.57	2.61	Training	—	80
178	Xanthene	92-83-1	2.71	3.32	Training	—	80
179	<i>N</i> -phenylantranilic acid	91-40-7	2.89	3.06	Training	—	79
180	Mephobarbital	115-38-8	3.16	3.03	Training	—	21
181	4- <i>n</i> -amylphenyl acetate	202831-79-6	3.80	3.69	Training	—	18
182	Flufenamic acid	530-78-9	3.10	3.08	Training	—	79
183	Meclofenamic acid	644-62-2	2.67	3.15	Training	—	79
184	Nitrazepam	146-22-5	1.97	1.70	Training	—	21
185	Flurbiprofen	5104-49-4	3.69	3.48	Training	—	21
186	Sulphaphenazole	526-08-9	2.35	1.97	Training	—	21
187	Bendroflumethiazide	200-800-1	1.90	2.02	Training	—	21
188	Mefenamic acid	61-68-7	2.49	3.26	Training	—	21
189	Acetohexamide	968-81-0	2.94	2.62	Test	0.047	21
190	Fludiazepam	3900-31-0	2.33	1.76	Training	—	21
191	Nimetazepam	2011-67-8	1.73	1.63	Training	—	21
192	Fenbufen	252-979-0	2.63	3.26	Training	—	21
193	Ketoprofen	22071-15-4	2.85	3.26	Training	—	21
194	Medazepam	2898-12-6	2.40	2.98	Training	—	21
195	Progabide	62666-20-0	2.53	2.80	Test	0.084	21
196	Griseofulvin	126-07-8	1.47	1.91	Training	—	21
197	Tolnaftate	2398-96-1	3.83	3.68	Training	—	21
198	Prostacyclin	35121-78-9	2.94	3.26	Training	—	21
199	Triamcinolone	124-94-7	3.37	3.92	Test	0.095	82
200	Cortisone	53-06-5	3.35	3.19	Training	—	21
201	Prednisolone	50-24-8	3.56	3.36	Test	0.105	82
202	Hydrocortisone	50-23-7	3.60	3.42	Training	—	21
203	Corticosterone	50-22-6	3.85	3.17	Test	0.106	82
204	Dexamethasone	50-02-2	3.65	4.48	Training	—	21
205	Betamethasone	378-44-9	3.73	4.03	Training	—	82
206	Paramethasone	53-33-8	3.40	3.32	Training	—	82
207	Cortisone-21-acetate	50-04-4	3.62	3.26	Training	—	82

(Continued)

Table 1. (Continued)

No.	Name	CAS	$\log K_o$	$\log K_p$	Partition	h	Refs.
208	Prednisolone-21-acetate	52-21-1	3.76	3.21	Training	—	82
209	Hydrocortisone-21-acetate	50-03-3	3.51	3.35	Training	—	82
210	Fluocinolone acetonide	67-73-2	3.48	3.60	Training	—	82
211	Triamcinolone acetonide	76-25-5	3.51	3.95	Training	—	82
212	Spirolactone	52-01-7	4.44	4.20	Training	—	82
213	Dehydrocholic acid	81-23-2	3.38	3.45	Training	—	82
214	Chenodeoxycholic acid	474-25-9	4.36 ^a	4.45	Training	—	82
215	Ursodeoxycholic acid	128-13-2	4.51 ^a	—	Training	—	82
216	Cholic acid	81-25-4	3.50	3.90	Test	0.184 ^b	82
217	Hydrocortisone-17-butyrate	237-093-4	3.23	3.04	Training	—	82
218	Cinnarizine	298-57-7	3.64	3.24	Training	—	21
219	Cycloheptanol	502-41-0	3.23	2.63	Training	—	18
220	2-Methoxyethanol	109-86-4	0.22	0.29	Test	0.051	18
221	3-Hydroxycinnamic acid	588-30-7	2.56	2.52	Training	—	21
222	Ethyl 4-hydroxybenzoate	120-47-8	3.01	2.59	Training	—	21
223	Ethyl 4-aminobenzoate	94-09-7	2.69	2.64	Test	0.036	21
224	4-Methylcinnamic acid	1866-39-3	2.65	2.91	Training	—	21
225	Sulphadiazine	68-35-9	2.52	2.01	Test	0.057	21
226	L- α -O-benzylglycerol	213458-77-6	2.11	2.55	Test	0.051	81
227	Sulphamerazine	127-79-7	1.97	2.30	Training	—	21
228	Butyl 4-hydroxybenzoate	94-26-8	3.39	3.20	Training	—	21
229	Butyl 4-aminobenzoate	94-25-7	3.19	3.25	Training	—	21
230	Benzidine	92-87-5	3.35	3.64	Test	0.111	21
231	Triflumizole	68694-11-1	2.66	2.28	Training	—	21
232	Diazepam	439-14-5	2.33	1.97	Training	—	21
233	Prostaglandin E2	363-24-6	3.09	3.18	Training	—	21

* β -CD complex stability constant (K_o , observed, K_p , predicted), then log-transformed ($\log K$).

^aChemicals 214 and 215 were replaced by only one compound (chemical 214) with an averaged $\log K$ value (=4.44).

^bChemicals 1 and 216 have leverage values above the threshold (0.129) and, for that reason, its predictions were not taken into account when calculating Q_{EXT}^2 .

- Subject the derived QSPR model to rigorous internal and external validation, thereby assessing the performance of the model in what concerns its applicability and predictive power.
- Compute the contribution of the different substructures to determine their quantitative contribution to the complexation of the studied molecules.

Variable Selection

Nowadays, there is a vast amount and wide range of molecular descriptors with which one can model the activity of interest. This makes the search for gathering the most suitable subset quite complicated and time consuming because of the many possible combinations, especially if one tries to define an accurate, robust, and (above all) interpretable model. For this reason, we applied the Genetic Algorithm (GA) procedure⁴⁷ for

selecting the variables, as implemented in the Mobydigs software (v1.0).⁴⁸ The particular GA simulation applied here resorted to the generation of 100 regression models, ordered according to their increased internal predictive performance (verified by leave one out cross-validation). First of all, models with one to two variables were developed by the variable subset selection procedure in order to explore all low combinations. The number of descriptors was subsequently increased one by one, and new models formed. The GA was stopped when further increments in the size of the model did not increase internal predictivity in any significant degree. Furthermore, the following GA simulation conditions were used: the maximum number of variables in a model was 10, the number of best retained models for each size was 5, the trade-off between crossovers and mutation parameter (T) was from 0.3 to 0.7, and selection bias ($B\%$) was from 30 to 90.

Model Validation

Two kinds of diagnostic statistical tools were used for evaluating the performance of our regression model: the so-called goodness of fit and goodness of the prediction. In the first case, attention is given to the fitting properties of the model, whereas in the second case attention is paid to the predictive power of the model (i.e., the model adequacy for describing new compounds). In this work, k -Means Cluster Analysis (k -MCA) was used to split the original data set of chemicals into training and test sets. On doing so, 186 of the 233 compounds were selected as the training set and the remaining 47 taken as the external test set. Full details of this partition can be found in our previous work.⁴⁹

Goodness of fit of the models was assessed by examining the determination coefficient (R^2), the standard deviation (s), the Fisher's ratio (F), and the ratio between the number of cases and the number of adjustable parameters in the model (known as the ρ statistics; notice that ρ should be ± 4).⁵⁰ Other important statistics, namely the Kubinyi function (FIT)^{51,52} and Akaike's information criteria (AIC)^{53,54} were taken into account, as they give enough criteria for comparing models with different parameters, numbers of variables, and numbers of chemicals.

As to the robustness and predictivity of the models, these were evaluated by means of cross-validation, basically leave-one-out (CV-LOO) and bootstrapping testing techniques, by looking to the outcome statistics of both techniques (i.e., Q_{LOO}^2 and Q_{boot}^2) as well as to the Q_{EXT}^2 values obtained with the test set substances that fall within the applicability domain of the model. Bootstrapping simulates what would happen if the data set were to be randomly resampled several times (here 5000 times), then deriving the all squared difference between the true and predicted responses by using predictive residual sum of squares (PRESS). The average predictive power is expressed as Q_{boot}^2 .⁵⁵ Further, the stability under heavy perturbations in the training set was checked by examining the outcome statistics of a response randomisation procedure (Y scrambling) for the training and test sets ($\alpha(r^2)$ and $\alpha(Q^2)$ values). The randomisation procedure was repeated 300 times. All these calculations were carried out with software Mobydigs (v1.0).⁴⁸

To sum up, good quality of the models is indicated by high F , FIT, and ρ values, by low s and AIC values, as well as by values closed to one

for R^2 , Q_{LOO}^2 , Q_{boot}^2 , and Q_{EXT}^2 (save for $\alpha(r^2)$ and $\alpha(Q^2)$ values, which check random correlations).

The spectral moments are inherently collinear. From the point of view of QSPR modelling, the main drawback of collinearity is that it increases the standard errors associated with the individual regression coefficients, thereby decreasing their value for purposes of interpretability. To overcome this problem, we have employed here the Randić's method of orthogonalisation.⁵⁶⁻⁶⁰ Firstly, one has to select the appropriate order of orthogonalisation, which, in this case, is the order of significance of the variables in the model. The first variable (v_1) is taken as the first orthogonal descriptor ($\Omega^1 v_1$). The second one (v_2) is orthogonalised with respect to it by taking the residual of its correlation with $\Omega^1 v_1$. The process is repeated until all variables are completely orthogonalised, after which they are further standardised. Orthogonal standardised variables are then used to obtain a new model. For extracting of the information contained in the orthogonalised descriptors, we followed the procedure reported by Estrada and Molina.²⁹

Structural Alerts Identification

The identification of structural alerts (fragment contribution) to the β -CD complexation is based on bond contributions. This procedure, implemented in MODESLAB software, consists in transforming a QSAR/QSPR model into a bond additive scheme. Then, by summing up bonds contributions, one can detect the fragments on a given molecule that contribute positively or negatively to the underlying property and forward an interpretation of their effects in terms of physicochemical properties. Bond contributions are derived from the local spectral moments. They are defined as the diagonal entries of the different powers of the weighted bond matrix (B):

$$\mu_k^T(i) = b_{ii}^k(T) \quad (3)$$

where $\mu_k^T(i)$ is the k th local spectral moment of the bond i , $b_{ii}^k(T)$ are the diagonal entries of the weighted B matrix, and T is the type of bond weight. For a given molecule, we can substitute the values of the local spectral moments computed by Eq. (3) into Eq. (4) and thus gather the total contribution to the complexation of its different bonds

$$P = b_0 + \sum_k a_k \mu_k^T \quad (4)$$

Since the activity modelled is expressed as $\log K$, positive bond contributions increase

the K value and increase the complexation and vice versa. The structural information highlighted by the bond contributions may allow, along with other theoretical and experimental data, for a better understanding of the mechanisms of complexation of the involved chemicals.

Applicability Domain of the Models

Given that the real utility of a QSAR/QSPR model relies on its ability to accurately predict the modelled activity/property for new chemicals, careful assessment of the model's true predictive power is a must. This includes the model validation but also the definition of the applicability domain of the model in the space of

of Q_{EXT}^2 were performed only for those substances that had a leverage value below the threshold h^* .

RESULTS AND DISCUSSION

QSPR Model

According to the strategy outlined before, we began by seeking the best linear model relating the complex stability with the TOPS-MODE descriptors for the training set. The resulting best-fit model (a 11-variable equation) is given below along with the MLR statistics. As seen, this model is good both statistical significance and goodness of fit.

$$\begin{aligned} \log K = & -1.44 \times 10^{-3} (\pm 7.34 \times 10^{-5}) \mu_1 \mu_2^{\text{Std}} + 3.95 \times 10^{-7} (\pm 2.14 \times 10^{-8}) \mu_{10}^{\text{Std}} \\ & - 1.50 \times 10^{-2} (\pm 1.18 \times 10^{-3}) \mu_5^{\text{Ab-R}_2} + 0.42 (\pm 3.57 \times 10^{-2}) \mu_1^{\text{Hyd}} \\ & - 0.25 (\pm 2.49 \times 10^{-2}) \mu_1^{\text{Dip}^2} + 1.10 \times 10^{-2} (\pm 1.52 \times 10^{-3}) \mu_3^{\text{Van}} \\ & + 2.42 \times 10^{-4} (\pm 3.90 \times 10^{-5}) \mu_1 \mu_4^{\text{Dip}^2} + 9.33 \times 10^{-3} (\pm 1.68 \times 10^{-3}) \mu_4^{\text{Ab-log } L^{16}} \\ & + 1.50 \times 10^{-2} (\pm 3.21 \times 10^{-3}) \mu_4^{\text{Ab-}\sum \beta_2^0} + 5.03 \times 10^{-7} (\pm 1.62 \times 10^{-7}) \mu_4^{\text{Pols}} \\ & - 0.55 (\pm 0.12) \end{aligned} \quad (5)$$

$$N = 185, \quad R^2 = 0.870, \quad Q_{\text{LOO}}^2 = 0.849, \quad s = 0.329, \quad F = 116.76, \quad \text{AIC} = 0.122, \quad \text{FIT} = 4.106,$$

$$Q_{\text{boot}}^2 = 0.825, \quad \alpha(r^2) = 0.021, \quad \alpha(Q^2) = -0.114, \quad Q_{\text{EXT}}^2 = 0.827$$

molecular descriptors used for deriving the model. There are several methods for assessing the applicability domain of QSAR/QSPR models,^{61,62} but the most common one encompasses determining the leverage values for each compound.⁶³ A Williams plot, that is, the plot of standardised residuals versus leverage values (h), can then be used for an immediate and simple graphical detection of both the response outliers and structurally influential chemicals in the model. In this plot, the applicability domain is established inside a squared area within $\pm x$ standard deviations and a leverage threshold h^* (h^* is generally fixed at $3\kappa/n$, where n is the number of training compounds and κ the number of model parameters, whereas $x=2$ or 3), lying outside this are (vertical lines) the outliers and (horizontal lines) the influential chemicals. For future predictions, only predicted complex stability constant data for chemicals belonging to the chemical domain of the training set should be proposed and used.⁶⁴ So, calculations

Another aspect deserving special attention is the degree of collinearity of the variables of the model, which can readily be diagnosed by analysing the cross-correlation matrix (Tab. 2). Rather than deleting any of these descriptors, it is of interest to examine the performance of orthogonal complements.

Following Randić's technique, we determined orthogonal complements for all variables of the above nonorthogonalised model. On doing so, variables $\Omega^2 \mu_{10}^{\text{Std}}$ and $\Omega^3 \mu_5^{\text{Ab-R}_2}$ were found to be not statistically significant ($p=0.189$ and 0.496 ; Tab. 3), most likely because the information contained in these variables is common to the information contained in other molecular descriptors. In addition, the significance of adding these two variables to the model remains unclear as seen from the modest improvement in R^2 on going from step 8 to step 9 and to step 10 (see in Tab. 3, ΔR^2 for those steps). So after eliminating these uninformative variables, further proceed to refitting and

Table 2. Correlation matrix for Intercorrelations among the 10 Variables of the Initial Model (Eq. 5)

	μ_{10}^{Std}	μ_1^{Dip2}	μ_1^{Hyd}	μ_4^{Pols}	μ_3^{Van}	$\mu_5^{\text{Ab-R}_2}$	$\mu_4^{\text{Ab-}\sum\beta_2^0}$	$\mu_4^{\text{Ab-log}L^{16}}$	$\mu_1\mu_2^{\text{Std}}$	$\mu_1\mu_4^{\text{Dip2}}$
μ_{10}^{Std}	1.000	—	—	—	—	—	—	—	—	—
μ_1^{Dip2}	0.737	1.000	—	—	—	—	—	—	—	—
μ_1^{Hyd}	-0.205	-0.062	1.000	—	—	—	—	—	—	—
μ_4^{Pols}	0.405	0.352	-0.333	1.000	—	—	—	—	—	—
μ_3^{Van}	0.885	0.801	0.059	0.516	1.000	—	—	—	—	—
$\mu_5^{\text{Ab-R}_2}$	0.974	0.760	-0.142	0.509	0.945	1.000	—	—	—	—
$\mu_4^{\text{Ab-}\sum\beta_2^0}$	0.941	0.782	-0.062	0.532	0.980	0.989	1.000	—	—	—
$\mu_4^{\text{Ab-log}L^{16}}$	0.929	0.761	-0.033	0.540	0.980	0.984	0.997	1.000	—	—
$\mu_1\mu_2^{\text{Std}}$	0.935	0.770	0.003	0.442	0.965	0.952	0.966	0.962	1.000	—
$\mu_1\mu_4^{\text{Dip2}}$	0.921	0.894	-0.152	0.376	0.878	0.916	0.904	0.885	0.907	1.000

Significant correlations are marked in bold.

orthogonalisation, the following QSPR model was obtained:

As can be seen in Table 3, removal of $\Omega^2\mu_{10}^{\text{Std}}$ and $\Omega^3\mu_5^{\text{Ab-R}_2}$ had little effect on the overall

$$\begin{aligned} \log K = & 0.379(\pm 0.024)\Omega^1\mu_1\mu_2^{\text{Std}} + 0.509(\pm 0.024)\Omega^4\mu_1^{\text{Hyd}} - 0.063(\pm 0.025)\Omega^5\mu_1^{\text{Dip2}} \\ & + 0.475 \times 10^{-2}(\pm 0.024)\Omega^6\mu_3^{\text{Van}} + 0.080(\pm 0.025)\Omega^7\mu_1\mu_4^{\text{Dip2}} + 0.177(\pm 0.025)\Omega^8\mu_4^{\text{Ab-log}L^{16}} \\ & + 0.105(\pm 0.024)\Omega^9\mu_4^{\text{Ab-}\sum\beta_2^0} + 0.078(\pm 0.025)\Omega^{10}\mu_4^{\text{Pols}} + 2.537(\pm 0.024) \end{aligned} \quad (6)$$

$$N = 185, \quad R^2 = 0.868, \quad Q_{\text{LOO}}^2 = 0.851, \quad s = 0.329, \quad F = 145.61, \quad \text{AIC} = 0.12, \quad \text{FIT} = 4.656,$$

$$Q_{\text{boot}}^2 = 0.845, \quad \alpha(r^2) = 0.007, \quad \alpha(Q^2) = -0.1, \quad Q_{\text{EXT}}^2 = 0.8341$$

where the symbol ${}^i\Omega X$ means the orthogonal complement of variable X , the superscript referring to followed order in the orthogonalisation process.

Table 3. Step-by-Step Analysis of the Forward Stepwise Process

Step	Variable Included	R^2	ΔR^2	p -Level
1	${}^4\Omega\mu_1^{\text{Hyd}}$	0.323	0.323	3.25×10^{-17}
2	${}^6\Omega\mu_3^{\text{Van}}$	0.608	0.285	$2.56 \times 10^{-}$
3	$\Omega\mu_1\mu_2^{\text{Std}}$	0.798	0.191	6.26×10^{-28}
4	${}^8\Omega\mu_4^{\text{Ab-log}L^{16}}$	0.833	0.035	5.21×10^{-9}
5	${}^9\Omega\mu_4^{\text{Ab-}\sum\beta_2^0}$	0.848	0.015	5.59×10^{-5}
6	${}^{10}\Omega\mu_4^{\text{Pols}}$	0.855	0.008	2.12×10^{-3}
7	${}^7\Omega\mu_1\mu_4^{\text{Dip2}}$	0.863	0.008	1.65×10^{-3}
8	${}^5\Omega\mu_1^{\text{Dip2}}$	0.869	0.005	8.35×10^{-3}
9	${}^2\Omega\mu_{10}^{\text{Std}}$	0.870	1.3×10^{-3}	0.189
10	${}^3\Omega\mu_5^{\text{Ab-R}_2}$	0.870	3.5×10^{-4}	0.496

Significant correlations are marked in bold.

fitness of the model as the statistics are as robust as before, and further, by comparing Eq. (5) with Eq. (6), one can see that there are no changes in either the sign of the regression coefficients. Nevertheless, the relative contributions of the variables in the orthogonal model are quite different from those related to the nonorthogonal model.

Their direct interpretation of these complex topological indices is rather difficult, considering that they essentially condense a large amount of topological and atomic property information into a single number. However, some indirect links between those descriptors and the physical phenomena involved in host-guest complexation might be suggested.

The variables weighted with hydrophobicity and van der Waals radii explained, respectively, 32.3% and 28.5% of the variance for this specific training set of chemicals (Tab. 3). Thus, hydrophobicity and van der Waals seem to be the main driving forces of the complexation of β -CDs for the molecules under study.

The variables weighted with standard distance, solute gas-hexadecane partition coefficient, effective hydrogen-bond basicity, polar surface, and dipole moment accounted for 19.1%, 3.5%, 1.5%, 0.8%, and 1.3% of the variance, respectively, therefore, although to a lesser extent, interactions due to the polarity (hydrogen bonding) also appear to influence complexation.

TOPS-MODE Structural Interpretation

Recently Katritzky et al.²⁵ presented a QSAR study predicting the free energies of inclusion complexation between diverse *guest* molecules and CDs using (i) CODESSA descriptors and (ii) counts of different molecular fragments. The first of them (the Hansch-type approach⁶⁵) uses as descriptors certain physicochemical parameters calculated either by quantum mechanical methods or by some empirical techniques. The second (the Free-Wilson-type approach⁶⁶) uses counts of different molecular fragments as variables in a multiple regression analysis. Both techniques have their advantages and disadvantages.²⁵ Generally, fragmental descriptors (Free-Wilson-type method) are more interpretable than CODESSA descriptors (Hansch-type method). However, the main disadvantage of QSPR methods based on counts of different molecular fragments is related to the fact that they generally use more variables than CODESSA descriptors, thus leading to smaller values of Fisher criterion (less robust models). Another problem of the fragment-based approach is related to molecules containing fragments of “rare” occurrence (i.e., found in a single molecule), which should be excluded from the training or test sets, thus reducing the number of treated compounds.²⁵ The last problem arises when we attempt to study heterogeneous data sets of organic molecules. In this case there is not necessarily an atomic/bond pattern, which is repeated in all the molecules under study. As a consequence is most adequate to use molecular descriptors like the electronic chemical potential, the molecular electronegativity, the chemical hardness, or other global molecular indices.

This question immediately poses another: can we obtain structural information at a local scale from the models developed using global molecular descriptors? The only information that we need to transform the global model into the atomic/bond contributions is the mathematical relationship

between the global molecular descriptor and the local contributions.⁶⁷

In this article, the TOPS-MODE approach has been used to account for the contributions of molecular parts to the global molecular properties. The main advances of using the TOPS-MODE approach to study complex stability constant between diverse guest molecules and β -CDs as compared with other approaches, such as CODESSA,²⁵ is twofold. On one hand, TOPS-MODE permits the development of robustness and predictive QSPR models in a similar way to those approaches using molecular descriptors, such as CODESSA. On the other hand, it permits the interpretation of the results in terms of fragment contribution identifying those groups, fragments, or molecular regions that can be responsible for the studied property in a similar way as fragmental descriptors does. To do this, fragmental descriptors needs to collect a significant amount of data for each kind of compounds while TOPS-MODE is able to recognise this structural pattern from only one compound present in the data set.⁶⁷ This is possible due to the nature of these descriptors. They describe the molecular structure as a whole in terms of hydrophobic, steric, and electronic characteristics of the molecules that can be transformed into local contributions. In addition to that, new hypothesis can be obtained with TOPS-MODE approach, which can form the basis for new structural interpretation after experimental confirmation.

Thus, TOPS-MODE approach let us to detect fragments that contribute positively or negatively to a particular target endpoint and their effects been interpreted in terms of physicochemical properties.⁶⁸ Specifically in our case, the contributions to the β -CD complex stability constant for each of the selected fragments (see Fig. 1) were extracted from the final orthogonal-descriptor model; these are shown in Table 4. A careful look at these values might allow us to find functional groups, fragments, or molecular regions that either hamper the inclusion phenomenon or enhance it. Further, it might lead us to design molecular structures that have a better profile for the phenomenon or to a rapid selection of the most favourable substance among a long list of substances.

The importance of hydrophobicity in predicting β -CD complexation is also demonstrated here if one looks at the contributions of fragments from F₉ to F₁₅ and F₁₉ to F₂₁, which have a large hydrophobic character. Clearly, their presence in

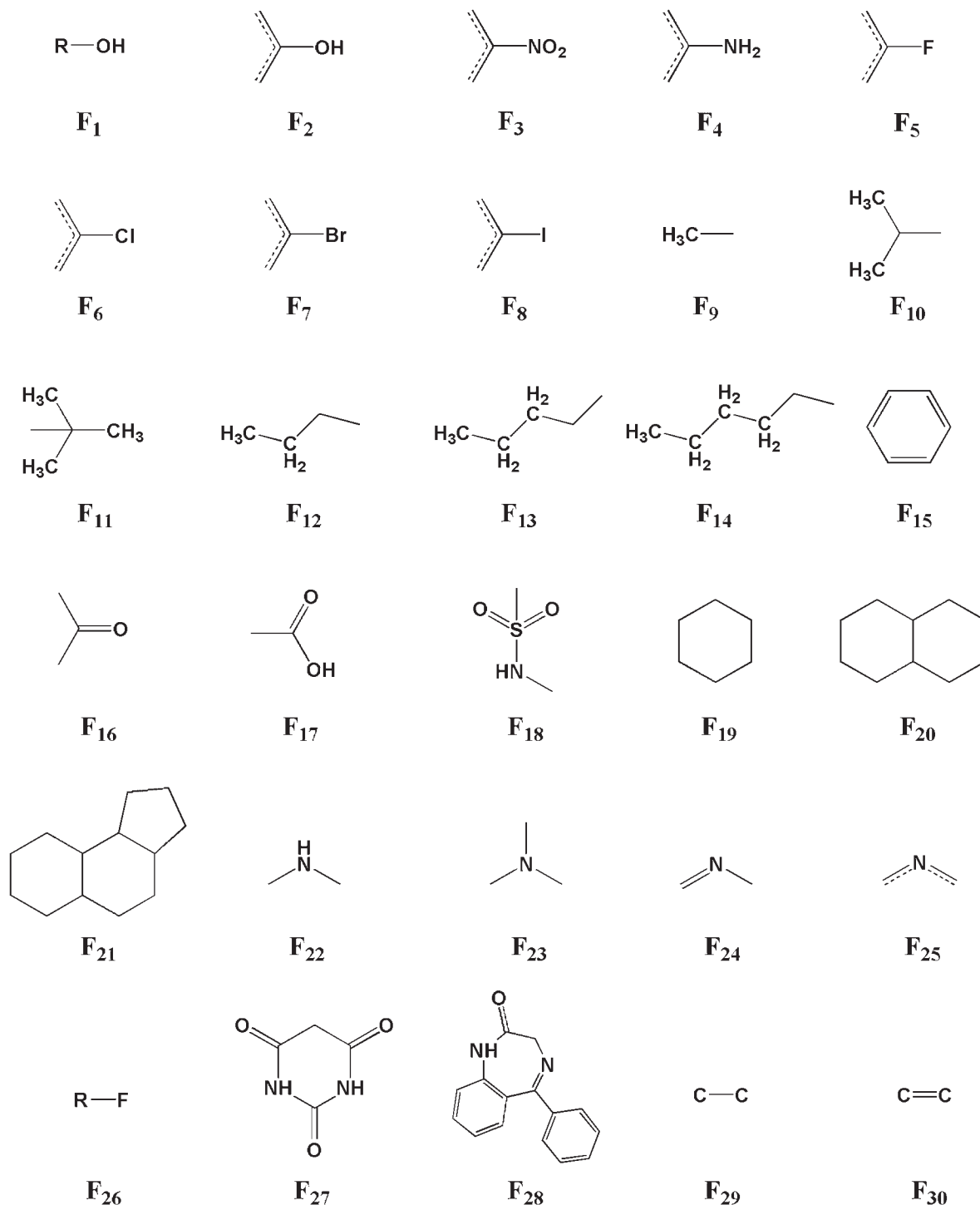


Figure 1. Selected molecular fragments (substructures) for which their contributions to the complexation with β -CD were calculated.

the molecule produces a significant improvement in the stability of the complex (Fig. 2). Indeed, an increase in the length of the hydrocarbon chain raises the stability of the complex because of the

greater hydrophobic character of the molecule (fragments F₉, F₁₂, F₁₃, and F₁₄). Additionally, the amount of branching can affect the complexation (fragments F₉–F₁₁). A certain degree of branching

Table 4. The Contributions of Different Structural Fragments to the Complex Stability Constant

Fragment	Contribution
F ₁	-0.361
F ₂	-0.081
F ₃	-0.062
F ₄	0.048
F ₅	-0.208
F ₆	0.066
F ₇	0.126
F ₈	0.627
F ₉	0.276
F ₁₀	0.611
F ₁₁	0.912
F ₁₂	0.363
F ₁₃	0.521
F ₁₄	0.685
F ₁₅	0.464
F ₁₆	-0.410
F ₁₇	-0.421
F ₁₈	-1.078
F ₁₉	0.598
F ₂₀	0.911
F ₂₁	1.454
F ₂₂	-0.156
F ₂₃	-0.116
F ₂₄	-0.587
F ₂₅	-0.178
F ₂₆	0.042
F ₂₇	-1.162
F ₂₈	-0.558
F ₂₉	0.152
F ₃₀	0.064

may be necessary to achieve optimal van der Waals contacts with the β -CD interior. However, an excess of branching could lead to steric clashes between the compound and the β -CD interior. Furthermore, for fragments F₁₉, F₂₀, and F₂₁, one can see that an increase in the hydrophobic cyclic framework for steroids makes possible the complexation, thus facilitating oral, bucal, or transdermal administration for these highly insoluble molecules.^{69–75}

On the other hand, the increasing flexibility or degrees of freedom in a guest molecule leads to a more favourable complexation entropy, since more of the possible “conformers” can fit properly into the cavity so the presence of unsaturated bonds reduces this flexibility and their chance of inclusion (fragments F₂₉ and F₃₀).⁷⁶

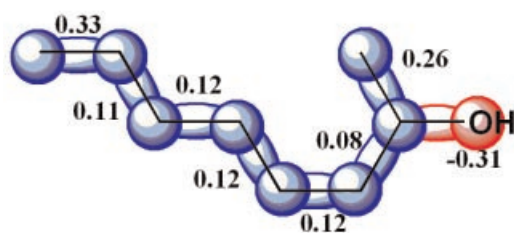
The negative contribution for oxygen and nitrogen containing groups (except aromatic amine) in the β -CD system can be assigned to

the possibility of the competitive interactions with the solvent as discussed by Park and Nah.²⁰ One can also state, by taking into account the negative sign of contributions from fragments F₁ and F₂, that the presence of hydroxyl groups hinders inclusion in the β -CD. As we have previously deduced the phenomenon of inclusion in the β -CD for this set of molecules is dominated mainly by hydrophobic interactions, so the presence of hydrophilic groups diminishes the ability of the molecules to go into the hydrophobic cavity of β -CD (Fig. 2). Notice also the differences between alcoholic and phenolic groups. Maybe these are due to the fact that, even though aliphatic hydroxyl groups can form hydrogen bonds to the peripheral hydroxyls of CD, these interactions are not as strong as those formed by phenolic hydroxyl groups.⁷⁶

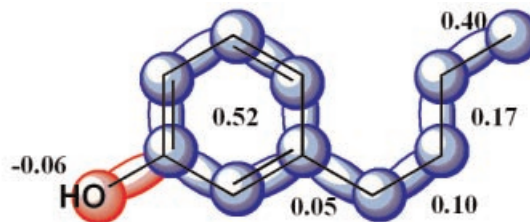
For other hydrophilic groups such as amines (F₄, F₂₂, F₂₃, F₂₄, and F₂₅) something similar could happen. The aromatic amines form a hydrogen bond to the peripheral hydroxyls of the CD stronger than aliphatic amines (Fig. 3). It is worth to note the higher values of the amine groups regarding the hydroxyl groups. Possibly, this suggests better hydrogen bonding formed between guest’s amine groups and the host’s hydroxyl groups than that formed between guest’s and host’s hydroxyl groups or guest’s amine groups may come to form hydrogen bonds with various host’s hydroxyl groups.

On another fragments like halogenated derivatives of benzilic group (F₅–F₈), complexity is enhanced by increased volumes of substituent, confirming what has been observed by Liu and Guo⁷⁷ where the increased volume and polarisability of the guest substituent can increase the stability of the complex due to the stronger van der Waals interactions. It is important to note that the F₅ fragment (Ar-F) makes a negative contribution to complexation with the β -CD while the rest of halogenated aromatic fragments make a positive one. In the cases of the Ar-F fragment should be considered another additional intermolecular force of attraction: hydrogen bonding. The hydrogen bonding formed between Ar-F and water (solvent) could be more powerful than that formed between Ar-F and β -CD and then it could justify the negative contribution of F₅. Therefore, the presence of Ar-F fragment in a molecule decreases the stability of the complex.

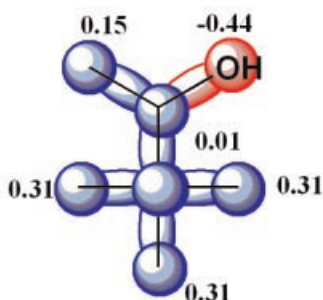
By analysing whole of these contributions, one might explore other situations in which drugs with low activity can be enhanced if they have a



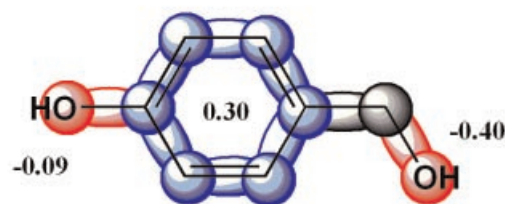
Compound 113



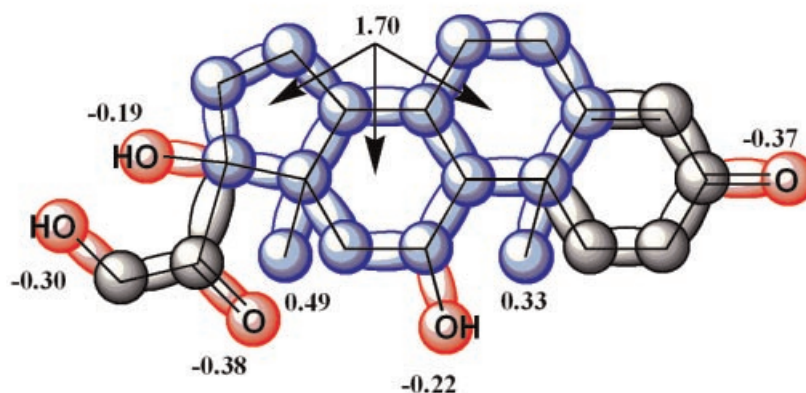
Compound 137



Compound 60



Compound 77



Compound 202

Figure 2. Contributions of the hydrophobic and hydroxyl groups. The red-coloured spheres represent the negative contributions to the complex stability constants whereas the blue spheres are those with positive contributions.

good complexation with the β -CD or facilitate their administration (as we have seen with steroids) or its bioavailability. For example, looking at the values of IC_{50} taste reported for benzodiazepines on the work of Sutherland et al.,⁷⁸ for which in compounds **191** and **232** (Fig. 4) the replacement of a nitro group by a chlorine entails a reduction of its power, but an

increase in their complexation with β -CD (fragments F_3 and F_6).

Suzuki¹⁶ found similar results when they used a group contribution model (GCM) for predicting free energies of complexation between guest molecules with β -CDs based on the same robust training set of 218 diverse ligands. In general, the presence of carbon, halogen, and sulphur results

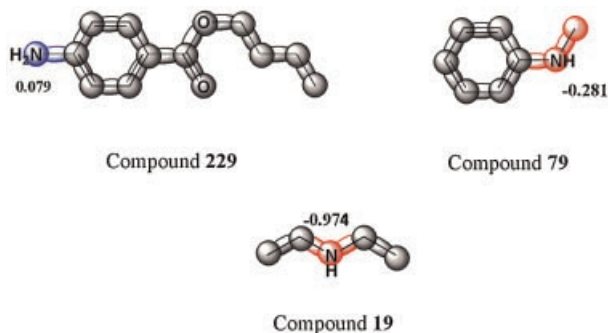


Figure 3. Contributions of the amine groups.

in an increase of the complex stability. In contrast, the presence of most oxygen and nitrogen containing groups (except $-\text{OH}$ (phenol), $-\text{O}$ -(ring), $>\text{C}=\text{O}$ (ring), $-\text{COOH}$, $-\text{NH}_2$, and $-\text{NO}_2$) decreases this one. In this technique, a molecule is analysed for the presence of certain predefined fragments or functional groups. Each group has a specific contribution to the overall value of the binding free energy, which is obtained by summing those individual contributions. After that, specific group contributions are used as descriptors for generating the QSAR model. This technique limits the types of compounds that can be evaluated. A molecule that contains very little or none of the fragments in the model training set cannot be properly analysed. However, this concern is not an issue for our model. TOPS-MODE descriptors describe the molecular structure in a global way, permitting to find the contribution of any fragment in the molecular structure to the complexation with the β -CD. Thus, by using QSAR model and TOPS-MODE descriptors you easily obtain the contribution of any fragment in training/test set to the property, which is an

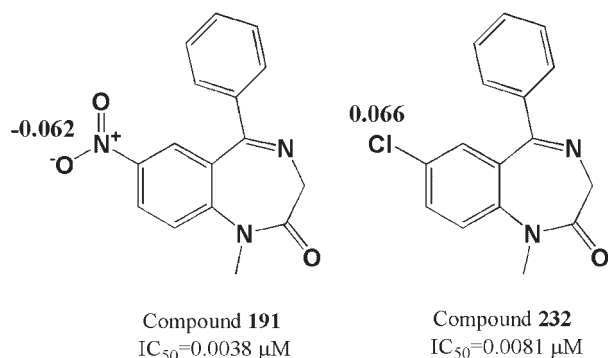


Figure 4. Fragment contributions of two benzodiazepines.

advantage of this work. Actually, in Figure 1 and Table 4 we show some simple molecular fragments ($\text{F}_1\text{--}\text{F}_9$, F_{16} , F_{17} , $\text{F}_{22}\text{--}\text{F}_{25}$) similar to Suzuki et al., and their fragment contributions, respectively. But we can also obtain new and more complex hypothesis (e.g., F_{21} , F_{18} , F_{27} , F_{28}) with TOPS-MODE approach, which can form the basis for new structural interpretation after experimental confirmation.

Applicability Domain

It would be very interesting to have a predictive model for the vast majority of chemicals, particularly for those which have not been tested and, therefore, with unknown $\log K$ values. Since this is usually not possible, one should define the applicability domain of the QSAR model, that is, the range within which the model bears a new compound.

For that purpose, we built a Williams plot using the leverage values calculated for each compound. As seen in Figure 5, most of the compounds of the test set are within the applicability domain covered by ± 3 times the standard residual (σ) and the leverage threshold h^* ($=0.129$), save for compounds 31, 53, 175, 197, 202, 204, 205, 207, 211, 214, 218, and 233. Even so, the latter should not be considered outliers but influential chemicals.⁶¹

Nevertheless, all evaluations pertaining to the external set were performed by taking into account the applicability domain of our QSAR model. So, if a chemical belonging to the test set

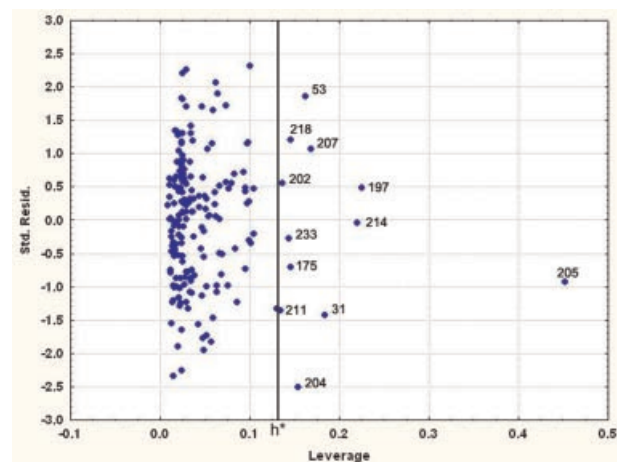


Figure 5. Williams plot based on Eq. (6), that is, plot of standardised residuals versus leverage values with a warning leverage of $h^* = 0.129$.

had a leverage value greater than h^* , we consider that this means that the prediction is the result of substantial extrapolation and therefore may not be reliable.⁶²

CONCLUSION

Due to the beneficial effects arising from the complexation of drugs with β -CDs, we have applied here a QSPR regression-based approach to a diverse set of 233 organic compounds with known complex stability constant (K) values. By means of k -MCA, 80% of these compounds were selected as the training set and the remaining as the external evaluation set. With regard to the QSPR modelling, the combination of multivariate data analysis in conjunction with a TOPS-MODE representation and the genetic selection algorithm was found to produce a final regression model with good accuracy, internal cross-validation statistics, and predictivity on the external data.

The analysis of the most frequent descriptors implicated in the final QSPR model afforded model interpretation in terms of chemical features influencing complexation with β -CD. The major driving forces for complexation, extracted from the model, were hydrophobicity and van der Waals interactions, and thus the presence of hydrophobic groups (hydrocarbon chains, aryl groups, etc.) and voluminous species (Cl, Br, I, etc.) in the molecule facilitate their complexation by β -CD, while possibly increasing the beneficial effects (solubility and bioavailability) derived from this. The final QSPR model was further used to collect effective information about what kinds of groups favour such complexation.

In summary, the information gathered by these descriptors given in the form of bond contributions provide valuable information for future use in drug design and other applications related to complexation with β -CDs.

ACKNOWLEDGMENTS

The authors acknowledge to MODESLAB 1.0 software owners for delivering a free copy of such program and the anonymous reviewers for their comments. A.M.H. acknowledges the *Portuguese Fundação para a Ciência e a Tecnologia* (FCT, Lisboa) (SFRH/BD/22692/2005) for financial support.

REFERENCES

1. Saenger W, Jacob J, Gessler K, Steiner T, Daniel S, Sanbe H, Koizumi K, Smith SM, Tanaka T. 1998. Structures of the common cyclodextrins and their larger analogues—Beyond the doughnut. *Chem Rev* 98:1787–1802.
2. Loftsson T, Brewster M, Masson M. 2004. Role of cyclodextrins in improving oral drug delivery. *Am J Drug Deliv* 2:261–275.
3. Davis ME, Brewster M. 2004. Cyclodextrin-based pharmaceuticals: Past, present and future. *Nat Rev Drug Discov* 3:1023–1035.
4. Avdeef A, Bendels S, Tsinman O, Tsinman K, Kansy M. 2007. Solubility excipient classification gradient maps. *Pharm Res* 24:530–545.
5. Kim C, Park J. 2004. Solubility enhancers for oral drug delivery. *Am J Drug Deliv* 2:113–130.
6. Loftsson T, Jarho P, Masson M, Jarvinen T. 2005. Cyclodextrins in drug delivery. *Expert Opin Drug Deliv* 2:335–351.
7. Kola I, Landis J. 2004. Can the pharmaceutical industry reduce attrition rates? *Nat Rev Drug Discov* 2:711–715.
8. Liu R. 2000. Water-insoluble drug formulation. Englewood: CO Interpharm Press.
9. Irie T, Uekama K. 1997. Pharmaceutical applications of cyclodextrins. iii. Toxicological issues and safety evaluation. *J Pharm Sci* 86:147–162.
10. Szejtli J. 1998. Introduction and general overview of cyclodextrin chemistry. *Chem Rev* 98:1743–1753.
11. Lantz A, Rodriguez M, Wetterer S, Armstrong D. 2006. Estimation of association constants between oral malodor components and various native and derivatized cyclodextrins. *Anal Chim Acta* 557:184–190.
12. Uekama K. 1999. Cyclodextrins in drug delivery. *Adv Drug Deliv Rev* 36:1–2.
13. Duchêne D, editor. 1987. Cyclodextrins and their industrial uses. Paris: Editions de Santé Paris.
14. Horvath G, Premkumar T, Boztas A, Lee E, Jon S, Geckeler KE. 2008. Supramolecular nanoencapsulation as a tool: Solubilization of the anti-cancer drug trans-dichloro(dipyridine)platinum(ii) by complexation with beta-cyclodextrin. *Mol Pharm* 5:358–363.
15. Lipkowitz KB. 1998. Applications of computational chemistry to the study of cyclodextrins. *Chem Rev* 98:1829–1873.
16. Suzuki T. 2001. A nonlinear group contribution method for predicting the free energies of inclusion complexation of organic molecules with α - and β -cyclodextrins. *J Chem Inf Comput Sci* 41:1266–1273.
17. Pérez F, Jaime C, Sánchez-Ruiz X. 1995. Mm2 calculations on cyclodextrins: Multimodel inclusion complexes. *J Org Chem* 60:3840–3845.

18. Matsui Y, Nishioka T, Fujita T. 1985. Quantitative structure-reactivity analysis of the inclusion mechanism by cyclodextrins. *Top Curr Chem* 128: 61–89.
19. Davis DM, Savage JR. 1993. Correlation analysis of the host-guest interaction of α -cyclodextrin and substituted benzenes. *J Chem Res-S* 94–95.
20. Park JH, Nah TH. 1994. Binding forces contributing to the complexation of organic molecules with β -cyclodextrin in aqueous solution. *J Chem Soc [Perkin Trans] 2*:1359–1362.
21. Klein CT, Polheim D, Viernstein H, Wolschann P. 2000. A method for predicting the free energies of complexation between β -cyclodextrin and guest molecules. *J Inclusion Phenom Macrocyclic Chem* 36:409–423.
22. Liu L, Guo QX. 1999. Wavelet neural network and its application to the inclusion of β -cyclodextrin with benzene derivatives. *J Chem Inf Comput Sci* 39:133–138.
23. Suzuki T, Ishida M, Fabian WMF. 2000. Classical QSAR and comparative molecular field analyses of the host-guest interaction of organic molecules with cyclodextrins. *J Comput Aided Mol Des* 14:669–678.
24. Cramer IRD, Patterson DE, Bunce JD. 1988. Comparative molecular field analysis (COMFA). 1. Effect of shape on binding of steroids to carrier proteins. *J Am Chem Soc* 110:5959–5967.
25. Katritzky AR, Fara DC, Yang HF, Karelson M, Suzuki T, Solov'ev VP, Varnek A. 2004. Quantitative structure-property relationship modelling of β -cyclodextrin complexation free energies. *J Chem Inf Comput Sci* 44:529–541.
26. Estrada E. 1996. Spectral moments of the edge adjacency matrix in molecular graphs. 1. Definition and applications to the prediction of physical properties of alkanes. *J Chem Inf Comput Sci* 36:844–849.
27. Estrada E. 1997. Spectral moments of the edge-adjacency matrix of molecular graphs. 2. Molecules containing heteroatoms and QSAR applications. *J Chem Inf Comput Sci* 37:320–328.
28. Estrada E. 1995. Edge adjacency relationships and a novel topological index related to molecular volume. *J Chem Inf Comput Sci* 35:31–33.
29. Estrada E, Molina E. 2006. Automatic extraction of structural alerts for predicting chromosome aberrations of organic compounds. *J Mol Graph Model* 25:275–288.
30. Estrada E, Patlewicz G, Gutierrez Y. 2004. From knowledge generation to knowledge archive. A general strategy using tops-mode with Derek to formulate new alerts for skin sensitization. *J Chem Inf Comput Sci* 44:688–698.
31. González MP, Dias L, Helguera AM. 2004. A topological sub-structural approach to the mutagenic activity in dental monomers. 2. Cycloaliphatic epoxides. *Polymer* 15:5353–5359.
32. González MP, Helguera AM, Molina R, Garca JR. 2004. A topological substructural approach of the mutagenic activity in dental monomers. 1. Aromatic epoxides. *Polymer* 45:2773–2779.
33. González MP, Helguera AM, Cabrera MA. 2005. Quantitative structureactivity relationship to predict toxicological properties of benzene derivative compounds. *Bioorg Med Chem* 13:1775–1781.
34. Helguera AM, Pérez MAC, Combes RD, González MP. 2006. Quantitative structure activity relationship for the computational prediction of nitrocompounds carcinogenicity. *Toxicology* 220:51–62.
35. Helguera AM, González MP, Cordeiro MNDS, Cabrera MA. 2007. Quantitative structure carcinogenicity relationship for detecting structural alerts in nitrosocompounds. *Toxicol Appl Pharmacol* 221: 189–202.
36. Helguera AM, González MP, Cordeiro MNDS, Cabrera MA. 2008. Quantitative structure-carcinogenicity relationship for detecting structural alerts in nitroso compounds: Species, rat; sex, female; route of administration, gavage. *Chem Res Toxicol* 21: 633–642.
37. González MP, Terán C, Teixeira M. 2006. A topological function based on spectral moments for predicting affinity toward $\alpha 3$ adenosine receptors. *Bioorg Med Chem Lett* 16:1291–1296.
38. González MP, Helguera AM, Collado IG. 2006. A topological substructural molecular design to predict soil sorption coefficients for pesticides. *Mol Divers* 10:109–118.
39. González MP, Díaz HG, Ruiz RM, Cabrera MA, Ramos de Armas R. 2003. Tops-mode based QSARs derived from heterogeneous series of compounds. Applications to the design of new herbicides. *J Chem Inf Comput Sci* 43:1192–1199.
40. Pérez-Garrido A, González MP, Escudero AG. 2008. Halogenated derivatives QSAR model using spectral moments to predict haloacetic acids (haa) mutagenicity. *Bioorg Med Chem* 16:5720–5732.
41. Helguera AM, Cordeiro MNDS, Cabrera MA, Combes RD, González MP. 2008. Quantitative structure carcinogenicity relationship for detecting structural alerts in nitroso-compounds species: Rat; sex: male; route of administration: water. *Toxicol Appl Pharmacol* 231:197–207.
42. Helguera AM, Pérez MAC, González MP, Ruiz RM, Díaz HG. 2005. A topological substructural approach applied to the computational prediction of rodent carcinogenicity. *Bioorg Med Chem* 13: 2477–2488.
43. Environment Directorate OECD. 2007. Guidance Document of the Validation of (Quantitative) Structure-Activity Relationships (Q)SAR Models. Environmental Health and Safety Publications, Series on Testing and Assessment No. 69.

44. Estrada E, Uriarte E, Gutierrez Y, González H. 2003. Quantitative structure-toxicity relationships using tops-mode. 3. Structural factors influencing the permeability of commercial solvents through living human skin. *SAR QSAR Environ Res* 14: 145–163.
45. Gutierrez Y, Estrada E. 2002. Modes Lab, Version 1.0.
46. Weininger D. 1988. Smiles, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J Chem Inf Comput Sci* 28:31–36.
47. Goldberg D. 1989. Genetic algorithms in search, optimization, and machine learning. USA: Addison-Wesley.
48. Todeschini R, Ballabio D, Consonni V, Mauri A, Pavan M. 2004. Mobydigs Computer Software. Milano: TALLETE SRL.
49. Pérez-Garrido A, Helguera AM, Cordeiro MNDS, Abellán A, Escudero AG. 2008. Convenient QSAR model for predicting the complexation of structurally diverse compounds with β -cyclodextrins. *Bioorg Med Chem* 17:896–904.
50. Garcia-Domenech R, Julian-Ortiz JV. 1998. Antimicrobial activity characterization in a heterogeneous group of compounds. *J Chem Inf Comput Sci* 38:445–449.
51. Kubinyi H. 1994. Variable selection in QSAR studies. 1. An evolutionary algorithm. *Quant Struct Act Relat* 13:285–294.
52. Kubinyi H. 1994. Variable selection in QSAR studies. 2. A highly efficient combination of systematic search and evolution. *Quant Struct Act Relat* 13: 393–401.
53. Akaike H. 1973. Information theory and an extension of the maximum likelihood principle. In *Proceedings of the Second International Symposium on Information Theory*. Budapest: Akademiai Kiado, pp 267–281.
54. Akaike H. 1974. New look at statistical-model identification. *IEEE Trans Automat Control* AC-19: 716–723.
55. Lučić B, Nikolić S, Trinajstić N, Jurić D. 1995. The structure-property models can be improved using the orthogonalized descriptors. *J Chem Inf Comput Sci* 35:532–538.
56. Todeschini R, Consonni V. 2000. *Handbook of molecular descriptors*. Mannheim: Wiley-VCH, 667p.
57. Klein D, Randić M, Babić D, Lučić B, Nikolić S, Trinajstić N. 1997. Hierarchical orthogonalization of descriptors. *Int J Quantum Chem* 63:215–222.
58. Randić M. 1991. Orthogonal molecular descriptors. *N J Chem* 15:517–525.
59. Randić M. 1991. Resolution of ambiguities in structure-property studies by use of orthogonal descriptors. *J Chem Inf Comput Sci* 31:311–320.
60. Randić M. 1991. Correlation of enthalpy of octanes with orthogonal connectivity indices. *J Mol Struct (Theochem)* 233:45–59.
61. Eriksson L, Jaworska J, Worth AP, Cronin MTD, McDowell RM, Gramatica P. 2003. Methods for reliability and uncertainty assessment and for applicability evaluations of classification- and regression-based QSARs. *Environ Health Perspect* 111:1361–1375.
62. Netzeva TI, Worth AP, Aldenberg T, Benigni R, Cronin MTD, Gramatica P, Jaworska JS, Kahn S, Klopman P, Marchant CA, Myatt G, Nikolova-Jeliazkova N, Patlewicz GY, Perkins R, Roberts DW, Schultz TW, Stanton DT, van de Sandt JJH, Tong W, Veith G, Yang C. 2005. Current status of methods for defining the applicability domain of (quantitative) structure activity relationships. *ATLA* 33:155–173.
63. Gramatica P. 2007. Principles of QSAR models validation: Internal and external. *QSAR Comb Sci* 26:1–9.
64. Vighi M, Gramatica P, Consolaro F, Todeschini R. 2001. QSAR and chemometrics approaches for setting water quality objectives for dangerous chemicals. *Ecotoxicol Environ Saf* 49:206–220.
65. Hansch C, Leo A, Hoekman DH. 1995. *Exploring QSAR fundamentals and applications in chemistry and biology*. In ACS professional reference book. Washington, DC: American Chemical Society, p 580.
66. Free SM, Wilson JW. 1964. A mathematical contribution to structure-activity studies. *J Med Chem* 7:395–399.
67. Estrada E. 2008. How the parts organize in the whole? A top-down view of molecular descriptors and properties for QSAR and drug design. *Mini Rev Med Chem* 8:213–221.
68. Benigni R, Giuliani A. 2003. Putting the predictive toxicology challenge into perspective: Reflections on the results. *Bioinformatics* 19:1194–1200.
69. Seo H, Tsuruoka M, Hashimoto T, Fujinaga T, Otagiri M, Uekama K. 1983. Enhancement of oral bioavailability of spironolactone by betacyclodextrin and gamma-cyclodextrin complexations. *Chem Pharm Bull* 31:286–291.
70. Pitha J, Harman SM, Michel ME. 1986. Hydrophilic cyclodextrin derivatives enable effective oral-administration of steroidal hormones. *J Pharm Sci* 75:165–167.
71. Uekama K, Fujinaga T, Otagiri M, Seo H, Tsuruoka M. 1981. Enhanced bioavailability of digoxin by gamma-cyclodextrin complexation. *J Pharmacobiodyn* 4:735–737.
72. Uekama K, Fujinaga T, Hirayama F, Otagiri M, Yamasaki M, Seo H, Hashimoto T, Tsuruoka M. 1983. Improvement of the oral bioavailability of digitalis glycosides by cyclodextrin complexation. *J Pharm Sci* 72:1338–1341.

73. Taylor GT, Weiss J, Pitha J. 1989. Testosterone in a cyclodextrin containing formulation—Behavioral and physiological-effects of episode like pulses in rats. *Pharm Res* 6:641–646.
74. Loftsson T, Olafsdottir BJ, Bodor N. 1991. The effects of cyclodextrins on transdermal delivery of drugs. *Eur J Pharm Biopharm* 37:30–33.
75. Uekama K, Arimori K, Sakai A, Masaki K, Irie T, Otagiri M. 1987. Improvement in percutaneous-absorption of prednisolone by betacyclodextrin and gamma-cyclodextrin complexations. *Chem Pharm Bull* 35:2910–2913.
76. Rekharsky MV, Inoue Y. 1998. Complexation thermodynamics of cyclodextrins. *Chem Rev* 98:1875–1917.
77. Liu L, Guo QX. 2002. The driving forces in the inclusion complexation of cyclodextrins. *J Inclusion Phenom Macrocyclic Chem* 42:1–14.
78. Sutherland JJ, O'Brien LA, Boztas A, Weaver DF. 2003. Spline-fitting with a genetic algorithm: A method for developing classification structure-activity relationships. *J Chem Inf Comput Sci* 43:1906–1915.
79. Inoue Y, Hakushi T, Liu Y, Tong LH, Shen BJ, Jin DS. 1993. Thermodynamics of molecular recognition by cyclodextrins. 1. Calorimetric titration of inclusion complexation of naphthalenesulfonates with α -, β -, and γ -cyclodextrins: Enthalpy-entropy compensation. *J Am Chem Soc* 115:475–481.
80. Carpignano R, Marzona M, Cattaneo E, Quaranta S. 1997. QSAR study of inclusion complexes of heterocyclic compounds with β -cyclodextrin. *Anal Chim Acta* 348:489–493.
81. Rekharsky MV, Goldberg RN, Schwarz FP, Tewari YB, Ross PD, Yamashoji Y, Inoue Y. 1995. Thermodynamic and nuclear magnetic resonance study of the interactions of α - and β -cyclodextrin with model substances: Phenethylamine, ephedrine, and related substances. *J Am Chem Soc* 117:8830–8840.
82. Wallimann P, Marti T, Fürer A, Diederich F. 1997. Steroids in molecular recognition. *Chem Rev* 97:1567–1608.

available at www.sciencedirect.comjournal homepage: www.intl.elsevierhealth.com/journals/dema

QSAR models to predict mutagenicity of acrylates, methacrylates and α,β -unsaturated carbonyl compounds

Alfonso Pérez-Garrido^{a,b,*}, Aliuska Morales Helguera^{c,d,e}, Francisco Girón Rodríguez^b, M.Natália D.S. Cordeiro^e

^a Environmental Engineering and Toxicology Dpt., Catholic University of San Antonio, Guadalupe, Murcia, Spain

^b Department of Food and Nutrition Technology, Catholic University of San Antonio, Guadalupe, Murcia, Spain

^c Department of Chemistry, Faculty of Chemistry and Pharmacy, Central University of Las Villas, Santa Clara, Villa Clara, Cuba

^d Molecular Simulation and Drug Design Group, Chemical Bioactive Center, Central University of Las Villas, Santa Clara, Villa Clara, Cuba

^e REQUIMTE, Chemistry Department, Faculty of Sciences, University of Porto, Porto, Portugal

ARTICLE INFO

Article history:

Received 30 May 2009

Received in revised form

8 September 2009

Accepted 26 November 2009

Keywords:

QSAR

Mutagenicity

α,β -Unsaturated carbonyl compounds

Dental restorative monomers

ABSTRACT

Objective. The purpose of this study is to develop a quantitative structure–activity relationship (QSAR) model that can distinguish mutagenic from non-mutagenic species with α,β -unsaturated carbonyl moiety using two endpoints for this activity – Ames test and mammalian cell gene mutation test – and also to gather information about the molecular features that most contribute to eliminate the mutagenic effects of these chemicals.

Methods. Two data sets were used for modeling the two mutagenicity endpoints: (1) Ames test and (2) mammalian cells mutagenesis. The first one comprised 220 molecules, while the second one 48 substances, ranging from acrylates, methacrylates to α,β -unsaturated carbonyl compounds. The QSAR models were developed by applying linear discriminant analysis (LDA) along with different sets of descriptors computed using the DRAGON software.

Results. For both endpoints, there was a concordance of 89% in the prediction and 97% confidentiality by combining the three models for the Ames test mutagenicity. We have also identified several structural alerts to assist the design of new monomers.

Significance. These individual models and especially their combination are attractive from the point of view of molecular modeling and could be used for the prediction and design of new monomers that do not pose a human health risk.

© 2010 Academy of Dental Materials. Published by Elsevier Ltd. All rights reserved.

1. Introduction

A matrix resin made of acrylate or methacrylate based monomers that are photo and/or chemically polymerizable is used usually for dental resin filling materials and adhesives. These dental restorative materials are prepared in situ and, as the polymerization is often not ideal, some unreacted

monomers will dribble from the restoration over time [1,2]. Any unpolymerized monomer in the composite has a potential biological liability if it leaches from the composite toward the pulp of the tooth [3]. Among other substances, triethylene glycol dimethacrylate (TEGDMA)(Fig. 1) causes DNA deletions in mammalian cells [4,5].

Actually, acrylates and methacrylates are recognized as a chemical category by the U.S. EPA in evaluation of

* Corresponding author at: Environmental Engineering and Toxicology Dpt., Catholic University of San Antonio, Guadalupe, Murcia, Spain. Tel.: +34 968 278 755.

E-mail address: Aperez@pdi.ucam.edu (A. Pérez-Garrido).

0109-5641/\$ – see front matter © 2010 Academy of Dental Materials. Published by Elsevier Ltd. All rights reserved.

doi:10.1016/j.dental.2009.11.158

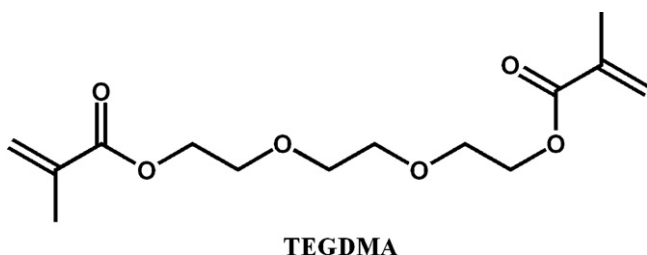


Fig. 1 – Molecular structure of triethylene glycol dimethacrylate (TEGDMA).

pre-manufacturing notices, specifically when considering environmental toxicity [6]. Similarly, the OECD (Organization for Economic Cooperation and Development) has assessed short-chain alkyl methacrylate esters as a chemical category within the HPV (High Productive Volume) program. The European Union has established a category of monoalkyl- and monoaryl-acrylates, which uses to regulate chemical classification and labeling within Annex 1 of the Dangerous Substances Directive [7]. In general, compounds with a α,β -unsaturated carbonyl moiety are particularly reactive, due to the positively polarized double bond, and interact with nucleophilic groups on peptides, proteins, or DNA, thus causing disruption of these biological macromolecules and hence cause toxic effects [8–13], particularly mutagenicity in which the major mechanistic domains are Michael type acceptor or Schiff base formation [14].

The commonly experimental assays employed for determining the mutagenicity of chemicals are however not easy or inexpensive to carry out, and usually too slow to cope with number of them that may have mutagenic effects. One way to alleviate such problems is to use alternative approaches, such as quantitative structure–activity relationships (QSAR) modeling, proven to be particularly useful in agrochemical, pharmaceutical chemistry and toxicology [15]. The formulation of thousands of equations using QSAR methodology attests to the validation of its concepts and its usefulness in the prediction of activity, as well as in the theoretical elucidation of the mechanisms of action at molecular levels. The use of QSAR models is also supported by the conceptual framework of the fifth level of the OECD [16], which foresees the use of *in vitro* tests and QSAR models before *in vivo* tests. Thus, this methodology is a recognized tool for prioritizing chemicals for subsequent experimental verification.

There have been many attempts to predict the mutagenic potency of these substances through QSAR models [17–20]. In many of them, the authors model the gradation of the potency of mutagenic compounds. Thus the models found are not useful to predict the mutagenic activity, i.e. whether the chemicals are mutagenic or not, but only a secondary information: potency. Some studies have shown that structural effects on mutagenic potency should be distinguished from effects on yes/no activity [21].

Attempts to model the mutagenic activity for these substances to be able to distinguish mutagenicity were carried out by Benigni et al. [22], who developed a prediction model for twenty-five α,β -unsaturated aldehydes by stepwise linear discriminant analysis based on data from *Salmonella*

typhimurium TA100 strain assays. The results of this study indicated a dependency between mutagenicity, hydrophobicity and molecular volume.

In addition to that, a study has appeared to model and predict the Ames TA100-derived mutagenicity for 45 α,β -unsaturated carbonyl compounds [14]. In this work, the set of compounds was divided into several sub-groups and a system of rules was developed for the classification based on the different mechanisms of action known. Like the study of Benigni et al. [22,23], the rules obtained establish a relationship between mutagenicity, hydrophobicity and molecular volume.

All these studies only considered the Ames test as exclusive endpoint for estimating the mutagenicity, though it is well-known that it is necessary to use information across multiple endpoints to get more reliable predictions [24]. In addition to that, the predictions for each substance can be different depending on the type of descriptors or methodology used to obtain the model. Recently Zhu et al. [25] and Brien et al. [26] demonstrated that both predictive performance and coverage of the final consensus QSAR models were superior as compared to these parameters for individual models.

The aim of this study is to build a reliable predictive QSAR model from α,β -unsaturated carbonyl data that could be used to probe mutagenic activity overall: Ames mutagenesis test (AMES) [27] and mammalian cells genetic mutation test (MCGM) [28]. We examined the use of linear discriminant analysis and feature selection algorithms in conjunction with a variety of molecular descriptors calculated with the DRAGON package [29]. DRAGON provides more than 1600 molecular descriptors divided into several families: OD (constitutional descriptors), 1D (e.g., functional group counts), 2D (e.g., topological descriptors and connectivity indices), and 3D (e.g., GETAWAY, WHIM, RDF and 3D-MORSE descriptors). These descriptors have proved to be reliable in QSAR modeling of properties as diverse as fragrance [30] to soil sorption [31] to complexation [32]. DRAGON descriptors have also been successfully applied on QSAR modeling of various target bioactivities like receptors' affinity [33–40] to enzymatic inhibition [41–49] to carcinogenicity [50–53], as well as in drug design [54–57]. This work also shows how mutagenesis models can be combined to shape predictions depending on one's needs. That is to say, we demonstrate how the judicious use of data and considered combination of predictions can produce models that provide truly useful answers.

2. Materials and methods

2.1. Data

The first set of data used comprises 220 compounds with α,β -unsaturated carbonyl moiety (Table 1) derived from the Ames test classification for mutagenicity made by Bursi and coworkers [58]. Notice that their analysis has been restricted to the standard plate or preincubation tests of *Salmonella typhimurium* strains TA98, TA100, TA1535, TA1537, TA97 either with or without a metabolic activation mixture. In addition, strains TA102 and TA1538 have been applied in cases where the results of other strains were equivocal. In such classification, a compound was categorized as a mutagen if at least one

Table 1 – CAS number, observed and predicted classification for the compounds in the two endpoint used in this study.

Compound no.	CAS	Name	Observed AMES class.	Partition	Predicted AMES class.	Observed MCGM class.	Partition	Predicted MCGM class.
1	87406-72-2	N,n-Diglycidylacrylamide	1	Training	-1	-	-	-
2	23282-20-4	Fusarenone x	-1	Training	-1	-	-	-
3	34807-41-5	Mezerein	-1	Training	-1	-	-	-
4	6379-69-7	Trichothecin	-1	Training	-1	-	-	-
5	63166-73-4	Phyllanthoside	-1	Training	-1	-	-	-
6	23246-96-0	Riddelline	1	Training	-1	-	-	-
7	130-01-8	Senecionine	-1	Training	-1	-	-	-
8	21794-01-4	Rubratoxin b	-1	Training	-1	-	-	-
9	61203-01-8	Methyl-1-bromovinyl ketone	1	Training	1	-	-	-
10	2849-98-1	Pentyl methacrylate	-1	Training	-1	-	-	-
11	97-90-5	Ethylene glycol dimethacrylate	-1	Training	-1	1	Training	1
12	4513-36-4	Neopentyl acrylate	-1	Training	-1	-	-	-
13	13675-34-8	1,5-Pentanediol dimethacrylate	-1	Training	-1	-	-	-
14	868-77-9	2-Hydroxyethyl methacrylate (HEMA)	-1	Training	-1	-	-	-
15	5466-77-3	2-Ethylhexyl p-methoxycinnamate	-1	Training	-1	-	-	-
16	123-73-9	(E)-Crotonaldehyde	1	Training	1	-	-	-
17	96910-73-5	9-Hydroxyvelleral	1	Training	1	-	-	-
18	14925-39-4	2-Bromoacrolein	1	Training	1	-	-	-
19	2397-76-4	Neopentyl methacrylate	-1	Training	-1	-	-	-
20	68162-37-8	α ,5-Dinitro-2-furanacrylic acid, methyl ester	1	Training	1	-	-	-
21	836-37-3	α -Cyano-5-nitro-2-furanacrylic acid, methyl ester	1	Training	1	-	-	-
22	29590-42-9	Isooctyl acrylate monomer	-1	Training	-1	-1	Training	-1
23	555-68-0	3-Nitrocinnamic acid	1	Training	1	-	-	-
24	58-54-8	Ethacrynic acid	-1	Training	-1	-	-	-
25	3688-53-7	Furylfuramide	1	Training	1	1	Training	1
26	18829-55-5	Trans-2-heptenal	1	Training	1	-	-	-
27	1013-96-3	O-Nitrocinnamic acid	-1	Training	-1	-	-	-
28	102059-18-7	4-Acetamidochalcone	1	Training	-1	-	-	-
29	7364-09-2	2-Chloro-3,3-dimethylacrolein	1	Training	1	-	-	-
30	499-12-7	Aconitic acid	-1	Training	-1	-	-	-
31	10443-65-9	2-Bromoacrylic acid	-1	Training	1	-	-	-
32	6281-23-8	5-Nitro-2-furanacrylic acid	1	Training	1	-	-	-
33	109460-96-0	Methyl 2-cyano-3-(2-bromophenyl)acrylate	-1	Training	-1	-	-	-
34	5443-49-2	α -Bromocinnamaldehyde	1	Training	1	-	-	-
35	645-62-5	2-Ethyl-2-hexenal	-1	Training	1	-	-	-
36	97-86-9	Isobutyl methacrylate	-1	Training	-1	-	-	-
37	137-05-3	Methyl 2-cyanoacrylate	1	Training	-1	-	-	-
38	20426-12-4	4-Hydroxychalcone	-1	Training	-1	-	-	-
39	104-55-2	Cinnamaldehyde	-1	Training	-1	-	-	-

Table 1 – (Continued)

Compound no.	CAS	Name	Observed AMES class.	Partition	Predicted AMES class.	Observed MCGM class.	Partition	Predicted MCGM class.
40	488-11-9	Mucobromic acid	1	Training	1	-	-	-
41	109-16-0	Triethylene glycol dimethacrylate (TEGDMA)	-1	Training	-1	1	Training	1
42	434-07-1	Oxymetholone	-1	Training	-1	-1	Training	-1
43	126572-80-3	(E)-2-Chloro-3-(dichloromethyl)-2-butenedioic acid	1	Training	1	-	-	-
44	13171-21-6	Phosphamidon	1	Training	1	-	-	-
45	141-32-2	N-Butyl acrylate	-1	Training	-1	1	Training	1
46	94-62-2	Piperine	-1	Training	-1	-	-	-
47	399-10-0	4'-Fluorochalcone	-1	Training	-1	-	-	-
48	2998-23-4	Pentyl acrylate	-1	Training	-1	-	-	-
49	2403-27-2	4'-Bromochalcone	-1	Training	-1	-	-	-
50	107-02-8	Acrolein	1	Training	1	1	Training	1
51	6606-59-3	1,6-Hexanediol dimethacrylate	-1	Training	-1	-	-	-
52	97-63-2	Ethyl methacrylate (EMA)	-1	Training	-1	-	-	-
53	1985-51-9	Neopentenediol dimethacrylate	-1	Training	-1	-	-	-
54	97055-37-3	3-(Dichloromethyl)-2,4,4-trichloro-2-butenic acid	1	Training	1	-	-	-
55	89811-25-6	N-Isobutyl-5-nitro-2-furanacrylamide	1	Training	1	-	-	-
56	623-15-4	Furfural acetone	-1	Training	-1	-	-	-
57	117823-31-1	(2Z)-2,4-Trichloro-3-formyl-2-butenic acid	1	Training	1	-	-	-
58	1774-66-9	4-Bromochalcone	-1	Training	-1	-	-	-
59	999-55-3	Allyl acrylate	-1	Training	-1	-	-	-
60	2657-25-2	4'-Hydroxychalcone	-1	Training	-1	-	-	-
61	1107-26-2	8'-Apo- β -carotenal	-1	Training	1	-	-	-
62	125974-06-3	4-(Chloromethyl)-5-hydroxy-2(5H)-furanone	1	Training	1	-	-	-
63	6197-30-4	2-Ethylhexyl	-1	Training	-1	-1	Training	-1
64	2358-84-1	2-cyano-3,3-diphenylacrylate	-1	Training	-1	-	-	-
65	2403-28-3	Diethylene glycol dimethacrylate	-1	Training	-1	-	-	-
66	683-51-2	4-Phenylchalcone	1	Training	1	-	-	-
67	90147-21-0	2-Chloroacrolein	1	Training	1	-	-	-
68	1466-88-2	5-Nitro-2-furanacrylic acid, isopropyl ester	1	Training	1	-	-	-
69	505-70-4	O-Nitrocinamaldehyde	1	Training	1	-	-	-
70	5234-68-4	Muconic acid	-1	Training	-1	-	-	-
71	97055-38-4	Carboxin	-1	Training	-1	-	-	-
		3-Chloro-4-(dichloromethyl)-5-methoxy-2(5H)-furanone	1	Training	1	-	-	-

72	2873-97-4	Diacetone acrylamide	-1	Training	-1	-	-	-	-
73	68053-32-7	Merulidial	1	Training	1	-1	-	Training	-1
74	1070-13-9	2-Propylacrolein	1	Training	1	-	-	-	-
75	19660-16-3	2,3-Dibromopropyl acrylate	1	Training	1	-	-	-	-
76	104-28-9	2-Ethoxyethyl	-1	Training	-1	-	-	-	-
77	3524-68-3	p-methoxycinnamate	-1	Training	-1	1	-	Training	1
78	142-09-6	Pentaerythritol triacrylate	-1	Training	-1	-	-	-	-
79	14129-84-1	N-Hexyl methacrylate	1	Training	1	-	-	-	-
80	122-57-6	2-(Dichloromethyl)-3,3-dichloro-2-propenal	1	Training	-1	-	-	-	-
81	24140-30-5	Methyl styryl ketone	1	Training	1	-	-	-	-
82	125974-08-5	(+)-2-Methylbutyl-4-methoxybenzylidene-4'-aminocinnamate	1	Training	1	1	-	Training	1
83	15625-89-5	3-Chloro-4-(chloromethyl)-5-hydroxy-2(5H)-furanone	1	Training	-1	-	-	-	-
84	126572-78-9	Trimethylolpropane triacrylate	1	Training	1	-1	1	Test	1
85	2223-82-7	(Z)-2-Chloro-3-(dichloromethyl)-4-hydroxybut-2-enoic acid	1	Training	1	-	-	-	-
86	710-25-8	Neopentanediol diacrylate	-1	Training	-1	-	-	-	-
87	2213-00-5	3-(5-Nitro-2-furyl)acrylamide	1	Training	1	-	-	-	-
88	614-47-1	Methyl marasmate	-1	Training	-1	-	-	-	-
89	13088-34-1	Benzylidene acetophenone	1	Training	1	-	-	-	-
90	4823-47-6	5-Dichloromethylene-2-furanone	1	Training	1	-	-	-	-
91	17831-71-9	2-Bromoethyl acrylate	-1	Training	-1	1	-	Training	1
92	1629-58-9	Tetraethyleneglycol diacrylate	1	Training	-1	-	-	-	-
93	2206-89-5	Ethylvinyl ketone	1	Training	-1	-	-	-	-
94	127072-60-0	2-Chloroethyl acrylate	1	Training	1	-	-	-	-
95	90147-18-5	10-Hydroxy-(2,4,6,8)dodecatetraenal	1	Training	1	-	-	-	-
96	110-17-8	5-Nitro-2-furanacrylic acid, propyl ester	-1	Training	-1	1	-	Training	1
97	6728-26-3	Fumaric acid	1	Training	1	-	-	-	-
98	62674-12-8	Trans-2-hexenal	1	Training	1	-	-	-	-
99	1576-87-0	3,4-Dichloro-2(5H)-furanone	1	Training	1	-	-	-	-
100	924-42-5	Trans-2-pentenal	-1	Training	-1	-	-	-	-
101	5392-40-5	N-Methylolacrylamide	-1	Training	-1	1	-	-	-
102	3695-86-1	Citral	-1	Training	-1	-	-	-	-
103	619-89-6	α -Cyano-2-furanacrylic acid, methyl ester	1	Training	1	-	-	-	-
104	79-10-7	4-Nitrocinnamic acid	-1	Training	-1	1	-	Training	1
105	2274-11-5	Acrylic acid	-1	Training	-1	-	-	Training	1
106	78-85-3	Ethylene glycol diacrylate	1	Training	1	-	-	Training	1
107	959-23-9	2-Methyl-2-propenal	-1	Training	-1	1	-	Training	1
		4'-Methoxychalcone	-1	Training	-1	-	-	-	-

Table 1 – (Continued)

Compound no.	CAS	Name	Observed AMES class.	Partition	Predicted AMES class.	Observed MCGM class.	Partition	Predicted MCGM class.
108	91134-58-6	5-Nitro-2-furanacrylic acid, isobutyl ester	1	Training	1	-	-	-
109	122-40-7	α -Amyl cinnamaldehyde	-1	Training	-1	-	-	-
110	97461-40-0	N-(1-Methylpropyl)-5-nitro-2-furanacrylamide	1	Training	1	-	-	-
111	90147-19-6	5-Nitro-2-furanacrylic acid, butyl ester	1	Training	1	-	-	-
112	96-33-3	Methyl acrylate	-1	Training	-1	1	Training	1
113	79-41-4	Methacrylic acid	-1	Training	-1	-	-	-
114	112309-61-2	3-Chloro-4-methyl-5-hydroxy-2(SH)-furanone	1	Training	1	-	-	-
115	2393-18-2	4-Aminocinnamic acid	-1	Training	-1	-	-	-
116	1152-48-3	4'-Nitrochalcone	1	Training	1	-	-	-
117	97461-41-1	N-(2,2-Dimethylpropyl)-5-nitro-2-furanacrylamide	1	Training	1	-	-	-
118	1222-98-6	4-Nitrochalcone	1	Training	1	-1	Training	-1
119	7085-85-0	Ethyl 2-cyanoacrylate	-1	Training	-1	-	-	-
120	2082-81-7	1,4-Butanediol dimethacrylate	-1	Training	-1	-	-	-
121	2499-95-8	Hexyl acrylate	-1	Training	-1	-	-	-
122	104-98-3	Urocanic acid	-1	Training	-1	-	-	-
123	6755-13-1	N,n-Dimethyl-5-nitro-2-furanacrylamide	1	Training	1	-	-	-
124	2157-01-9	N-Octyl methacrylate	-1	Training	-1	-	-	-
125	25870-67-1	4,4'-Dinitrochalcone	1	Training	1	-	-	-
126	557-48-2	Trans,cis-2,6-nonadienal	-1	Training	1	-	-	-
127	3160-37-0	Piperonyl acetone	-1	Training	-1	-	-	-
128	97461-42-2	N-Pentyl-5-nitro-2-furanacrylamide	1	Training	1	-	-	-
129	140-10-3	Trans-cinnamic acid	-1	Training	-1	1	Training	1
130	91642-47-6	5-Nitro-2-furanacrylic acid, pentyl ester	1	Training	1	-	-	-
131	1565-94-2	Bisphenol A diglycidyl dimethacrylate (bisGMA)	-1	Training	-1	-	-	-
132	6923-22-4	(E)-Monocrotophos	1	Training	1	-	-	-
133	147151-67-5	(3 β)-3-[[3-[9-[Bis(2-chloroethyl)amino]phenyl]-1-oxo-2-propenyl]oxy]-17 α -aza-d-homoandrost-5-en-1	1	Training	1	-	-	-
134	28564-83-2	2,3-Dihydro-3,5-dihydroxy-6-methyl-4h-pyran-4-one	1	Training	-1	-	-	-
135	142438-64-0	4-Ethyl-3-(methoxycarbonyl)-5-methyl-3,4-didehydro-gamma-butyrolactone	1	Training	-1	-	-	-
136	63-75-2	Arecoline	1	Training	-1	-	-	-

137	55557-02-3	N-Nitrosoguvacoline	1	Training	-1	-	-	-
138	2210-28-8	N-Propyl methacrylate	-1	Training	-1	-	-	-
139	97461-43-3	α -Methyl-5-nitro-2-furanacrylic acid, methyl ester	1	Training	1	-	-	-
140	125973-99-1	4-Chloromethyl-2(5H)-furanone	1	Training	1	-	-	-
141	3179-47-3	N-Decyl methacrylate	-1	Training	-1	-	-	-
142	78-94-4	Methyl vinyl ketone	1	Training	-1	-	-	-
143	101-39-3	α -Methylcinnamaldehyde	-1	Training	-1	-	-	-
144	90-65-3	Penicillic acid	-1	Training	-1	-	-	-
145	13048-33-4	1,6-Hexanediol diacrylate	-1	Training	-1	-	-	-
146	36840-85-4	1,5-Pentanediol diacrylate	-1	Training	-1	-	-	-
147	110-26-9	N,n'-Methylene-bis-acrylamide	1	Training	-1	-	-	-
148	14308-65-7	N-Methyl-5-nitro-2-furanacrylamide	1	Training	1	-	-	-
149	14901-07-6	β -Ionone	-1	Training	-1	-	-	-
150	97461-38-6	5-Nitro-2-furanacrylic acid, 1-methylpropyl ester	1	Training	1	-	-	-
151	53175-28-3	2-Chlorocrotonaldehyde	1	Training	1	-	-	-
152	3787-28-8	2,3,3-Trichloropropenal	1	Training	1	-	-	-
153	103-11-7	Mono(2-ethylhexyl) acrylate	-1	Training	-1	1	Training	-1
154	18031-40-8	1-Perillaldehyde	-1	Training	1	-	-	-
155	78-59-1	Isophorone	-1	Training	-1	-	Test	-1
156	959-33-1	4-Methoxychalcone	-1	Training	-1	-	-	-
157	1193-54-0	3,4-Dichloro-1h-pyrrole-2,5-dione	1	Training	1	-	-	-
158	458-37-7	Curcumin	-1	Training	-1	-	-	-
159	1615-02-7	P-Chlorocinnamic acid	-1	Training	-1	-	-	-
160	614-48-2	3-Nitrochalcone	-1	Training	1	-	-	-
161	555-66-8	Shogaol	1	Training	-1	-	-	-
162	6755-16-4	N-Ethyl-5-nitro-2-furanacrylamide	1	Training	1	-	-	-
163	122-69-0	Cinnamyl cinnamate	1	Training	-1	-	-	-
164	1663-39-4	Tert-butyl acrylate	-1	Training	-1	-	-	-
165	15743-13-2	2,2,4,5-Tetrachlorocyclopentene-1,3-dione	-1	Training	-1	-	-	-
166	97461-39-7	5-Nitro-2-furanacrylic acid, 1,1-dimethylethyl ester	1	Training	1	-	-	-
167	39965-42-9	N-(1,1-Dimethylethyl)-5-nitro-2-furanacrylamide	1	Training	1	-	-	-
168	31876-38-7	Moniliformin	-1	Training	-1	-	-	-
169	19337-19-0	4'-Acetamidochalcone	-1	Training	-1	-	-	-
170	122551-89-7	3-Chloro-4-(dichloromethyl)-2(5H)-furanone	1	Training	1	-	-	-
171	541-59-3	Maleimide	1	Training	-1	1	Test	1
172	2177-18-6	Acrylic acid vinyl ester	-1	Training	-1	-	-	-
173	137-66-6	Ascorbyl palmitate	-1	Training	-1	-	-	-
174	89-65-6	Erythorbic acid	-1	Training	-1	-	-	-

Table 1 – (Continued)

Compound no.	CAS	Name	Observed AMES class.	Partition	Predicted AMES class.	Observed MCGM class.	Partition	Predicted MCGM class.
175	80-71-7	3-Methylcyclopentane-1,2-dione hydrate	-1	Training	-1	-	-	-
176	97-88-1	N-Butyl methacrylate (BMA)	-1	Training	-1	-	-	-
177	23255-69-8	Nivalenol	-1	Test	-1	-	-	-
178	480-81-9	Seneciphylline	1	Test	-1	-	-	-
179	2439-35-2	2-(Dimethylamino)ethyl acrylate	-1	Test	-1	-	-	-
180	17341-40-1	1,1-Dimethyl-1-(2-hydroxypropylamine)methacrylimide	-1	Test	-1	-	-	-
181	1070-70-8	1,4-Butanediol diacrylate	-1	Test	-1	-	-	-
182	37962-27-9	5-Nitro-2-furanacrylic n-(5-nitro-2-furfurylidene)hydrazide	1	Test	1	-	-	-
183	2154-67-8	3-Carboxy-2,5-dihydro-2,2,5,5-tetramethyl-1h-pyrrol-1-yloxy	-1	Test	-1	-	-	-
184	514-78-3	Canthaxanthin	-1	Test	-1	-	-	-
185	584-79-2	Bioallethrin	1	Test	-1	-	-	-
186	1135-24-6	Ferulic acid	-1	Test	-1	-	-	-
187	1608-51-1	4-Fluorochalcone	-1	Test	-1	-	-	-
188	497-23-4	Butenolide	-1	Test	-1	-	-	-
189	7473-93-0	2-Nitrochalcone	-1	Test	1	-	-	-
190	1609-93-4	Cis-3-chloropropenoic acid	1	Test	1	-	-	-
191	87-56-9	Muochloric acid	1	Test	1	1	Training	1
192	1734-79-8	P-Nitrocinnamaldehyde	1	Test	1	1	Training	1
193	331-39-5	Caffeic acid	-1	Test	-1	1	Training	1
194	6203-18-5	P-Dimethylaminocinnamaldehyde	-1	Test	-1	-	-	-
195	129401-88-3	N-Acryloyl-n'-phenylpiperazine	-1	Test	-1	-	-	-
196	3066-70-4	2,3-Dibromopropyl methacrylate	1	Test	1	-	-	-
197	585-07-9	Tert-butyl methacrylate	-1	Test	-1	-	-	-
198	1874-12-0	5-Nitro-2-furanacrylic acid, ethyl ester	1	Test	1	-	-	-
199	1030-27-9	4-Dimethylaminochalcone	-1	Test	-1	-	-	-
200	818-61-1	2-Hydroxyethyl acrylate	-1	Test	-1	1	Training	1
201	766-40-5	3,4-Dichloro-5-hydroxy-2(5H)-furanone	1	Test	1	-	-	-
202	91182-09-1	N-Butyl-5-nitro-2-furanacrylamide	1	Test	1	-	-	-
203	105-76-0	Dibutyl (Z)-but-2-enedioate	-1	Test	-1	-	-	-
204	623-30-3	β-2-Furylacrolein	-1	Test	1	-	-	-
205	142-83-6	Trans,trans-2,4-hexadienal	1	Test	1	1	Training	-1
206	106-63-8	Isobutyl acrylate	-1	Test	-1	-	-	-

207	90147-31-2	N-Propyl-5-nitro-2-furanacrylamide	1	Test	1	-	-	-	-
208	1951-56-0	N-Isopropyl-5-nitro-2-furanacrylamide	1	Test	1	1	Test	1	1
209	2648-51-3	3,3-Dichloropropenal	1	Test	1	-	-	-	-
210	3290-92-4	Trimethylolpropane trimethacrylate	1	Test	-1	1	Training	1	1
211	1874-22-2	5-Nitro-2-furanacrolein	1	Test	1	-	-	-	-
212	108-31-6	Maleic anhydride	-1	Test	-1	-	-	-	-
213	77439-76-0	3-Chloro-4-(dichloromethyl)-5-hydroxy-2(5H)-furanone	1	Test	1	1	Training	1	1
214	4655-34-9	Isopropyl methacrylate	-1	Test	-1	-	-	-	-
215	4074-88-8	Diethylene glycol diacrylate	-1	Test	-1	-	-	-	-
216	1874-24-4	5-Nitro-2-furanacrylic acid, methyl ester	1	Test	1	-	-	-	-
217	96910-71-3	9-Hydroxyisovalleral	1	Test	1	-1	Training	-1	-1
218	125974-01-8	3-Chloro-4-(chloromethyl)-2(5H)-furanone	1	Test	1	-	-	-	-
219	922-63-4	2-Ethylacrolein	1	Test	1	-	-	-	-
220	327-97-9	Chlorogenic acid	-1	Test	-1	-1	Training	1	1
221	4170-30-3	Crotonaldehyde	-	-	-	-1	Training	1	1
222	303-34-4	Lasiocarpine	-	-	-	1	Training	-1	-1
223	14371-10-9	Trans-cinnamaldehyde	-	-	-	1	Training	1	1
224	110-44-1	Sorbic acid	-	-	-	-1	Training	-1	-1
225	108893-54-5	Acetylmerulidial	-	-	-	-1	Training	-1	-1
226	18409-46-6	(E,E)-Muconaldehyde	-	-	-	1	Training	1	1
227	33118-34-2	Polygodial	-	-	-	-1	Training	-1	-1
228	3588-17-8	(E,E)-Muconic acid	-	-	-	-1	Training	-1	-1
229	5956-39-8	Epipolygodial	-	-	-	-1	Training	-1	-1
230	62966-21-6	Fumaryl acetone	-	-	-	1	Training	1	1
231	120-57-0	Piperonal	-	-	-	1	Test	1	1
232	140-88-5	Ethyl acrylate	-	-	-	1	Test	1	1
233	79-06-1	Acrylamide	-	-	-	1	Test	1	1
234	37841-91-1	Isovelleral	-	-	-	1	Test	1	1
235	50656-61-6	Velleral	-	-	-	-1	Test	-1	-1

the Ames test result was positive while a compound was categorized as nonmutagen if exclusively negative Ames test results one or more were reported [58].

The other set employed comprises 48 compounds with α,β -unsaturated carbonyl moiety (Table 1), with published results from mammalian cells mutagenesis in L5178Y mouse lymphoma cells, CHO, AS52 and V79 lines of Chinese hamster cells, and was extracted from the Chemical Carcinogenesis Research Information System (<http://toxnet.nlm.nih.gov/cgi-bin/sis/htmlgen?CCRIS>). The compounds of this set were classified in the same way as the Ames test classification made by Bursi and coworkers.

In order to obtain validated QSAR models the data should be divided into training and test sets. In this work, we have applied the *k*-means cluster analysis technique towards designing both training and test sets that are representative of the entire experimental universe.

2.2. Computational strategies

The structures of all molecules were first drawn with the aid of Viewer ACD/3D software [59] followed by a fully optimization with the quantum-mechanics semi-empirical PM3 method implemented in MOPAC 6.0 [60]. Different families of descriptors were calculated using DRAGON software package [29]. Variables with constants or closed to constants values were deleted.

Stepwise linear discriminant analysis was applied to find the mathematical models that discriminate between actives and inactives. The replacement technique was used to select the variables (descriptors) with the highest influence on mutagenicity [61], but in contrast to regression analysis, which minimizes the standard deviation, we minimized the Wilk's Lambda. All these calculations were carried out with STATISTICA software [62].

One should remark that a statistical parameter is required, alike FIT in regression [63,64], which could be used to compare the quality of the different QSAR discriminant models. Here, we used a similar parameter, the so-called FIT(λ), defined by:

$$FIT(\lambda) = \frac{(1 - \lambda)(n - l - 1)}{(n + l^2)\lambda} \quad (1)$$

where *n* is the number of cases, *l* is the number of parameters of our model and λ is the Wilk's Lambda. So, reaching a higher value for this parameter means improving the usefulness of our model.

To measure the goodness of the training set and of the predictions we employed the following statistical measures:

- Sensitivity: the percentage of positives correctly predicted as positives.
- Specificity: the percentage of negatives correctly predicted as negatives.
- Concordance: the percentage of compounds correctly classified.
- Kappa (*K*) [65]: The kappa index excludes matching due solely to chance. The maximum possible agreement is $K = 1$. The value $K = 0$ is obtained when the agreement observed is that expected exclusively by chance. If the agreement

Table 2 – Interpretation of Kappa

Kappa	Agreement
<0	Less than chance agreement
0.01–0.20	Slight agreement
0.21–0.40	Fair agreement
0.41–0.60	Moderate agreement
0.61–0.80	Substantial agreement
0.81–0.99	Almost perfect agreement

is higher than expected simply because of chance, $K > 0$, while if it is less, $K < 0$. However, a commonly cited scale is represented in Table 2[66].

Good overall quality of the models is indicated by a small value of λ and high values of FIT(λ) and Kappa.

2.3. *k*-Means cluster analysis

Developing rational approaches for the selection of training and test set compounds is an active area of research. These approaches range, for instance, from straightforward random selection [67] to various clustering techniques [68]. The main goal of *k*-means cluster analysis (*k*-MCA) is to partition the original series of compounds into several statistically representative classes of chemicals, among which one might then select the training and test set compounds. The training set contained 80% (176/220 for AMES and 39/48 for MCGM) of the original data whereas the test set the remaining 20%. The *k*-MCA analysis was separately made for each group: mutagenic and non-mutagenic.

Starting from all descriptors of all 0–3D family types, those that produce the greatest separation of clusters meanwhile ensuring a statistically acceptable data partition were selected. In so doing, we took into account the number of members in each cluster and the standard deviation of the variables in the cluster (as low as possible). For AMES mutagenicity, *k*-MCA split the mutagenic compounds into five clusters comprising 21, 16, 27, 22 and 18 members with standard deviations of 0.14, 0.08, 0.12, 0.21 and 0.24, respectively, and the non-mutagenic compounds into six clusters comprising 27, 27, 16, 24, 12 and 10 members with standard deviations of 0.07, 0.11, 0.09, 0.11, 0.20 and 0.18, respectively. For MCGM mutagenicity, *k*-MCA split the positive compounds into five clusters comprising 6, 5, 9, 5 and 9 members with standard deviations of 0.12, 0.03, 0.11, 0.04 and 0.09, respectively, and the negative compounds into two clusters comprising 7 members each one with standard deviations of 0.26 and 0.27, respectively. Selection of the training and test sets was then carried out by taking compounds belonging to each cluster, proportionally to the size of the cluster. We also made an inspection of the standard deviation between and within clusters, the respective Fisher ratio and *p* level of significance (ought to be lower than 0.05) [69,70] (see Tables 3 and 4).

2.4. Applicability domain of the model

Given that the real utility of a QSAR model relies on its ability to accurately predict the modeled activity for new chemicals, careful assessment of the model's true predictive power is a must. This includes the model validation but also the defi-

Table 3 – Standard deviation between and within clusters, degrees of freedom (df), Fisher ratio (F) and level of significance (p) of the variables in the k-means cluster analysis for Ames test mutagenicity (AMES).

	Variable	Between SS	df	Within SS	df	F	Signif. p
Mutagenic	IC1	60.39	4	33.98	99.00	43.99	< 10 ⁻⁵
	ATS2v	83.36	4	11.58	99.00	178.23	< 10 ⁻⁵
	ATS5v	98.52	4	13.74	99.00	177.48	< 10 ⁻⁵
	ATS6v	97.23	4	6.96	99.00	345.85	< 10 ⁻⁵
	ATS2p	78.24	4	13.16	99.00	147.15	< 10 ⁻⁵
	ATS6p	95.45	4	7.58	99.00	311.45	< 10 ⁻⁵
	ATS3v	106.51	4	13.95	111.00	211.84	< 10 ⁻⁵
	ATS4v	92.85	4	10.97	111.00	234.82	< 10 ⁻⁵
Non-mutagenic	ATS3p	108.72	4	15.21	111.00	198.29	< 10 ⁻⁵
	ATS4p	93.73	4	11.08	111.00	234.82	< 10 ⁻⁵
	BELe8	97.82	4	42.14	111.00	64.41	< 10 ⁻⁵
	DP02	87.71	4	25.54	111.00	95.30	< 10 ⁻⁵
	Mor30v	42.83	4	71.62	111.00	16.60	< 10 ⁻⁵
	Mor15p	87.81	4	61.29	111.00	39.76	< 10 ⁻⁵

nition of the applicability domain of the model in the space of molecular descriptors used for deriving the model. There are several methods for assessing the applicability domain of QSAR/QSPR models [71,72] but the most common one encompasses determining the leverage values for each compound [73]. A Williams plot, i.e. the plot of standardized residuals versus leverage values (h), can then be used for an immediate and simple graphical detection of both the response outliers and structurally influential chemicals in the model. In this plot, the applicability domain is established inside a squared area within $\pm x$ standard deviations and a leverage threshold h^* (h^* is generally fixed at $3p/n$, where n is the number of training compounds and p the number of model parameters, whereas $x = 2$ or 3), lying outside this area (vertical lines) the outliers and (horizontal lines) influential chemicals. For future predictions, only predicted mutagenicity for chemicals belonging to the chemical domain of the training set should be proposed and used [74]. So, calculations of validation set classifications were performed only for those substances that had a leverage value below the threshold h^* .

2.5. Dofetilide displacement

In order to improve the predictions of this study, a series of combinations of the four best models [26] will be carried out to thus obtain new models superior in some respects, depending on the desired applications.

Following the O'Brien et al. report [26], we combined the individual models in each of three ways as follows:

- Recover positive model: this model returns a positive prediction if either model predicts a positive compound.
- Recover negative: this model returns a negative prediction if either model predicts a negative compound.
- Consensus model: this model returns predictions only if all models do agree. Two sets of statistics can be extracted from this model (consensus overall and consensus predicted). The first reflects the accuracy of the prediction of all the molecules in the validation set and the second reflects the accuracy of the prediction of the compounds which has the consensus overall forecasts.

3. Results and discussion

3.1. QSAR models

The task of building the best QSAR discriminatory models for a particular endpoint is a search for the best subset of descriptors that characterize the endpoint and the selection of the type and optimal parameters of the model. Achieving this requires experimentation to obtain the best solution. Of the eighteen families of different descriptors available from DRAGON, the best models attained from the cluster analysis for the two endpoints selected (AMES and MCGM) are shown in Tables 5 and 6. The meaning of each descriptor included in such models is shown in Table 7.

As can be judged, the best models for AMES mutagenicity were developed from only functional groups count variables, whereas for MCGM the best models were obtained from atom centered fragments pertaining variables. Both families, unlike

Table 4 – Standard deviation between and within clusters, degrees of freedom (df), Fisher ratio (F) and level of significance (p) of the variables in the k-means cluster analysis for mammalian cell gene mutagenicity test (MCGM).

	Variable	Between SS	df	Within SS	df	F	Signif. p
Mutagenic	BEHv5	31.98	4	0.76	29.00	304.66	< 10 ⁻⁵
	BELv5	31.66	4	1.39	29.00	165.67	< 10 ⁻⁵
	BEHe5	32.43	4	1.18	29.00	198.72	< 10 ⁻⁵
	BEHp5	31.75	4	0.77	29.00	297.44	< 10 ⁻⁵
	H1v	26.40	4	3.91	29.00	48.98	< 10 ⁻⁵
Non-mutagenic	Mor19u	10.34	1	4.04	12.00	30.74	1.27 × 10 ⁻⁴
	Mor19e	10.37	1	3.83	12.00	32.54	9.85 × 10 ⁻⁵

Table 5 – Best models derived for each family of descriptors for Ames test mutagenicity.

Family	No. variables	Training sens.	Training spec.	Training concor.	λ	Test sens.	Test spec.	Test concor.	Test sens.	Test spec.	Test concor.	No. compounds excluded	FIT (λ)	Kappa (κ)
Functional groups count	8	81	91	86	0.493	86	91	89	86	91	89	0	0.714	0.772
Atom centered fragments	8	80	90	85	0.500	67	74	70	74	85	79	5	0.695	0.588
3DMorSE	8	81	84	82	0.533	86	74	80	90	81	85	3	0.610	0.708
GETAWAY	7	75	88	82	0.545	71	87	80	75	87	81	1	0.624	0.624
2Dautocorrelations	7	78	88	84	0.582	76	83	80	76	83	80	0	0.536	0.589
Geometrical	6	80	80	80	0.584	71	78	75	75	82	79	2	0.568	0.569
RDF	8	81	78	80	0.624	86	70	77	86	80	83	3	0.420	0.658
WHIM	7	71	86	79	0.630	81	65	73	81	75	78	3	0.439	0.560
Constitutional	5	72	87	80	0.642	67	61	64	78	67	72	5	0.473	0.439
Burden eigenvalues	7	71	81	76	0.650	76	74	75	80	77	77	1	0.402	0.536
Information	4	72	81	77	0.655	76	74	75	76	74	75	0	0.469	0.500
Topologicos	3	71	80	76	0.736	57	70	64	63	70	67	2	0.333	0.327
Eigenvalue-based	2	64	81	73	0.827	72	81	80	65	77	80	2	0.200	0.425
Walk and path counts	4	55	84	70	0.844	67	74	70	74	81	78	4	0.164	0.548
Connectivity	2	54	84	70	0.887	57	78	68	67	78	73	3	0.122	0.452
Galvez topological charge	2	69	69	69	0.918	62	65	64	62	68	65	1	0.086	0.301
Molecular properties	2	86	44	64	0.921	81	17	48	85	18	50	2	0.083	0.031
Randit molecular profiles	1	58	56	57	0.949	48	52	50	53	60	56	5	0.053	0.126

* Results obtained taking into account all the substances of the test set.

** Results obtained taking into account only those substances that are within the applicability domain.

Table 6 – Best models derived for each family of descriptors for mammalian cells mutagenesis.

Family	No. variables	Training sens.	Training spec.	Training concor.	λ	Test sens.	Test spec.	Test concor.	Test sens.	Test spec.	Test concor.	No. compounds excluded	FIT (λ)	Kappa (κ)
Atom centered fragments	4	89	83	87	0.360	86	100	89	86	100	89	0	1.100	0.727
Functional groups count	4	100	67	90	0.397	86	100	89	86	100	89	0	0.938	0.727
Topologicos	4	96	75	90	0.401	86	100	89	86	100	89	0	0.921	0.727
Connectivity	4	100	75	92	0.420	86	100	89	86	100	89	0	0.855	0.727
GETAWAY	4	93	83	90	0.426	86	0	67	86	0	67	0	0.832	0.727
RDF	4	93	83	90	0.430	86	50	78	86	50	78	0	0.818	-0.174
Burden eigenvalues	4	85	100	90	0.447	71	100	78	83	100	88	1	0.357	0.766
2Dautocorrelations	4	93	83	90	0.451	71	50	67	71	100	75	1	0.752	0.714
Constitutional	4	96	67	87	0.467	86	100	89	86	100	89	0	0.706	0.385
Walk and path counts	4	93	75	87	0.474	71	100	78	81	100	78	0	0.686	0.727
Eigenvalue-based	4	89	75	85	0.492	86	100	89	86	100	89	0	0.639	0.526
Geometrical	4	93	67	85	0.505	71	0	56	71	0	56	0	0.606	0.727
3DMorSE	2	93	67	85	0.516	86	100	89	86	100	89	0	0.785	-0.286
WHIM	4	89	83	87	0.517	71	50	67	71	100	75	1	0.578	0.385
Galvez topological charge	3	89	75	85	0.574	71	50	67	71	50	67	0	0.542	0.182
Information	2	78	75	77	0.602	86	100	89	86	100	89	0	0.554	0.727
Molecular properties	2	85	75	82	0.612	86	100	89	86	100	89	0	0.531	0.727
Randit molecular profiles	4	85	58	77	0.620	86	50	78	86	50	78	0	0.379	0.357

* Results obtained taking into account all the substances of the test set.

** Results obtained taking into account only those substances that are within the applicability domain.

Table 7 – Symbols and definition for the QSAR variables involved in the models Eqs. (2) and (3) and in *k*-means cluster analysis.

Variable	Definition
IC1	Information content index (neighborhood symmetry of 1-order)
ATS2v	Broto-Moreau autocorrelation of a topological structure– lag 2/weighted by atomic van der Waals volumes
ATS3v	Broto-Moreau autocorrelation of a topological structure– lag 3/weighted by atomic van der Waals volumes
ATS4v	Broto-Moreau autocorrelation of a topological structure– lag 4/weighted by atomic van der Waals volumes
ATS5v	Broto-Moreau autocorrelation of a topological structure– lag 5/weighted by atomic van der Waals volumes
ATS6v	Broto-Moreau autocorrelation of a topological structure– lag 6/weighted by atomic van der Waals volumes
ATS2p	Broto-Moreau autocorrelation of a topological structure– lag 2/weighted by atomic polarizabilities
ATS3p	Broto-Moreau autocorrelation of a topological structure– lag 3/weighted by atomic polarizabilities
ATS4p	Broto-Moreau autocorrelation of a topological structure– lag 4/weighted by atomic polarizabilities
ATS6p	Broto-Moreau autocorrelation of a topological structure– lag 6/weighted by atomic polarizabilities
BEHe5	Highest eigenvalue n. 5 of Burden matrix/weighted by atomic Sanderson electronegativities
BELe8	Lowest eigenvalue n. 8 of Burden matrix/weighted by atomic Sanderson electronegativities
BEHv5	Highest eigenvalue n. 5 of Burden matrix/weighted by atomic van der Waals volumes
BELv5	Lowest eigenvalue n. 5 of Burden matrix/weighted by atomic van der Waals volumes
BEHp5	Highest eigenvalue n. 5 of Burden matrix/weighted by atomic polarizabilities
DP02	Molecular profile no. 02
Mor30v	3D-MoRSE – signal 30/weighted by atomic van der Waals volumes
Mor15p	3D-MoRSE– signal 15/weighted by atomic polarizabilities
Mor19u	3D-MoRSE – signal 19/unweighted
Mor19e	3D-MoRSE – signal 19/weighted by atomic Sanderson electronegativities
H1v	H autocorrelation of lag 1/weighted by atomic van der Waals volumes
nCb ⁻	Number of substituted benzene C(sp ²)
nRCHO	Number of aldehydes (aliphatic)
nArC=N	Number of imines (aromatic)
nArNO ₂	Number of nitro groups (aromatic)
nPO ₄	Number of phosphates
nCH ₂ RX	Number of CH ₂ RX
nCXr	Number of X–C– on ring
nCconjX	Number of X–C on conjugated C
C-015	=CH ₂
C-016	=CHR
C-039	Ar–C(=X)–R
H-046	H attached to C0(sp ³) no X attached to next C

3D descriptor variables, do not consider information on conformational aspects. Yet they can be derived from molecular structures using low computational resources, making them remarkable attractive in molecular modeling. Our resulting best-fit models are given below together with the statistical parameters of the classification.

$$\begin{aligned} \text{AMES} = & -0.499 \cdot \text{nCb}^- + 2.400 \cdot \text{nRCHO} + 8.099 \cdot \text{nArC=N} + \\ & + 5.533 \cdot \text{nArNO}_2 + 4.671 \cdot \text{nPO}_4 + 4.795 \cdot \text{nCH}_2\text{RX} - \\ & - 3.830 \cdot \text{nCXr} + 2.861 \cdot \text{nCconjX} - 1.969 \end{aligned} \quad (2)$$

$$N = 176; \lambda = 0.493; p < 10^{-5};$$

$$F = 21.431; \text{FIT}(\lambda) = 0.714; K = 0.772$$

$$\begin{aligned} \text{MCGM} = & 1.812 \cdot (\text{C-015}) - 1.165 \cdot (\text{C-016}) - 10.278 \cdot (\text{C-039}) - \\ & - 0.649 \cdot (\text{H-046}) + 5.564 \end{aligned} \quad (3)$$

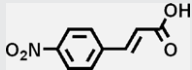
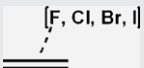
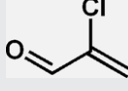
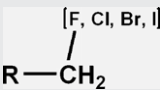
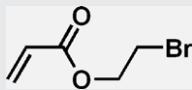
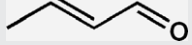
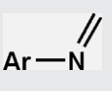
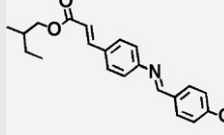
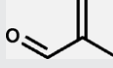

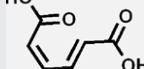

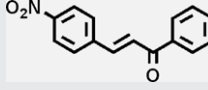
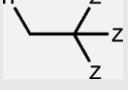
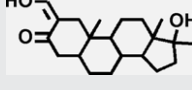
$$N = 39; \lambda = 0.359; p < 10^{-5};$$

$$F = 15.123; \text{FIT}(\lambda) = 1.100; K = 0.727$$

As seen, the models are reasonable in both statistical significance and goodness of prediction. A close inspection of the descriptors involved as well as the sign of their coefficients reveals also that the models are scientifically plausible. Indeed, one can establish how the presence or absence of a certain kind of bond affects the mutagenicity and thereby obtain a series of structural alerts for this activity in this family of compounds. For interpretation of the results, we will focus on the most significant descriptors that discriminate the two mutagenicity endpoints, namely for AMES: nArNO₂, nCconjX, nCH₂RX, nRCHO and nArC=N; and for MCGM: C-015, C-016, C-039 and H-046. The substructure representations pertaining to each of these descriptors are depicted in Table 8 along with examples of compounds containing them.

Regarding Ames mutagenicity Eq. (2), one can see that the presence of well-known structural alerts such as aromatic nitro groups (nArNO₂) and aromatic amines, more specifically imines (nArC=N), making such compounds mutagenic in the Ames test, through known enzymatic routes of metabolic activation (nitroreductases for nitroaromatic and cytochrome P-450 for the aromatic amines [75]) to hydroxylamine. Another well-known structural alert is the aliphatic halogenated derivatives, especially the primary known alkylating agents (nCH₂RX). Moreover, the halogen atom presence in the *alpha* or *beta* position of the double bound (nCconjX)

Table 8 – Substructure representations and example of compounds of the variables involved in both QSAR models.

Variable	Substructure representation*	Example compound
nArNO ₂	Ar-NO ₂	
nCconjX		
nCH ₂ RX		
nRCHO	R-CHO	
nArC=N		
C-015	=CH ₂	
C-016		
C-039		
H-046		

*X = heteroatom, Z = no heteroatom, Ar = aromatic and R = aliphatic

increases the mutagenicity in the Ames test [76], due to the cross-linking potential with another DNA or protein nucleophilic center [77]. Both relationships were previously reached by us [78] and by other authors [20,79–81]. Furthermore, as expected, aldehydes (nRCHO) are more mutagenic than other carbonyl groups, and the presence of this functional group increases the mutagenicity of the substance. Koleva et al. [14] arrived at the same conclusion by using a system of rules generated for the prediction of mutagenicity in *Salmonella typhimurium* strain TA100. The reactivity of the carbonyl group in electrophilic addition processes is influenced by the size and electronic effects of the substituents. Both steric and electronic factors favor aldehydes to be more reactive [14].

Looking now at the MCGM mutagenicity Eq. (3), one can see that mainly an increase of the number of hydrogen's attached to carbon sp³ (H-046) leads to an increase of the size of the molecule, and hence a reduction in the mutagenicity, most likely due to additional difficulty of passage of the substances into the cell and therefore their access to genetic material. Another features one can draw from this model, further supporting the mechanism of action of Michael type addition to the sulfhydryl of glutathione (GSH) or by an enzymatic reaction catalyzed by GSH transferase [82], is that the presence of a double bond in the terminal position of the chain (C-015)

favors mutagenicity, whereas alkyl substituent's at the double bond hinders this activity (C-016) by reducing the positive charge on the terminal carbon, the preferred site of nucleophilic attack [83,84]. Besides the presence of aromatic rings in carbonyl group (C-039) negatively affects the mutagenicity, possibly due to the charge delocalization by the aromatic ring that reduces its reactivity. This kind of structural rules are the heart of expert systems, such as DEREK [85], TOPKAT [86], Tox-tree [87,88], CASE [89,90] and MCASE [91,92], used to evaluate the toxicological profile of chemicals [93], so either models or structural features extracted in this work could be included in these expert systems for mutagenicity.

3.2. Applicability domain

It would be very interesting to have a predictive model for the vast majority of chemicals, particularly for those who have not been tested and, therefore, with unknown mutagenicity. Since this is usually not possible, one should define the applicability domain of the QSAR model, that is, the range within which it bears a new compound. For that purpose, we built a Williams plot using the leverage values calculated for each compound. As seen in Figs. 2 and 3, most of the compounds of the test set are within the applicability domain covered by ± 3 times

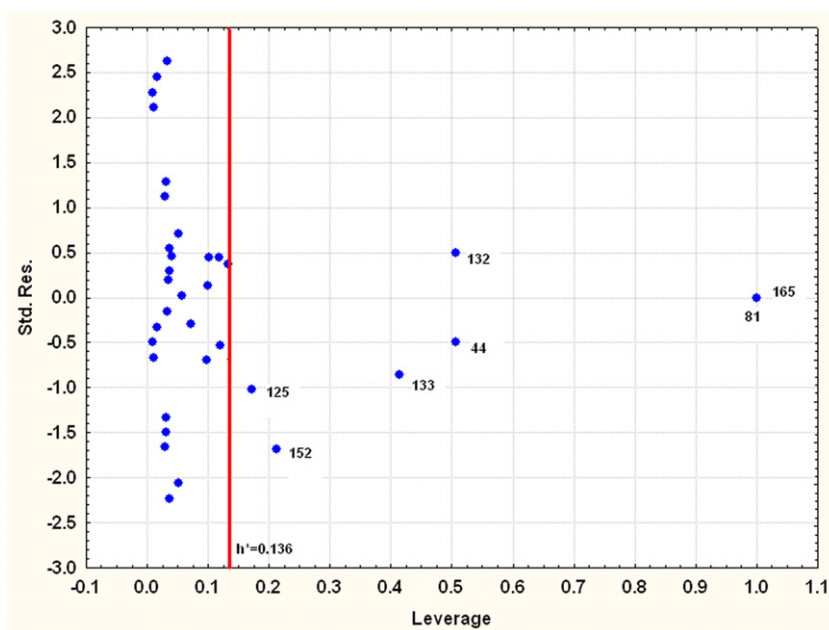


Fig. 2 – Applicability domain of the QSAR model for AMES Eq. (2).

the standard residual (σ) and the leverage threshold h^* ($=0.136$) and ($=0.307$) for Ames and MCGM endpoints, respectively, save for compounds 44, 81, 125, 132, 133, 152 and 165 (AMES) as well as compound 118 (MCGM). Even so, these compounds should not be considered outliers but influential chemicals [71].

Nevertheless, all evaluations pertaining to the external set were performed by taking into account the applicability domain of our QSAR model. So, if a chemical belonging to the test set had a leverage value greater than h^* , we considered that this means that the prediction is the result of substantial extrapolation and therefore may not be reliable [72].

3.3. Dofetilide displacement assay

Even though the predictive ability of individual models was good, we have applied the dofetilide displacement in an attempt to improve coverage on the predictions. By this method, we combined the best models to create more useful ones, since substances will be classified accurately by a family of descriptors and not by the others. The models chosen and the best results for the above combinations pertaining to Ames mutagenesis are shown in Table 9 and depicted in Fig. 4. Notice that the results for MCGM are not included since there were no noticeable differences

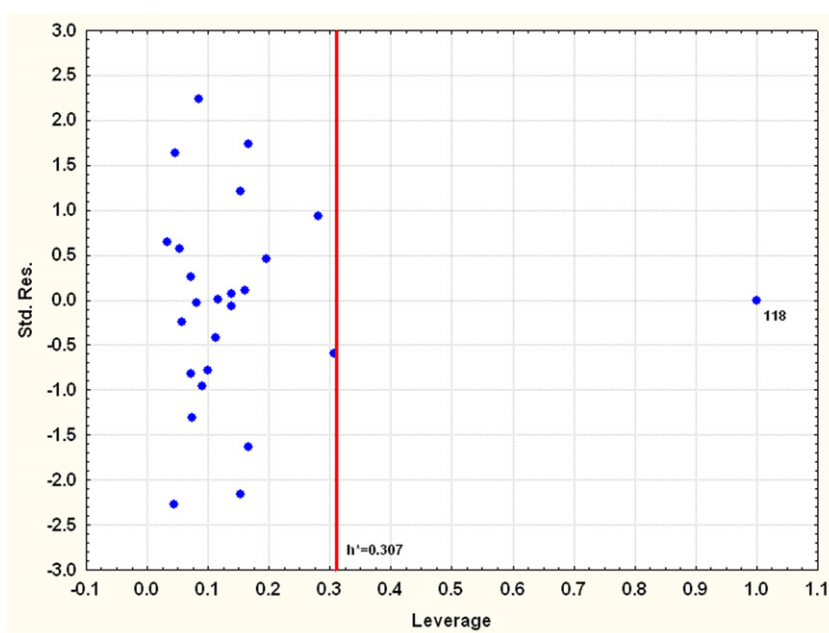


Fig. 3 – Applicability domain of the QSAR model for MCGM Eq. (3).

Table 9 – Selected model statistics for the dofetilide displacement assay prediction based on the Ames mutagenicity models taking into account all substances.

Family	Sensitivity (%)	Specificity (%)	Concordance (%)	Kappa
Functional groups counts	86	91	89	0.77
Atom centered	67	74	70	0.51
3DMoRSE	86	74	80	0.64
GETAWAY	71	87	80	0.60
Recover positive ^a	95	78	86	0.73
Recover negative ^b	76	100	89	0.77
Consensus-overall ^c	71	70	70	0.60
Consensus-predicted ^d	94	100	97	0.94

^a Model combines functional groups and 3DMoRSE

^b Model combines functional groups and GETAWAY

^c Model combines functional groups, 3DMoRSE and GETAWAY

^d Model combines functional groups, 3DMoRSE and GETAWAY and 76 % of positives and 70% of negatives predicted

with respect to the predictions attained by the individual models.

The recover positive model has a concordance of 86% and correctly classifies 95% of the mutagenic substances. This is better than all the individual models, making it extremely powerful if one is interested in identifying all possible positive substances or to find definitive non-mutagenic compounds. The recover negative correctly classifies all of the non-mutagenic substances and has a concordance equal to the best individual models. This model is then useful when one wishes to identify definitive mutagenic compounds. The consensus-overall model shows the lowest concordance of all the models (70%). However, though 27% of molecules are unclassified (as conflicting predictions have been made by the three individual models), when a prediction is made,

94% of positive compounds and 100% of negative compounds are accurately classified (Table 9), thus providing significantly increased confidence in prediction. Moreover, the false positive (5%) and false negative (0%) rates were reduced with this consensus model compared to individual models (Fig. 4).

In contrast to previous studies [17,18] the predictions can be made by the consensus model. However, we considered that our combinatorial QSAR modeling could be integrated with the previous QSAR models [17,18] in order to be used as a tool for the risk assessment of the other chemicals lacking experimental data. With these QSARs, a two-step prediction of mutagenicity it is possible: Step 1: yes/no activity from our discriminant functions, Step 2: if the answer from Step 1 is yes, then we carry out with the prediction of the degree of the mutagenic potency.

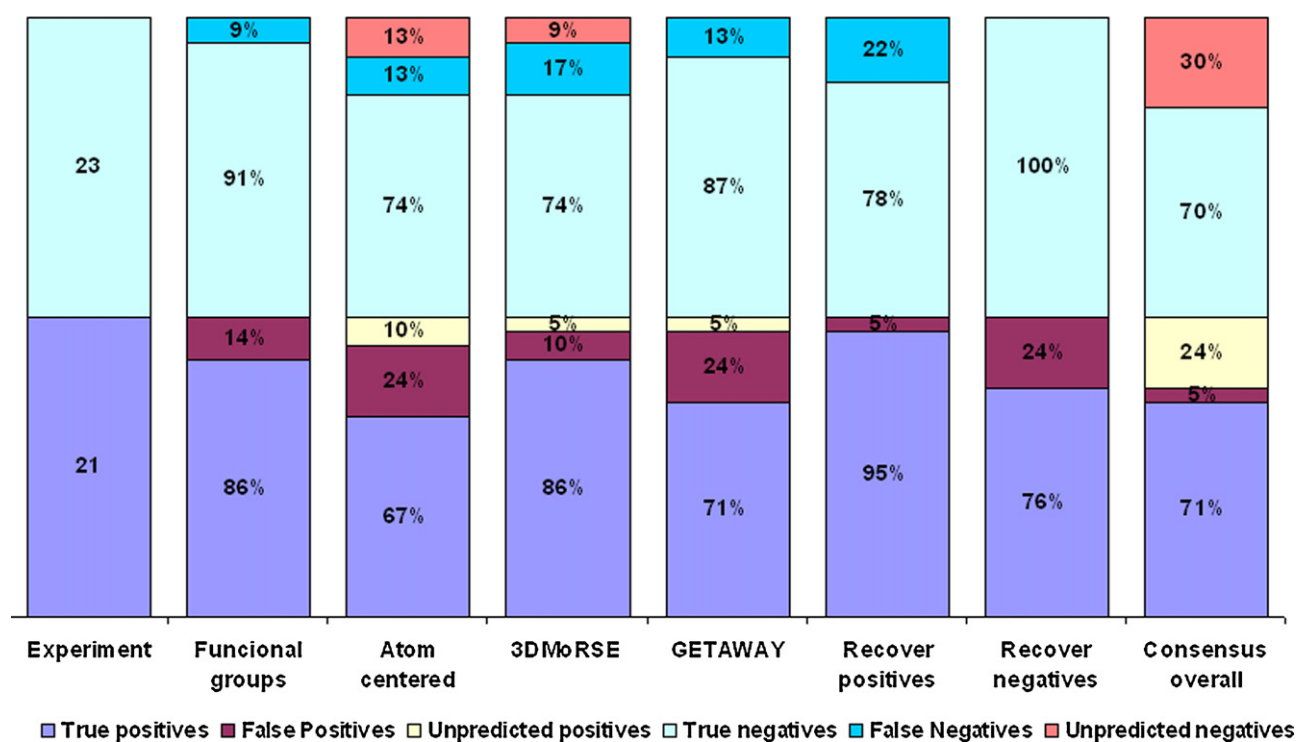


Fig. 4 – Statistics of dofetilide displacement assay prediction models.

4. Conclusions

In this study, we accomplished to model the mutagenic activity of acrylates, methacrylates and α,β -unsaturated carbonyl compounds using two different end points for estimating this toxicological activity.

The structural alerts obtained from the models indicate that the presence of nitroaromatic, aromatic imine, double primary bonds, aldehyde, primary monohalogenated moiety or halogens in the unsaturation adjacent to the carbonyl group in α or β position foster mutagenicity of these substances, thus their presence should be avoided in the design of new monomers. While features such as increased volume, the presence of alkyl groups in the unsaturation adjacent to the carbonyl group or benzenic rings in the carbonyl group reduce their mutagenicity. Therefore, one should increase the presence of these features in the new dental monomers to make them less genotoxic.

Besides the sub-structural features that were extracted from the resulting models, we have attained a high degree of success in predicting substances that were not used in the generation of the models (89% for both endpoints), as well as a confidentiality of 97% by combining the three best models for the Ames test mutagenicity.

In conclusion, this work has provided further evidence that the present QSAR models are attractive from the point of view of molecular modeling (low computational resources), and could be used for screening new candidates as dental monomers. Moreover, the structural features gathered from our models shall aid in the future design of new dental monomers that do not pose a human health risk and could be included in experts systems like DEREK, TOPKAT, Toxtree, CASE and MCASE for mutagenicity. Moreover, the consensus model could be integrated with the previous QSAR models to be used as tools for the risk assessment in a two-step prediction of mutagenicity.

Acknowledgement

A.M.H. acknowledges the *Portuguese Fundação para a Ciência e a Tecnologia* (FCT - Lisboa) (SFRH/BD/22692/2005) for financial support.

REFERENCES

- [1] Asmussen E. Factors affecting the quantity of remaining double bonds in restorative resin polymers. *Scand J Dent Res* 1982;90:490–6.
- [2] Imazato S, McCabe JF, Tarumi H, Ehara A, Ebisu S. Degree of conversion of composites measured by DTA and FTIR. *Dent Mater* 2001;17:178–83.
- [3] Hume WR, Gerzina TM. Bioavailability of components of resin-based materials which are applied to teeth. *Crit Rev Oral Biol Med* 1996;7:172–9.
- [4] Schweikl H, Schmalz G, Rackebrandt K. The mutagenic activity of unpolymerized resin monomers in *Salmonella typhimurium* and v79 cells. *Mutat Res* 1998;415:119–30.
- [5] Schweikl H, Schmalz G. Triethylene glycol dimethacrylate induces large deletions in the hprt gene of v79 cells. *Mutat Res* 1999;438:71–8.
- [6] EPA Chemical categories report. Available from: <http://www.epa.gov/opptintr/newchems/pubs/chemcat.htm>. 2006.
- [7] (EEC) EEC European council directive 1967/548/eec. 1967.
- [8] Chung FL, Roy KR, Hecht SS. A study of reactions of α,β -unsaturated carbonyl compounds with deoxyguanosine. *J Org Chem* 1988;53:14–7.
- [9] Dittberner U, Schmetzer B, Goélzer P, Eisenbrand G, Zankl H. Genotoxic effects of 2-trans-hexenal in human buccal mucosa cells in vivo. *Mutat Res* 1997;390:161–5.
- [10] Glaab V, Collins AR, Eisenbrand G, Janzowski C. DNA damaging potential and glutathione depletion of 2-cyclohexene-1-one in mammalian cells, compared to food relevant 2-alkenals. *Mutat Res* 2001;497:185–97.
- [11] Goélzer P, Janzowski C, Pool-Zobel BL, Eisenbrand G. (e)-2-Hexenal-induced dna damage and formation of cyclic 1,N²-(1,3-propano)-2'-deoxyguanosine adducts in mammalian cells. *Chem Res Toxicol* 1996;9:1207–13.
- [12] Ichihashi K, Osawa T, Toyokuni S, Uchida K. Endogenous formation of protein adducts with carcinogenic aldehydes: implications for oxidative stress. *J Biol Chem* 2001;276:23903–13.
- [13] Kautiainen A. Determination of hemoglobin adducts from aldehydes formed during lipid peroxidation in vitro. *Chem Biol Interact* 1992;83:55–63.
- [14] Koleva KY, Madden JC, Cronin MTD. Formation of categories from structure-activity relationships to allow read-across for risk assessment: Toxicity of α,β -unsaturated carbonyl compounds. *Chem Res Toxicol* 2008;21:2300–12.
- [15] Hansch C, Leo A. *Substituent constants for correlation analysis in chemistry and biology*. New York: John Wiley & Sons; 1979.
- [16] OECD The Report from the Expert Group on (Quantitative) structure activity relationship ([Q]SARs) on the principles for the validation of (Q)SARs, 49, OECD Series on Testing and Assessment. Paris: OECD, 2004.
- [17] Yourtee D, Holder AJ, Smith R, Morrill JA, Kostoryz E, Brockmann W, et al. Quantum mechanical quantitative structure-activity relationships to avoid mutagenicity in dental monomers. *J Biomater Sci Polymer Edn* 2001;12:89–105.
- [18] Holder AJ, Ye L. Quantum mechanical quantitative structure-activity relationships to avoid mutagenicity. *Dent Mater* 2009;25:20–5.
- [19] Helguera AM, González MP, Rieumont JB. Tops-mode approach to predict mutagenicity in dental monomers. *Polymer* 2004;45:2045–50.
- [20] González MP, Teran MCT, Fall Y, Dias LC, Helguera AM. A topological sub-structural approach to the mutagenic activity in dental monomers. 3. Heterogeneous set of compounds. *Bioorg Med Chem* 2005;46:2783–90.
- [21] Benigni R, Passerini L, Gallo G, Giorgi F, Cotta-Ramusino M. QSAR models for discriminating between mutagenic and nonmutagenic aromatic and heteroaromatic amines. *Environ Mol Mutagen* 1998;32:75–83.
- [22] Benigni R, Conti L, Crebelli R, Rodomonte A, Vari MR. Simple and α,β -unsaturated aldehydes: correct prediction of genotoxicity activity through structure-activity relationship models. *Environ Mol Mutagen* 2005;46:000–10.
- [23] Benigni R, Passerini L, Rodomonte A. Structure-activity relationships for the mutagenicity and carcinogenicity of simple and α,β -unsaturated aldehydes. *Environ Mol Mutagen* 2003;42:136–43.
- [24] Yang C, Hasselgren CH, Boyer S, Arvidson K, Aveston S, Dierkes P, et al. Understanding genetic toxicity through data mining: The process of building knowledge by integrating

- multiple genetic toxicity databases. *Toxicol Mech Meth* 2008;18:277–95.
- [25] Zhu H, Tropsha A, Fourches D, Varnek A, Papa E, Gramatica P, et al. Combinatorial QSAR modeling of chemical toxicants tested against *Tetrahymena pyriformis*. *J Chem Inf Model* 2008;48:766–84.
- [26] O'Brien SE, de Groot MJ. Greater than the sum of its parts: combining models for useful admet prediction. *J Med Chem* 2005;48:1287–91.
- [27] Mortelmans K, Zeiger E. The ames Salmonella/microsome mutagenicity assay. *Mutat Res* 2000;455:29–60.
- [28] Moore MM, DeMarini DM, DeSerres FJ, Tindall KR, editors. *Banbury Report 28: Mammalian Cell Mutagenesis*. New York: Cold Spring Harbor Laboratory, 1987.
- [29] Todeschini R, Consonni V, Pavan M. Dragon software version 5.4; 2002.
- [30] Kovatcheva A, Golbraikh A, Oloff S, Xiao YD, Zheng WF, Wolschann P, et al. Combinatorial QSAR of ambergris fragrance compounds. *J Chem Inf Comput Sci* 2004;44:582–95.
- [31] Duchowicz PR, González MP, Helguera AM, Cordeiro M, Castro EA. Application of the replacement method as novel variable selection in QSPR. 2. Soil sorption coefficients. *Chemometr Intell Lab Syst* 2007;88:197–203.
- [32] Pérez-Garrido A, Helguera AM, Cordeiro MNDS, Abellán A, Escudero AG. Convenient QSAR model for predicting the complexation of structurally diverse compounds with β -cyclodextrins. *Bioorg Med Chem* 2009;17: 896–904.
- [33] González MP, Suárez PL, Fall Y, Gómez G. Quantitative structure–activity relationship studies of vitamin d receptor affinity for analogues of $1\alpha,25$ -dihydroxyvitamin D_3 . *Bioorg Med Chem Lett* 2005;15:5165–9.
- [34] González MP, Terán C, Teijeira M, Helguera AM. Radial distribution function descriptors: an alternative for predicting α_2 adenosine receptors agonists. *Eur J Med Chem* 2006;41:56–62.
- [35] Lapinsh M, Prusis P, Mutule I, Mutulis F, Wikberg JES. QSAR and proteo-chemometric analysis of the interaction of a series of organic compounds with melanocortin receptor subtypes. *J Med Chem* 2003;46:2572–9.
- [36] Hemmateenejad B, Akhond M, Miri R, Shamsipur M. Genetic algorithm applied to the selection of factors in principal component-artificial neural networks: application to QSAR study of calcium channel antagonist activity of 1,4-dihydropyridines (nifedipine analogous). *J Chem Inf Comput Sci* 2003;43:1328–34.
- [37] Carotti A, Altomare C, Savini L, Chiasserini L, Pellerano C, Mascia MP, et al. High affinity central benzodiazepine receptor ligands. Part 3. Insights into the pharmacophore and pattern recognition study of intrinsic activities of pyrazolo[4,3-c]quinolin-3-ones. *Bioorg Med Chem* 2003;11:5259–72.
- [38] Wang XH, Tang Y, Xie Q, Qiu ZB. QSAR study of 4-phenylpiperidine derivatives as μ opioid agonists by neural network method. *Eur J Med Chem* 2006;41: 226–32.
- [39] Zheng F, Bayram E, Sumithran SP, Ayers JT, Zhan CG, Schmitt JD, et al. QSAR modeling of mono- and bis-quaternary ammonium salts that act as antagonists at neuronal nicotinic acetylcholine receptors mediating dopamine release. *Bioorg Med Chem* 2006;14:3017–37.
- [40] Liu HX, Gramatica P. QSAR study of selective ligands for the thyroid hormone receptor beta. *Bioorg Med Chem* 2007;15:5251–61.
- [41] Kline T, Andersen NH, Harwood EA, Bowman J, Malanda A, Endsley S, et al. Potent, novel in vitro inhibitors of the *Pseudomonas aeruginosa* deacetylase Ipxc. *J Med Chem* 2002;45:3112–29.
- [42] McElroy NR, Jurs PC. QSAR and classification of murine and human soluble epoxide hydrolase inhibition by urea-like compounds. *J Med Chem* 2003;46:1066–80.
- [43] Pirrung MC, Tumej LN, McClerren A, Raetz CRH. High-throughput catch-and-release synthesis of oxazoline hydroxamates. Structure–activity relationships in novel inhibitors of *Escherichia coli* Ipxc: in vitro enzyme inhibition and antibacterial properties. *J Am Chem Soc* 2003;125:1575–86.
- [44] Jain HK, Mourya VK, Agrawal RK. Inhibitory mode of 2-acetoxyphenyl alkyl sulfides against cox-1 and cox-2: QSAR analyses. *Bioorg Med Chem Lett* 2006;16:5280–4.
- [45] Hayatshahia SHS, Abdolmalekia P, Ghiasib M, Safarian S. QSARs and activity predicting models for competitive inhibitors of adenosine deaminase. *FEBS Lett* 2007;581:506–14.
- [46] Rameshwar U, Kadam RU, Garg D, Chavan A, Roy N. Evaluation of *Pseudomonas aeruginosa* deacetylase Ipxc inhibitory activity of dual pde4-tnfx inhibitors: A multiscreening approach. *J Chem Inf Model* 2007;47:1188–95.
- [47] Casanola-Martin GM, Marrero-Ponce Y, Khan MTH, Ather A, Khan KM, Torrens F, et al. Dragon method for finding novel tyrosinase inhibitors: biosilico identification and experimental in vitro assays. *Eur J Med Chem* 2007;42:1370–81.
- [48] Li JZ, Du J, Xi LL, Liu HX, Yao XJ, Liu MC. Validated quantitative structure–activity relationship analysis of a series of 2-aminothiazole based p56(lck) inhibitors. *Anal Chim Acta* 2009;631:29–39.
- [49] Goodarzi M, Freitas MP, Jensen R. Feature selection and linear/nonlinear regression methods for the accurate prediction of glycogen synthase kinase-3 beta inhibitory activities. *J Chem Inf Model* 2009;49:824–32.
- [50] Helguera AM, Duchowicz PR, Cabrera MA, Castro EA, Cordeiro N, González MP. Application of the replacement method as a novel variable selection strategy in QSAR. 1. Carcinogenic potential. *Chemometr Intell Lab Syst* 2006;81:180–7.
- [51] Helguera AM, Cabrera MA, González MP. Radial distribution function approach to predict rodent carcinogenicity. *J Mol Model* 2006;19:1–12.
- [52] Toropov AA, Rasulev BF, Leszczynski J. QSAR modeling of acute toxicity for nitrobenzene derivatives towards rats: comparative analysis by mlra and optimal descriptors. *QSAR Comb Sci* 2007;26:686–769.
- [53] Helguera AM, Cordeiro M, Cabrera MA, Combes RD, González MP. QSAR modelling of the rodent carcinogenicity of nitrocompounds. *Bioorg Med Chem* 2008;15:3395–407.
- [54] Helguera AM, Cordeiro M, González MP, Cabrera MA. Applications of 2d descriptors in drug design: a dragon tale. *Curr Top Med Chem* 2008;8:1628–55.
- [55] Helguera AM, Rodríguez-Borges J, García-Mera X, Fernández F, Cordeiro M. Probing the anticancer activity of nucleoside analogues: a QSAR model approach using an internally consistent training set. *J Med Chem* 2007;50:1537–45.
- [56] Cruz-Monteagudo M, Borges F, Cordeiro MNDS. Desirability-based multiobjective optimization for global QSAR studies: application to the design of novel NSAIDs with improved analgesic, antiinflammatory, and ulcerogenic profiles. *J Comput Chem* 2008;29:2445–59.
- [57] Cruz-Monteagudo M, Borges F, Cordeiro MNDS, Fajin JLC, Morell C, Ruiz RM, et al. Desirability-based methods of multiobjective optimization and ranking for global QSAR studies. Filtering safe and potent drug candidates from combinatorial libraries. *J Comb Chem* 2008;10: 897–913.
- [58] Kazius J, McGuire R, Bursi R. Derivation and validation of toxicophores for mutagenicity prediction. *J Med Chem* 2005;48:312–20.

- [59] AcD/chemsketch 5.12 and acd/3d viewer are freeware from advanced chemistry development inc. isis(tm) draw 2.4 was obtained on internet from mdl information systems, Inc.
- [60] Frank J. MOPAC. Seiler Research Laboratory, US Air Force Academy, Colorado, Springs Co., 1993.
- [61] Duchowicz PR, Castro EA, Fernndez FM. Alternative algorithm for the search of an optimal set of descriptors in QSAR-QSPR studies. *MATCH Commun Math Comput Chem* 2006;55:179–92.
- [62] Frank J. STATISTICA. Statsoft, Inc., 6.0 edition, 2002.
- [63] Kubinyi H. Variable selection in QSAR studies.1. an evolutionary algorithm. *Quant Struct Act Relat* 1994;13:285.
- [64] Kubinyi H. Variable selection in QSAR studies. 2. A highly efficient combination of systematic search and evolution. *Quant Struct Act Relat* 1994;13:393.
- [65] Cohen J. A coefficient of agreement for nominal scales. *J Educat Psychol Measure* 1960;20:37–46.
- [66] Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977;33:159–74.
- [67] Yasri A, Hartsough DJ. Toward an optimal procedure for variable selection and QSAR model building. *J Chem Inf Comput Sci* 2001;41:1218–27.
- [68] Gore PAJ. Handbook of applied multivariate statistics and mathematical modeling. USA: Academic Press; 2000. p. 298–318.
- [69] McFarland JW, Gans DJ. Chemometric methods in molecular design. Weinheim: VCH; 1995. p. 295–307.
- [70] Johnson RA, Wichern DW. Applied multivariate statistical analysis. New York: Prentice-Hall; 1988.
- [71] Eriksson L, Jaworska J, Worth AP, Cronin MTD, McDowell RM, Gramatica P. Methods for reliability and uncertainty assessment and for applicability evaluations of classification- and regression-based QSARs. *Environ Health Perspect* 2003;111:1361–75.
- [72] Netzeva TI, Worth AP, Aldenberg T, Benigni R, Cronin MTD, Gramatica P, et al. Current status of methods for defining the applicability domain of (quantitative) structure–activity relationships. *ATLA* 2005;33:155–73.
- [73] Gramatica P. Principles of QSAR models validation: internal and external. *QSAR Comb Sci* 2007;00:1–9.
- [74] Vighi M, Gramatica P, Consolaro F, Todeschini R. QSAR and chemometrics approaches for setting water quality objectives for dangerous chemicals. *Ecotoxicol Environ Saf* 2001;49:206–20.
- [75] Benigni R, Alessandro G, Franke R, Gruska A. Quantitative structure–activity relationships of mutagenic and carcinogenic aromatic amines. *Chem Rev* 2000;100:3697–714.
- [76] Eder E, Weinfurter E. Mutagenic and carcinogenic risk of oxygen containing chlorinated c-3 hydrocarbons: putative secondary products of c-3 chlorohydrocarbons and chlorination of water. *Chemosphere* 1994;29:2455–66.
- [77] Van Beerendonk GJM, Nivard MJM, Vogel EW, Nelson SD, Meerman JHN. Formation of thymidine adducts and cross-linking potential of 2-bromoacrolein, a reactive metabolite of tris(2,3-dibromopropyl)phosphate. *Mutagenesis* 1992;7:19–24.
- [78] Pérez-Garrido A, González MP, Escudero AG. Halogenated derivatives QSAR model using spectral moments to predict haloacetic acids (haa) mutagenicity. *Bioorg Med Chem* 2008;16:5720–32.
- [79] Mekenyan OG, Dimitrov SD, Pavlov TS, Veith GD. A systems approach to simulating metabolism in computational toxicology. I. The times heuristic modelling framework. *Curr Pharm Des* 2004;10:1273–93.
- [80] Benigni R. Chemical structure of mutagens and carcinogens and the relationship with biological activity. *J Exp Clin Cancer Res* 2004;23:5–8.
- [81] Mekenyan OG, Dimitrov SD, Schmieder P, Veith GD. In silico modelling of hazard endpoints: current problems and perspectives. *SAR QSAR Environ Res* 2003;14:361–71.
- [82] Ciaccio PJ, Gicquel E, O'Neill PJ, Scribner HE, Vandenberghe YL. Investigation of the positive response of ethyl acrylate in the mouse lymphoma genotoxicity assay. *Toxicol Sci* 1998;46:324–32.
- [83] Feron VJ, Til HP, de Vrijer F, Woutersen RA, Cassee FR, van Bladeren PJ. Aldehydes: occurrence, carcinogenic potential, mechanism of action and risk assessment. *Mutat Res* 1991;259:363–85.
- [84] Dearfield KL, Harrington-Brock K, Doerr CL, Rabinowitz JR, Moore MM. Genotoxicity in mouse lymphoma cells of chemicals capable of michael addition. *Mutagenesis* 1991;6:519–25.
- [85] Sanderson DM, Earnshaw CG. Computer prediction of possible toxic action from chemical structure; the derek system. *Human Exp Toxicol* 1991;10:261–73.
- [86] Enslein K, Gombar VK, Blake BW. Use of sar in computer-assisted prediction of carcinogenicity and mutagenicity of chemicals by the topkat program. *Mutat Res* 1994;305:47–61.
- [87] Benigni R, Bossa C. structure alerts for carcinogenicity, and the Salmonella assay system: a novel insight through the chemical relational databases technology. *Mutat Res* 2008;659:248–61.
- [88] Benigni R, Bossa C, Jeliakova NG, Netzeva TI, Worth AP, editors. The Benigni/Bossa rulebase for mutagenicity and carcinogenicity—a module of Toxtree. Luxembourg: EUR 23241 EN, EUR-Scientific and Technical Report Series, Office for the Official Publications of the European Communities, 2008.
- [89] Klopman G. Artificial intelligence approach to structure–activity studies. Computer automated structure evaluation of biological activity of organic molecules. *J Am Chem Soc* 1984;106:7315–21.
- [90] Klopman G, Rosenkranz HS. Structural requirements for the mutagenicity of environmental nitroarenes. *Mutation Res* 1984;126:227–38.
- [91] Klopman G. Multicase 1. A hierarchical computer automated structure evaluation program. *Quant Struct Activity Relationships* 1992;11:176–84.
- [92] Klopman G, Rosenkranz HS. Prediction of carcinogenicity/mutagenicity using multicase. *Mutat Res* 1994;305:33–46.
- [93] Tunkel T, Mayo K, Austin C, Hickerson A, Howard P. Practical considerations on the use of predictive models for regulatory purposes. *Environ Sci Technol* 2005;39:2188–99.



Contents lists available at ScienceDirect

Toxicology

journal homepage: www.elsevier.com/locate/toxicol

A topological substructural molecular design approach for predicting mutagenesis end-points of α , β -unsaturated carbonyl compounds

Alfonso Pérez-Garrido^{a,b,*}, Aliuska Morales Helguera^{c,d,e}, Gabriel Caravaca López^b, M.Natália D.S. Cordeiro^e, Amalio Garrido Escudero^a

^a Environmental Engineering and Toxicology Dpt., Catholic University of San Antonio, Guadalupe, Murcia, C.P. 30107, Spain

^b Department of Food and Nutrition Technology, Catholic University of San Antonio, Guadalupe, Murcia, C.P. 30107, Spain

^c Department of Chemistry, Faculty of Chemistry and Pharmacy, Central University of Las Villas, Santa Clara, 54830 Villa Clara, Cuba

^d Molecular Simulation and Drug Design Group, Chemical Bioactive Center, Central University of Las Villas, Santa Clara, 54830 Villa Clara, Cuba

^e REQUIMTE, Chemistry Department, Faculty of Sciences, University of Porto, 4169-007 Porto, Portugal

ARTICLE INFO

Article history:

Received 20 October 2009

Received in revised form

29 November 2009

Accepted 30 November 2009

Available online 11 December 2009

Keywords:

QSAR

Mutagenicity

TOPS-MODE

α , β -Unsaturated carbonyl compounds

REACH

ABSTRACT

Chemically reactive, α , β -unsaturated carbonyl compounds are common environmental pollutants able to produce a wide range of adverse effects, including, e.g. mutagenicity. This toxic property can often be related to chemical structure, in particular to specific molecular substructures or fragments (alerts), which can then be used in specialized software or expert systems for predictive purposes. In the past, there have been many attempts to predict the mutagenicity of α , β -unsaturated carbonyl compounds through quantitative structure activity relationships (QSAR) but considering only one exclusive endpoint: the Ames test. Besides, even though those studies give a comprehensive understanding of the phenomenon, they do not provide substructural information that could be useful for improving expert systems based on structural alerts (SAs). This work reports an evaluation of classification models to probe the mutagenic activity of α , β -unsaturated carbonyl compounds over two endpoints – the Ames and mammalian cell gene mutation tests – based on linear discriminant analysis along with the topological Substructure molecular design (TOPS-MODE) approach. The obtained results showed the better ability of the TOPS-MODE approach in flagging structural alerts for the mutagenicity of these compounds compared to the expert system TOXTREE. Thus, the application of the present QSAR models can aid toxicologists in risk assessment and in prioritizing testing, as well as in the improvement of expert systems, such as the TOXTREE software, where SAs are implemented.

© 2009 Elsevier Ireland Ltd. All rights reserved.

1. Introduction

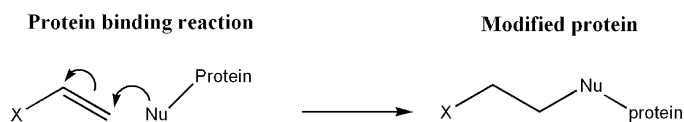
α , β -Unsaturated carbonyl compounds are common environmental pollutants, often used in the synthesis of chemicals, solvents, food additives, disinfectants and dental restorative materials (Feron et al., 1991; Boelens and Gemert, 1987; van Noort et al., 1990). These compounds possess a strongly polarized carbon-oxygen double bond due to the presence of an additional double bond between carbons 2 and 3 (i.e. α and β), which makes them even more reactive than simple carbonyls. Because of their particular reactivity, they are able to interact with electron-rich biological macromolecules and a wide range of adverse effects has been reported, including for instance mutagenicity. In general, these mutagenic compounds act through a Michael type addition mech-

anism (see Fig. 1) but the type of substituents in the α or β -carbons of the unsaturated carbonyl moiety significantly affects the effectiveness of the reaction (Aptula and Roberts, 2006).

Regulatory bodies use various endpoints as standard tests for screening chemicals for potential mutagenic effects. The four common assay types employed are bacterial mutagenesis, mammalian mutagenesis, *in vitro* chromosome aberration and *in vivo* micronucleous. All have in common the ability to identify those substances that produce some sort of alteration on DNA. Owing to the cost in both resources and time required in such mutagenic assays, there has been a remarkable upsurge in interest in alternative non-animal approaches as tools for speeding up, at least, priority setting and risk assessment. One such set of tools, strongly encouraged under the framework of the European Union's REACH (Registration, Evaluation, Authorisation and restriction of Chemicals) legislation, comprises *in silico* prediction of mutagenicity, based on (Quantitative) Structure–Activity Relationships [(Q)SAR] modelling (OECD, 2007). QSAR modelling seeks to discover and use mathematical relationships between molecular properties of the compounds (descriptors) and the activity or property of interest.

* Corresponding author at: Environmental Engineering and Toxicology Dpt., Catholic University of San Antonio, Guadalupe, Murcia, C.P. 30107, Spain. Tel.: +34 968 278 755.

E-mail address: Aperez@pdi.ucam.edu (A. Pérez-Garrido).



Double or triple bond with electron-withdrawing substituent X, such as $-\text{CHO}$, $-\text{COR}$, $-\text{CO}_2\text{R}$, $-\text{CN}$, $-\text{SO}_2\text{R}$, $-\text{NO}_2$, etc. Includes ortho- or para-quinones, often formed by oxidation of ortho- or para-dihydroxy aromatics acting as pro-Michael acceptors. X can also be a heterocyclic group such as ortho- or para-pyridino.

Fig. 1. Michael type addition reaction and corresponding applicability domain.

In the past, there have been several QSAR studies aimed at modelling the mutagenicity of α , β -unsaturated carbonyl compounds. These QSARs can be subdivided into two types of models, i.e. for (1) the gradation of potency of active chemicals (Yourtee et al., 2001; Holder and Ye, 2009; Helguera et al., 2004; González et al., 2005b) and for (2) the discrimination between active (positive) and inactive (negative) chemicals (Benigni et al., 2005; Koleva et al., 2008). One should notice here that, previous work has shown that the structural effects that modulate the mutagenic potency are normally different from those that distinguish positive/negative active chemicals, and so the first type of models are not useful to set up mutagenic activity the first issue in risk assessment (Benigni et al., 1998).

The only attempts towards modelling the mutagenic activity of this family of substances have been based on data collected from the Salmonella typhimurium Ames test. Benigni et al. (2005) developed a QSAR prediction model for 25 α , β -unsaturated aldehydes by stepwise linear discriminant analysis based on data from TA100 strain assays. The results of this study indicated a dependency between mutagenicity, hydrophobicity and molecular volume. More recently, a study has appeared to model and predict the Ames TA100-derived mutagenicity for 45 α , β -unsaturated carbonyl compounds (Koleva et al., 2008). In this study, the set of compounds was divided into several sub-groups, namely: (1) halogenated derivatives, (2) nitro derivatives of cinnamaldehyde and (3) related acroleins, and a system of rules was developed for the classification based on their reactivity and mechanisms of action.

All the above studies only considered as endpoint the Ames test for estimating the mutagenic activity. Even though that is understandable for two of the other endpoints – *in vitro* chromosome aberration and *in vivo* micronucleus – since they do not provide enough data (only for 6 and 7 compounds, respectively) to set up valid QSAR models, that is not so for the MGGM endpoint (data for 45 compounds). What's more, information across multiple endpoints seems to be needed to reach more realistic predictions about mutagenicity (Yang et al., 2008). In addition, those studies resorted to global molecular descriptors (i.e.: molecular refractivity, the logarithm of the octanol/water partition coefficient $\log P$, etc.) and many of them were unable to perceive how each fragment or functional group influence a particular molecular structure of interest. One way to overcome the latter problem is to use Structural Alerts-based SAR studies such as the recently implemented expert system: TOXTREE (Benigni and Bossa, 2008; Benigni et al., 2008). Structural Alerts (SA) are molecular substructures or functional groups that are related to the toxic properties of the chemicals, that is to say, a sort of “codes” embodying long series of studies aimed at highlighting their mechanisms of action. This SA-based expert system has already pulled together very good results in predictions of mutagenic/carcinogenic chemicals (Benigni and Bossa, 2008). Even so, it has been recognised also that further work is still required to improve the knowledge about modulating factors (Benigni and Bossa, 2008; Benigni et al., 2008), that is to say, about the chemical functionalities that may annihilate the toxic effects of the SAs when they are present simultaneously in the same molecule.

This work aims at discriminating the mutagenic activity of α , β -unsaturated carbonyl compounds, based on two different endpoints together with a substructural approach such as the Topological Substructural Molecular Design (TOPS MODE) (Estrada, 1996, 1997, 1995) descriptors. These descriptors can and have proved to be very useful in QSAR modelling of a broad range of toxicities (Estrada et al., 2001, 2003a,b, 2004; Estrada and Uriarte, 2001; González et al., 2005a, 2006; Helguera et al., 2005, 2006, 2007, 2008a,b; Estrada and Molina, 2006; García-Lorenzo et al., 2008; Sosted et al., 2004), including mutagenicity (González et al., 2004a,b, 2005b; Helguera et al., 2004; Pérez-Garrido et al., 2008). Furthermore, the TOPS MODE approach is able to transform simple molecular descriptors, such as $\log P$, polar surface area, charges, etc., into series of descriptors that account for the distribution of the related characteristics (hydrophobicity, polarity, electronic effects, etc.) across the molecule. In fact, such approach has been recognised recently by the Organisation for Economic Co-operation and Development (OECD) as providing “a mechanistic interpretation at a bond level” and enabling “the generation of new hypotheses such as structural alerts” (OECD, 2007). Thus, by gathering structural information at a local level from the models developed, we will be able to identify SAs as well as to quantify their accompanying modulating factors. The results of this work can then improve the expert systems where these SAs are implemented.

2. Materials and methods

2.1. Mutagenicity data sets

The two sets of data include various substances (220 for the Ames test mutagenicity -AMES-, and 48 for the mammalian cell gene mutation test -MCGM-) with α , β -unsaturated carbonyl moiety (Table 1). AMES data was derived from the Ames test classification made by Kazius et al. (2005) for mutagenicity, being their analysis restricted to the *Salmonella typhimurium* strains TA98, TA100, TA1535 and either TA1537 or TA97, performed with the standard plate or preincubation method either with or without a metabolic activation mixture. In that classification, a compound was categorized as a mutagen if at least one Ames test result was positive and non-mutagen if exclusively negative Ames test results -one or more- were reported (Kazius et al., 2005).

On the other hand, MCGM data was collected from compounds with published results from mammalian cells mutagenesis in L5178Y mouse lymphoma cells, CHO, AS52 and V79 lines of Chinese hamster cells extracted from the Chemical Carcinogenesis Research Information System (available at <http://toxnet.nlm.nih.gov/cgi-bin/sis/htmlgen?CCRIS>). The classification was performed in the same manner as the Ames test classification made by Kazius et al. (2005).

2.2. The TOPS-MODE approach

The TOPS-MODE approach is based on computing the spectral moments of the topological bond matrix (Estrada, 1995). The mathematical details of this approach have been documented in detail previously (Estrada, 1996, 1997), but an overview highlighting only the most important aspects will be given here.

Firstly, the molecular structure of each compound is represented by its molecular graph and then, the bond adjacency matrix (B) is derived. B is a squared symmetric matrix whose entries are ones or zeros if the corresponding bonds are adjacent or not. The order of this matrix (m) is the number of bonds in the molecular graph, being two bonds adjacent if they are incident to a common atom. Furthermore, weights are introduced in the diagonal entries of this matrix to mirror fundamental physicochemical properties that might relate to the target endpoint

Table 1
CAS number, observed and predicted classification, and leverage values for the compounds used for obtaining the final QSAR models for the two endpoints (AMES: Eq. (7) and MCGM: Eq. (8)).

Compound no.	CAS	Observed		TOPS-MODE		TOXTREE	Observed		Predicted	
		AMES Class.	Partition	AMES Class.	Leverage		AMES Class.	MCGM Class.	Partition	MCGM Class.
1	87406-72-2	1	Training	1	–	1	–	–	–	–
2	23282-20-4	–1	Training	–1	–	1	–	–	–	–
3	34807-41-5	–1	Training	–1	–	1	–	–	–	–
4	6379-69-7	–1	Training	–1	–	1	–	–	–	–
5	63166-73-4	–1	Training	–1	–	1	–	–	–	–
6	23246-96-0	1	Training	–1	–	–1	–	–	–	–
7	130-01-8	–1	Training	–1	–	–1	–	–	–	–
8	21794-01-4	–1	Training	–1	–	–1	–	–	–	–
9	61203-01-8	1	Training	1	–	1	–	–	–	–
10	2849-98-1	–1	Training	–1	–	–1	–	–	–	–
11	97-90-5	–1	Training	–1	–	–1	1	Training	1	–
12	4513-36-4	–1	Training	–1	–	–1	–	–	–	–
13	13675-34-8	–1	Training	–1	–	–1	–	–	–	–
14	868-77-9	–1	Training	–1	–	–1	–	–	–	–
15	5466-77-3	–1	Training	–1	–	–1	–	–	–	–
16	123-73-9	1	Training	1	–	1	–	–	–	–
17	96910-73-5	1	Training	1	–	1	–	–	–	–
18	14925-39-4	1	Training	1	–	1	–	–	–	–
19	2397-76-4	–1	Training	–1	–	–1	–	–	–	–
20	68162-37-8	1	Training	1	–	1	–	–	–	–
21	836-37-3	1	Training	1	–	1	–	–	–	–
22	29590-42-9	–1	Training	–1	–	–1	–1	Training	1	–
23	555-68-0	1	Training	1	–	1	–	–	–	–
24	58-54-8	–1	Training	1	–	1	–	–	–	–
25	3688-53-7	1	Training	1	–	1	1	Training	1	–
26	18829-55-5	1	Training	–1	–	1	–	–	–	–
27	1013-96-3	–1	Training	1	–	1	–	–	–	–
28	102059-18-7	1	Training	–1	–	1	–	–	–	–
29	7364-09-2	1	Training	1	–	1	–	–	–	–
30	499-12-7	–1	Training	–1	–	–1	–	–	–	–
31	10443-65-9	–1	Training	–1	–	–1	–	–	–	–
32	6281-23-8	1	Training	1	–	1	–	–	–	–
33	109460-96-0	–1	Training	–1	–	–1	–	–	–	–
34	5443-49-2	1	Training	1	–	1	–	–	–	–
35	645-62-5	–1	Training	–1	–	1	–	–	–	–
36	97-86-9	–1	Training	–1	–	–1	–	–	–	–
37	137-05-3	1	Training	1	–	–1	–	–	–	–
38	20426-12-4	–1	Training	–1	–	–1	–	–	–	–
39	104-55-2	–1	Training	–1	–	–1	–	–	–	–
40	488-11-9	1	Training	1	–	1	–	–	–	–
41	109-16-0	–1	Training	–1	–	–1	1	Training	1	–
42	434-07-1	–1	Training	–1	–	1	–1	Training	–1	–
43	126572-80-3	1	Training	1	–	1	–	–	–	–
44	13171-21-6	1	Training	1	–	1	–	–	–	–
45	141-32-2	–1	Training	–1	–	–1	1	Training	1	–
46	94-62-2	–1	Training	1	–	1	–	–	–	–
47	399-10-0	–1	Training	–1	–	–1	–	–	–	–
48	2998-23-4	–1	Training	–1	–	–1	–	–	–	–
49	2403-27-2	–1	Training	–1	–	–1	–	–	–	–
50	107-02-8	1	Training	1	–	1	1	Training	1	–
51	6606-59-3	–1	Training	–1	–	–1	–	–	–	–
52	97-63-2	–1	Training	–1	–	–1	–	–	–	–
53	1985-51-9	–1	Training	–1	–	–1	–	–	–	–
54	97055-37-3	1	Training	1	–	1	–	–	–	–
55	89811-25-6	1	Training	1	–	1	–	–	–	–
56	623-15-4	–1	Training	–1	–	–1	–	–	–	–
57	117823-31-1	1	Training	1	–	1	–	–	–	–
58	1774-66-9	–1	Training	–1	–	–1	–	–	–	–
59	999-55-3	–1	Training	–1	–	–1	–	–	–	–
60	2657-25-2	–1	Training	–1	–	–1	–	–	–	–
61	1107-26-2	–1	Training	–1	–	–1	–	–	–	–
62	125974-06-3	1	Training	1	–	1	–	–	–	–
63	6197-30-4	–1	Training	–1	–	–1	–1	Training	–1	–
64	2358-84-1	–1	Training	–1	–	–1	–	–	–	–
65	2403-28-3	–1	Training	–1	–	–1	–	–	–	–
66	683-51-2	1	Training	1	–	1	–	–	–	–
67	90147-21-0	1	Training	1	–	1	–	–	–	–
68	1466-88-2	1	Training	1	–	1	–	–	–	–
69	505-70-4	–1	Training	–1	–	–1	–	–	–	–
70	5234-68-4	–1	Training	1	–	1	–	–	–	–
71	97055-38-4	1	Training	1	–	1	–	–	–	–
72	2873-97-4	–1	Training	–1	–	1	–	–	–	–

Table 1 (Continued)

Compound no.	CAS	Observed		TOPS-MODE		TOXTREE	Observed		Predicted	
		AMES Class.	Partition	AMES Class.	Leverage		AMES Class.	MCGM Class.	Partition	MCGM Class.
73	68053-32-7	1	Training	1	–	1	–1	Training	–1	–
74	1070-13-9	1	Training	1	–	1	–	–	–	–
75	19660-16-3	1	Training	1	–	1	–	–	–	–
76	104-28-9	–1	Training	–1	–	–1	–	–	–	–
77	3524-68-3	–1	Training	–1	–	–1	1	Training	1	–
78	142-09-6	–1	Training	–1	–	–1	–	–	–	–
79	14129-84-1	1	Training	1	–	1	–	–	–	–
80	122-57-6	1	Training	–1	–	–1	–	–	–	–
81	24140-30-5	1	Training	–1	–	–1	–	–	–	–
82	125974-08-5	1	Training	1	–	1	1	Training	1	–
83	15625-89-5	1	Training	–1	–	–1	1	Test	1	0.139
84	126572-78-9	1	Training	1	–	1	–	–	–	–
85	2223-82-7	–1	Training	–1	–	–1	–	–	–	–
86	710-25-8	1	Training	1	–	1	–	–	–	–
87	2213-00-5	1	Training	1	–	1	–	–	–	–
88	614-47-1	–1	Training	–1	–	–1	–	–	–	–
89	13088-34-1	1	Training	1	–	–1	–	–	–	–
90	4823-47-6	1	Training	1	–	1	–	–	–	–
91	17831-71-9	–1	Training	–1	–	–1	1	Training	1	–
92	1629-58-9	1	Training	–1	–	1	–	–	–	–
93	2206-89-5	1	Training	1	–	1	–	–	–	–
94	127072-60-0	1	Training	–1	–	–1	–	–	–	–
95	90147-18-5	1	Training	1	–	1	–	–	–	–
96	110-17-8	–1	Training	–1	–	–1	1	Training	1	–
97	6728-26-3	1	Training	1	–	1	–	–	–	–
98	62674-12-8	1	Training	1	–	–1	–	–	–	–
99	1576-87-0	1	Training	1	–	1	–	–	–	–
100	924-42-5	–1	Training	–1	–	1	–	–	–	–
101	5392-40-5	–1	Training	–1	–	–1	–	–	–	–
102	3695-86-1	–1	Training	–1	–	–1	–	–	–	–
103	619-89-6	1	Training	1	–	1	–	–	–	–
104	79-10-7	–1	Training	–1	–	–1	1	Training	1	–
105	2274-11-5	–1	Training	–1	–	–1	1	Training	1	–
106	78-85-3	1	Training	1	–	1	1	Training	1	–
107	959-23-9	–1	Training	–1	–	–1	–	–	–	–
108	91134-58-6	1	Training	1	–	1	–	–	–	–
109	122-40-7	–1	Training	–1	–	–1	–	–	–	–
110	97461-40-0	1	Training	1	–	1	–	–	–	–
111	90147-19-6	1	Training	1	–	1	–	–	–	–
112	96-33-3	–1	Training	1	–	–1	1	Training	1	–
113	79-41-4	–1	Training	–1	–	–1	–	–	–	–
114	112309-61-2	1	Training	1	–	–1	–	–	–	–
115	2393-18-2	–1	Training	–1	–	1	–	–	–	–
116	1152-48-3	1	Training	1	–	1	–	–	–	–
117	97461-41-1	1	Training	1	–	1	–	–	–	–
118	1222-98-6	1	Training	1	–	1	–1	Training	–1	–
119	7085-85-0	–1	Training	–1	–	–1	–	–	–	–
120	2082-81-7	–1	Training	–1	–	–1	–	–	–	–
121	2499-95-8	–1	Training	–1	–	–1	–	–	–	–
122	104-98-3	–1	Training	–1	–	–1	–	–	–	–
123	6755-13-1	1	Training	1	–	1	–	–	–	–
124	2157-01-9	–1	Training	–1	–	–1	–	–	–	–
125	25870-67-1	1	Training	1	–	1	–	–	–	–
126	557-48-2	–1	Training	–1	–	–1	–	–	–	–
127	3160-37-0	–1	Training	1	–	–1	–	–	–	–
128	97461-42-2	1	Training	1	–	1	–	–	–	–
129	140-10-3	–1	Training	–1	–	–1	1	Training	1	–
130	91642-47-6	1	Training	1	–	1	–	–	–	–
131	1565-94-2	–1	Training	–1	–	–1	–	–	–	–
132	6923-22-4	1	Training	1	–	1	–	–	–	–
133	147151-67-5	1	Training	1	–	1	–	–	–	–
134	28564-83-2	1	Training	1	–	1	–	–	–	–
135	142438-64-0	1	Training	1	–	–1	–	–	–	–
136	63-75-2	1	Training	1	–	–1	–	–	–	–
137	55557-02-3	1	Training	1	–	1	–	–	–	–
138	2210-28-8	–1	Training	–1	–	–1	–	–	–	–
139	97461-43-3	1	Training	1	–	1	–	–	–	–
140	125973-99-1	1	Training	1	–	1	–	–	–	–
141	3179-47-3	–1	Training	–1	–	–1	–	–	–	–
142	78-94-4	1	Training	1	–	1	–	–	–	–
143	101-39-3	–1	Training	–1	–	–1	–	–	–	–
144	90-65-3	–1	Training	–1	–	1	–	–	–	–
145	13048-33-4	–1	Training	–1	–	–1	–	–	–	–
146	36840-85-4	–1	Training	–1	–	–1	–	–	–	–

Table 1 (Continued)

Compound no.	CAS	Observed		TOPS-MODE		TOXTREE	Observed			Predicted	
		AMES Class.	Partition	AMES Class.	Leverage		AMES Class.	MCGM Class.	Partition	MCGM Class.	Leverage
147	110-26-9	1	Training	-1	-	1	-	-	-	-	-
148	14308-65-7	1	Training	1	-	1	-	-	-	-	-
149	14901-07-6	-1	Training	-1	-	1	-	-	-	-	-
150	97461-38-6	1	Training	1	-	1	-	-	-	-	-
151	53175-28-3	1	Training	1	-	1	-	-	-	-	-
152	3787-28-8	1	Training	1	-	1	-	-	-	-	-
153	103-11-7	-1	Training	-1	-	-1	1	Training	1	-	-
154	18031-40-8	-1	Training	-1	-	1	-	-	-	-	-
155	78-59-1	-1	Training	-1	-	1	-1	Test	-1	0.046	-
156	959-33-1	-1	Training	-1	-	-1	-	-	-	-	-
157	1193-54-0	1	Training	1	-	1	-	-	-	-	-
158	458-37-7	-1	Training	-1	-	-1	-	-	-	-	-
159	1615-02-7	-1	Training	-1	-	-1	-	-	-	-	-
160	614-48-2	-1	Training	1	-	1	-	-	-	-	-
161	555-66-8	1	Training	-1	-	1	-	-	-	-	-
162	6755-16-4	1	Training	1	-	1	-	-	-	-	-
163	122-69-0	1	Training	Outlier	-	-1	-	-	-	-	-
164	1663-39-4	-1	Training	-1	-	-1	-	-	-	-	-
165	15743-13-2	-1	Training	-1	-	1	-	-	-	-	-
166	97461-39-7	1	Training	1	-	1	-	-	-	-	-
167	39965-42-9	1	Training	1	-	1	-	-	-	-	-
168	31876-38-7	-1	Training	-1	-	-1	-	-	-	-	-
169	19337-19-0	-1	Training	-1	-	1	-	-	-	-	-
170	122551-89-7	1	Training	1	-	1	-	-	-	-	-
171	541-59-3	1	Training	-1	-	1	1	Test	1	0.142	-
172	2177-18-6	-1	Training	-1	-	1	-	-	-	-	-
173	137-66-6	-1	Training	-1	-	-1	-	-	-	-	-
174	89-65-6	-1	Training	-1	-	-1	-	-	-	-	-
175	80-71-7	-1	Training	1	-	1	-	-	-	-	-
176	97-88-1	-1	Training	-1	-	-1	-	-	-	-	-
177	23255-69-8	-1	Test	-1	0.183 ^a	1	-	-	-	-	-
178	480-81-9	1	Test	-1	0.068	-1	-	-	-	-	-
179	2439-35-2	-1	Test	-1	0.012	-1	-	-	-	-	-
180	17341-40-1	-1	Test	1	0.050	1	-	-	-	-	-
181	1070-70-8	-1	Test	-1	0.018	-1	-	-	-	-	-
182	37962-27-9	1	Test	1	0.103	1	-	-	-	-	-
183	2154-67-8	-1	Test	-1	0.065	-1	-	-	-	-	-
184	514-78-3	-1	Test	-1	0.127 ^a	-1	-	-	-	-	-
185	584-79-2	1	Test	-1	0.048	1	-	-	-	-	-
186	1135-24-6	-1	Test	-1	0.018	-1	-	-	-	-	-
187	1608-51-1	-1	Test	-1	0.016	-1	-	-	-	-	-
188	497-23-4	-1	Test	1	0.016	-1	-	-	-	-	-
189	7473-93-0	-1	Test	1	0.027	1	-	-	-	-	-
190	1609-93-4	1	Test	1	0.015	-1	-	-	-	-	-
191	87-56-9	1	Test	1	0.032	1	1	Training	1	-	-
192	1734-79-8	1	Test	1	0.025	1	1	Training	-1	-	-
193	331-39-5	-1	Test	-1	0.021	-1	1	Training	1	-	-
194	6203-18-5	-1	Test	-1	0.025	1	-	-	-	-	-
195	129401-88-3	-1	Test	-1	0.013	1	-	-	-	-	-
196	3066-70-4	1	Test	1	0.021	1	-	-	-	-	-
197	585-07-9	-1	Test	-1	0.044	-1	-	-	-	-	-
198	1874-12-0	1	Test	1	0.023	1	-	-	-	-	-
199	1030-27-9	-1	Test	-1	0.023	1	-	-	-	-	-
200	818-61-1	-1	Test	-1	0.014	-1	1	Training	1	-	-
201	766-40-5	1	Test	1	0.020	-1	-	-	-	-	-
202	91182-09-1	1	Test	1	0.017	1	-	-	-	-	-
203	105-76-0	-1	Test	-1	0.023	-1	-	-	-	-	-
204	623-30-3	-1	Test	-1	0.017	-1	-	-	-	-	-
205	142-83-6	1	Test	1	0.021	1	1	Training	-1	-	-
206	106-63-8	-1	Test	-1	0.011	-1	-	-	-	-	-
207	90147-31-2	1	Test	1	0.018	1	-	-	-	-	-
208	1951-56-0	1	Test	1	0.020	1	1	Test	1	0.177	-
209	2648-51-3	1	Test	1	0.035	1	-	-	-	-	-
210	3290-92-4	1	Test	-1	0.068	-1	1	Training	1	-	-
211	1874-22-2	1	Test	1	0.031	1	-	-	-	-	-
212	108-31-6	-1	Test	-1	0.032	-1	-	-	-	-	-
213	77439-76-0	1	Test	1	0.029	1	1	Training	1	-	-
214	4655-34-9	-1	Test	-1	0.011	-1	-	-	-	-	-
215	4074-88-8	-1	Test	-1	0.025	-1	-	-	-	-	-
216	1874-24-4	1	Test	1	0.022	1	-	-	-	-	-
217	96910-71-3	1	Test	1	0.181 ^a	1	-1	Training	-1	-	-
218	125974-01-8	1	Test	1	0.025	1	-	-	-	-	-
219	922-63-4	1	Test	1	0.019	1	-	-	-	-	-
220	327-97-9	-1	Test	-1	0.092	-1	-1	Training	1	-	-

Table 1 (Continued)

Compound no.	CAS	Observed		TOPS-MODE		TOXTREE	Observed		Predicted	
		AMES Class.	Partition	AMES Class.	Leverage		AMES Class.	MCGM Class.	Partition	MCGM Class.
221	4170-30-3	–	–	–	–	–	–1	Training	1	–
222	303-34-4	–	–	–	–	–	1	Training	1	–
223	14371-10-9	–	–	–	–	–	1	Training	1	–
224	110-44-1	–	–	–	–	–	–1	Training	–1	–
225	108893-54-5	–	–	–	–	–	–1	Training	–1	–
226	18409-46-6	–	–	–	–	–	1	Training	1	–
227	33118-34-2	–	–	–	–	–	–1	Training	–1	–
228	3588-17-8	–	–	–	–	–	–1	Training	1	–
229	5956-39-8	–	–	–	–	–	–1	Training	–1	–
230	62966-21-6	–	–	–	–	–	1	Training	1	–
231	120-57-0	–	–	–	–	–	1	Test	1	0.178
231	140-88-5	–	–	–	–	–	1	Test	1	0.055
233	79-06-1	–	–	–	–	–	1	Test	1	0.079
234	37841-91-1	–	–	–	–	–	1	Test	–1	0.142
235	50656-61-6	–	–	–	–	–	–1	Test	–1	0.089

^a Chemicals with leverage values above the threshold (0.120 for AMES mutagenicity and 0.307 for MCGM) and, for that reason, its predictions were not taken into account.

being modelled. In this work, the weights included the standard bond distance (Std), standard bond dipole moments (Dip, Dip₂), as well as contributions from the following atomic properties: hydrophobicity (Hyd), polar surface area (Pols), polarizability (Pol), molar refractivity (Mol), van der Waals radii (vdW), Gasteiger–Marsilli charges (Gas), atomic masses (Ato), solute excess molar refraction (Ab-R₂), solute dipolarity/polarizability (Ab- π_2^H), effective hydrogen-bond basicity (Ab- $\sum \beta_2^O$, Ab- $\sum \beta_2^H$) and solute gas hexadecane partition coefficient (Ab-log L¹⁶). As described previously (Estrada et al., 2003b), the atomic contributions are transformed into bond weight contributions – $w(i, j)$ – as follows:

$$w(i, j) = \frac{w_i}{\delta_i} + \frac{w_j}{\delta_j} \quad (1)$$

where w_i and δ_i stand for the atomic weight and vertex degree of the atoms i and j , respectively. Finally, the spectral moments are defined as the traces (i.e., the sum of the main diagonal elements) of the different powers of the weighted B matrix.

In this work, these graph-based descriptors were computed with the MODESLAB software (<http://www.modeslab.com>) (Gutierrez and Estrada, 2002), using the SMILES (Simplified Molecular Input Line Entry System) notation available for each compound (Weininger, 1988).

Explicitly, we have calculated the first 15 spectral moments (μ_1 – μ_{15}) for each bond weight and the number of bonds in the molecules (μ_0), excluding the hydrogen atoms. We have also multiplied μ_0 and μ_1 for the first 15 spectral moments, obtaining therefore 30 new variables. Be aware that, these variables might offset the linear approximation assumption of the model. As to the modelling technique, we opted for building discriminant functions able of classifying the chemicals as actives or inactives. This was attained by the Linear Discriminant Analysis (LDA) technique implemented in STATISTICA software 8.0 (Frank, 2002).

To summarize, the following three-stage procedure was adopted to develop the structure–activity relationships.

Stage 1: Model selection. This proceeds as follows.

1. Select a small subset from the total chemicals to act as “test” set (for AMES: 44 from the total 220 chemicals; for MCGM: 9 from the total 48 chemicals; see Pérez-Garrido et al., in press). The remaining chemicals form the “training” set for QSAR modelling.
2. Draw the molecular graphs for each molecule included in the training and test sets.
3. Compute the spectral moment's descriptors using an appropriate set of weights.
4. Find an adequate QSAR model from the training set by a discriminant-based approach. The task here is to obtain a mathematical function (see Eq. (2) below) that best describes the studied activity P (in our case, the mutagenicity) as a linear combination of the X -predictor variables (the k -spectral moments μ_k), with the coefficients a_k . Such coefficients are to be optimized by means of LDA.
5. The QSAR model is subjected to rigorous internal and external validation, thereby assessing the performance of the model in what concerns its applicability and predictive power.
6. Compute the contribution of the different substructures to determine their quantitative contribution to the mutagenicity of the studied molecules.

$$P = a_0\mu_0 + a_1\mu_1 + a_2\mu_2 + \dots + a_k\mu_k + b \quad (2)$$

P values of +1 and –1 were assigned to active and inactive compounds, respectively. Moreover, several models were first obtained by forward stepwise LDA and then, the best of them was improved by the replacement method (RM) (Duchowicz et al., 2006).

Stage 2: Model validation. Assess the performance of the derived QSAR model by rigorous internal and external validation, looking in particular to its applicability and predictive power.

Stage 3: SAs identification. Identify one or more critical SAs for mutagenesis. That is to say, compute the contribution of different selected substructures and determine their quantitative contribution to the mutagenicity of the studied chemicals.

2.3. Model validation

Two kinds of diagnostic statistical tools were used for evaluating the performance of our discriminant model: the so-called goodness of fit and goodness of the prediction. In the first case, attention is given to the fitting properties of the model, whereas in the second case attention is paid to the predictive power of the model (i.e., the model adequacy for describing new compounds). In this work, k-Means Cluster Analysis (k-MCA) was used to split the original dataset of chemicals into training and an external validation test set. Full details of this partition can be found in our previous work (Pérez-Garrido et al., in press).

Measures of goodness of fit have been estimated by computing standard statistics such as the Mahalanobis distance (D^2), the Wilks' lambda (λ), the Fisher's statistic (F), and the corresponding p -level (p), as well as the percentage of correct classifications (accuracy). One should mention in particular that, the Mahalanobis distance shows whether the model has an adequate discriminatory power for differentiating between the two respective groups – active and inactive chemicals – whereas Wilks' lambda takes values in the range of zero (perfect discrimination) to one (no discrimination at all). Also, it should be remarked here that we minimized precisely the statistic when using the RM, not the standard deviation as it is done in linear regression analysis. Furthermore, similarly to the FIT statistic used in regression analysis (Kubinyi, 1994a,b), which allows comparing models with different number of variables (p) and cases (n), we have employed a new statistical parameter, FIT(λ), defined by:

$$FIT(\lambda) = \frac{(1 - \lambda)(n - p - 1)}{(n + p^2)\lambda} \quad (3)$$

Goodness-of-prediction for both the training and test sets was assessed by the following statistical measures:

- Accuracy: the percentage of chemicals correctly classified.
- Sensitivity: the percentage of toxicologically active chemicals (positives) correctly predicted as positives (calculated out of the total number of positives).
- Specificity: the percentage of toxicologically inactive chemicals (negatives) correctly predicted as negatives (calculated out of the total number of negatives).
- Kappa (K) (Cohen, 1960): The kappa index excludes matching due solely to chance. The maximum possible agreement is $K = 1$. $K = 0$ is obtained when the agreement observed is that expected exclusively by chance. If the agreement is higher than

Table 2
Interpretation of kappa

Kappa	Agreement
< 0	Less than chance agreement
0.01–0.20	Slight agreement
0.21–0.40	Fair agreement
0.41–0.60	Moderate agreement
0.61–0.80	Substantial agreement
0.81–0.99	Almost perfect agreement

Table 3

Results of the classification (%) of compounds in the training and external test sets, according to the TOPS-MODE models obtained here.

	AMES test (Eq. (4))	AMES test Cross val.	MCGM (Eq. (5))	MCGM Cross val.	AMES test (Eq. (6))	AMES test Cross val.
Training						
Sensitivity	85.54	85.47	92.59	93.35	86.59	86.34
Specificity	90.32	90.57	75.00	75.17	91.40	90.95
Accuracy	88.07	88.16	87.18	87.67	89.14	88.79
Test						
Sensitivity	85.00	85.71	85.71	89.87	85.00	86.13
Specificity	80.95	88.00	100.00	69.67	85.71	88.46
Accuracy	82.93	86.92	88.89	84.10	85.37	87.36

expected simply because of chance, $K > 0$, while if it is less, $K < 0$. However, a commonly cited scale is represented in Table 2 (Landis and Koch, 1977).

In addition, we carried out a cross-validation procedure on the training set. Specifically, the leave-group-out (LGO) procedure was applied, leaving out 20% of the training set by random extraction and then recalculating the model and the statistics with the remaining chemicals. This LGO procedure was repeated 300 times. The mean values of the accuracy, sensitivity, and specificity for both training and test sets, as well as the mean values of Wilk's λ (λ_{Cross}) and squared Mahalanobis distances (D_{Cross}^2), are reported.

In summary, good overall quality of the models is indicated by small values of λ , λ_{Cross} along with high values of FIT(λ), D^2 , F and Kappa.

The spectral moments are inherently collinear. From the point of view of QSAR modelling, the main drawback of collinearity is that it increases the standard errors associated with the individual coefficients, thereby decreasing their value for purposes of interpretability. To overcome this problem, we have employed here the Randić's method of orthogonalisation (Lucic et al., 1995; Klein et al., 1997; Randić, 1991b,c,a). Firstly, one has to select the appropriate order of orthogonalisation, which, in this case, is the order of significance of the variables in the model. The first variable (ν_1) is taken as the first orthogonal descriptor $\Omega\nu_1$ and the second one is orthogonalised with respect to it by taking the residual of its correlation with $\Omega\nu_1$. The process is repeated until all variables are completely orthogonalised. For extracting of the information contained in the orthogonalised descriptors we followed the procedure reported by Estrada and Molina (2006).

2.4. Applicability domain of the models

The utility of a QSAR model is its ability to accurately predict activity for new substances, which requires a careful assessment of the true predictive ability of models. This includes the model validation but also the definition of the applicability domain of the model in the space of molecular descriptors used for deriving the model. There are several methods for assessing the applicability domain of QSAR/QSPR models (Eriksson et al., 2003; Netzeva et al., 2005) but the most common one encompasses determining the leverage values for each compound (Gramatica, 2007). A Williams plot, i.e. the plot of standardized residuals versus leverage values (h), can then be used for an immediate and simple graphical detection of both the response outliers and structurally influential chemicals in the model. In this plot, the applicability domain is established inside a squared area within $\pm x$ standard deviations and a leverage threshold h^* (h^* is generally fixed at $3p/n$, where n is the number of training compounds and p the number of model parameters, whereas $x = 2$ or 3), lying outside this area (vertical lines) the outliers and (horizontal lines) influential chemicals. For future predictions, only predicted mutagenicity for chemicals belonging to the chemical domain of the training set should be proposed and used (Vighi et al., 2001). So, calculations of validation set classifications were performed only for those substances that had a leverage value below the threshold h^* .

3. Results and discussion

3.1. QSAR models

Following the computational strategies outlined in the previous section, the best model obtained for each of the chosen endpoints we can see in the following equations and in Table 3. As seen, these model is good both statistical significance and goodness of fit and prediction.

$$\begin{aligned} \text{AMES} = & 1.758 + 1.691\mu_1^{\text{Dip}2} - 3.399 \times 10^{-3}\mu_6^{\text{Dip}2} \\ & - 1.564 \times 10^{-2}\mu_5^{\text{Hyd}} + 8.557\mu_1^{\text{Gas}} + 3.341 \times 10^{-2}\mu_5^{\text{Ab}-\pi^H} \\ & - 6.858 \times 10^{-12}\mu_0\mu_{15}^{\text{Pol}} + 1.338 \times 10^{-2}\mu_1\mu_3^{\text{Hyd}} \quad (4) \end{aligned}$$

$$N = 176(83 \text{ positives}, 93 \text{ negatives}); \quad \lambda = 0.482;$$

$$D^2 = 4.259; \lambda_{\text{Cross}} = 0.478; \quad D_{\text{Cross}}^2 = 4.349; \quad p < 10^{-5};$$

$$F = 25.770; \quad \text{FIT} = 0.766; \quad K = 0.659$$

$$\begin{aligned} \text{MCGM} = & 4.143 - 2.548\mu_2^{\text{Dip}} + 3.012\mu_2^{\text{Dip}2} \\ & - 1.54 \times 10^{-4}\mu_7^{\text{Pol}} + 5.271\mu_1^{\text{Gas}} \quad (5) \end{aligned}$$

$$N = 39(26 \text{ positives}, 13 \text{ negatives}); \quad \lambda = 0.412; \quad D^2 = 6.343;$$

$$\lambda_{\text{Cross}} = 0.399; \quad D_{\text{Cross}}^2 = 6.756; \quad p < 10^{-5}; \quad F = 12.107;$$

$$\text{FIT} = 0.881; \quad K = 0.727$$

Then, we search for the presence of outliers that might be distorting these models. The high value found for the standard residual (> 3) of chemical 163 (i.e. of cinnamyl cinnamate) suggests that it could be an outlier. In general the cinnamyl derivatives were not mutagenic in the Ames test and their metabolism *in vivo* is usually to hippuric acid (Belsito et al., 2007). Therefore the derived AMES model (Eq. (4)) is not able to predict the mutagenicity of this chemical since it does not follow the general pattern of cinnamyl derivatives, thus being an outlier. If we remove this chemical from the training set and further proceed to refitting, we obtained the AMES model shown below.

$$\begin{aligned} \text{AMES} = & 1.864 + 1.882\mu_1^{\text{Dip}2} - 3.757 \times 10^{-3}\mu_6^{\text{Dip}2} \\ & - 1.734 \times 10^{-2}\mu_5^{\text{Hyd}} + 9.489\mu_1^{\text{Gas}} + 3.705 \times 10^{-2}\mu_5^{\text{Ab}-\pi^H} \\ & - 7.534 \times 10^{-12}\mu_0\mu_{15}^{\text{Pol}} + 1.426 \times 10^{-2}\mu_1\mu_3^{\text{Hyd}} \quad (6) \end{aligned}$$

$$N = 175 (82 \text{ positives}, 93 \text{ negatives}); \quad \lambda = 0.451;$$

$$D^2 = 4.818; \lambda_{\text{Cross}} = 0.448; \quad D_{\text{Cross}}^2 = 4.913; \quad p < 10^{-5};$$

$$F = 28.958; \quad \text{FIT} = 0.869; \quad K = 0.707$$

Eliminating the outlier produces an appreciable improvement in the statistical parameters as well as the percentages of classification (see Eq. (6) and Table 3). Another aspect deserving special attention is the degree of collinearity of the variables of the model, which can readily be diagnosed by analyzing the cross-correlation

Table 4

Intercorrelation among the descriptors selected for the initial AMES model (Eq. (4)).

	$\mu_1^{\text{Dip}2}$	$\mu_6^{\text{Dip}2}$	μ_5^{Hyd}	μ_1^{Gas}	$\mu_5^{\text{Ab}-\pi^H}$	$\mu_0\mu_{15}^{\text{Pol}}$	$\mu_1\mu_3^{\text{Hyd}}$
$\mu_1^{\text{Dip}2}$	1.00	-	-	-	-	-	-
$\mu_6^{\text{Dip}2}$	0.82	1.00	-	-	-	-	-
μ_5^{Hyd}	0.49	0.86	1.00	-	-	-	-
μ_1^{Gas}	-0.57	-0.63	-0.53	1.00	-	-	-
$\mu_5^{\text{Ab}-\pi^H}$	0.56	0.91	0.97	-0.60	1.00	-	-
$\mu_0\mu_{15}^{\text{Pol}}$	0.50	0.84	0.87	-0.56	0.91	1.00	-
$\mu_1\mu_3^{\text{Hyd}}$	0.18	0.34	0.57	-0.36	0.42	0.44	1.00

Table 5

Intercorrelation among the descriptors selected for the initial MCGM model (Eq. (5)).

	μ_2^{Dip}	$\mu_2^{\text{Dip}2}$	μ_7^{Pol}	μ_1^{Gas}
μ_2^{Dip}	1.00	–	–	–
$\mu_2^{\text{Dip}2}$	0.99	1.00	–	–
μ_7^{Pol}	0.89	0.89	1.00	–
μ_1^{Gas}	–0.71	–0.72	–0.49	1.00

matrix. Tables 4 and 5 show that, for both models, there are several descriptor variables highly correlated with each other. Rather than deleting any of these descriptors, it is of interest to examine the performance of orthogonal complements.

Following Randić's technique, we have determined orthogonal complements for all variables of the above non-orthogonalised models (Eqs. (5) and (6)), the following QSAR models were obtained:

$$\begin{aligned} \text{AMES} = & 2.188 + 0.230\Omega^3\mu_1^{\text{Dip}2} - 3.827 \times 10^{-3}\Omega^4\mu_6^{\text{Dip}2} \\ & + 1.133 \times 10^{-2}\Omega^6\mu_1\mu_3^{\text{Hyd}} + 4.050\Omega\mu_1^{\text{Gas}} \\ & - 5.036 \times 10^{-3}\Omega^5\mu_5^{\text{Hyd}} + 3.007 \times 10^{-3}\Omega^2\mu_5^{\text{Ab}-\pi^H} \\ & - 7.534 \times 10^{-12}\Omega^7\mu_0\mu_{15}^{\text{Pol}} \end{aligned} \quad (7)$$

$$N = 175 (82 \text{ positives}, 93 \text{ negatives}); \quad \lambda = 0.451; \quad D^2 = 4.818;$$

$$\lambda_{\text{Cross}} = 0.448; \quad D_{\text{Cross}}^2 = 4.913; \quad p < 10^{-5};$$

$$F = 28.958; \text{FIT} = 0.869; \quad K = 0.707$$

$$\begin{aligned} \text{MCGM} = & 4.183 - 2.193\Omega^2\mu_2^{\text{Dip}} - 0.062\Omega\mu_2^{\text{Dip}2} \\ & - 1.088 \times 10^{-4}\Omega^3\mu_7^{\text{Pol}} + 5.271\Omega^4\mu_1^{\text{Gas}} \end{aligned} \quad (8)$$

$$N = 39 (26 \text{ positives}, 13 \text{ negatives}); \quad \lambda = 0.412;$$

$$D^2 = 6.343; \lambda_{\text{Cross}} = 0.406; \quad D_{\text{Cross}}^2 = 6.548; \quad p < 10^{-5};$$

$$F = 12.107; \quad \text{FIT} = 0.881; \quad K = 0.727$$

The percentages of classifications of the derived orthogonalised models were found to be the same as the non-orthogonalised models (results not shown).

3.2. Comparison between the TOPS-MODE QSAR model and the TOXTREE software for predictions of AMES mutagenicity

The TOXTREE software is capable of making structure-based predictions for a number of toxicological endpoints but one of its modules aims at predicting carcinogenicity and mutagenicity. Mutagenic predictions from this expert software system are based in a revised list of structural alerts, taken from several literature sources, and in one QSAR model for α , β -unsaturated aldehydes (Benigni and Bossa, 2008; Benigni et al., 2008). The present analysis concerned comparing the predictions of our TOPS-MODE model against those of TOXTREE with respect to the AMES mutagenicity for this family of compounds. Thus, a chemical is considered to be predicted as positive (i.e., potentially mutagenic) either if it contains one genotoxic (DNA reactive) structural alert or if it belongs to the applicability domain of the relating QSAR model, otherwise it was identified as negative. For an easy visual comparison, the results are expressed as a Receive Operating Characteristics (ROC) graph (Fig. 2). A ROC graph reports true positive rate (sensitivity) on the Y-axis, and false positive rate (1-specificity) on the X-axis. In a ROC graph, perfect performance is located at the left upper corner, and random results lying on the diagonal line (Provost and

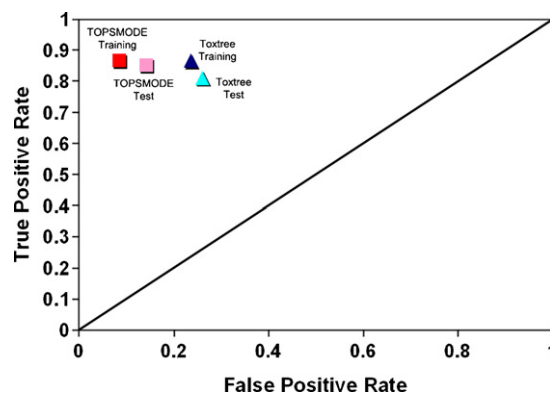


Fig. 2. Receiver operating characteristic graph for mutagenicity predictions of α , β -unsaturated carbonyl compounds given by TOXTREE and TOPS-MODE based on Ames data. The diagonal line corresponds to random responses whereas the top left corner to ideal performance.

Fawcett, 2001). Further details of the obtained results are collected in Table 6 (see also Table 1).

This analysis shows that the present QSAR model (Eq. (7)) has higher specificity and accuracy than the TOXTREE software for the training set, though identical sensitivity. However, for the external test set, the percentage of chemicals correctly identified by the TOPS-MODE model is superior to the one attained by TOXTREE. These results imply that the performance of the TOPS-MODE model is better than that of the TOXTREE software, not only in the percentage of overall classifications, but also and more important, in terms of the SA modulating factors for the mutagenicity of this chemical class. This is clearly due to the ability of TOPS-MODE approach in modelling the mutagenic activity at a local scale, which further allows quantifying how each alert is modulated by several molecular environments (modulating factors). More details about this issue will be given in next subsection.

3.3. Identification of structural alerts

One advantage of the present approach for QSAR studies is that it can provide information explaining how structural features of molecules can account for the endpoint activities (Estrada, 2008). It is then possible to detect fragments that contribute positively or negatively to a particular target endpoint and their effects been interpreted in terms of physicochemical properties.

Specifically in our case, the contributions to the mutagenicity for each of the selected fragments (see Fig. 3) were extracted from the final orthogonal-descriptor models related to the two endpoints. Table 7 shows the particular numerical values of the contributions of such fragments. A careful look at these values allows us to find functional groups that either hamper the toxicity or enhance it. Further, it might lead us to design molecular structures that are less

Table 6

Results of the classification (%) of compounds in the training and external test sets, according to the TOPS-MODE model (Eq. (7)) and the TOXTREE software.

	TOPS-MODE	TOXTREE
Training		
Sensitivity	86.59	86.59
Specificity	91.40	76.34
Accuracy	89.14	81.14
Test		
Sensitivity	85.00	80.95
Specificity	85.71	73.91
Accuracy	85.37	77.27

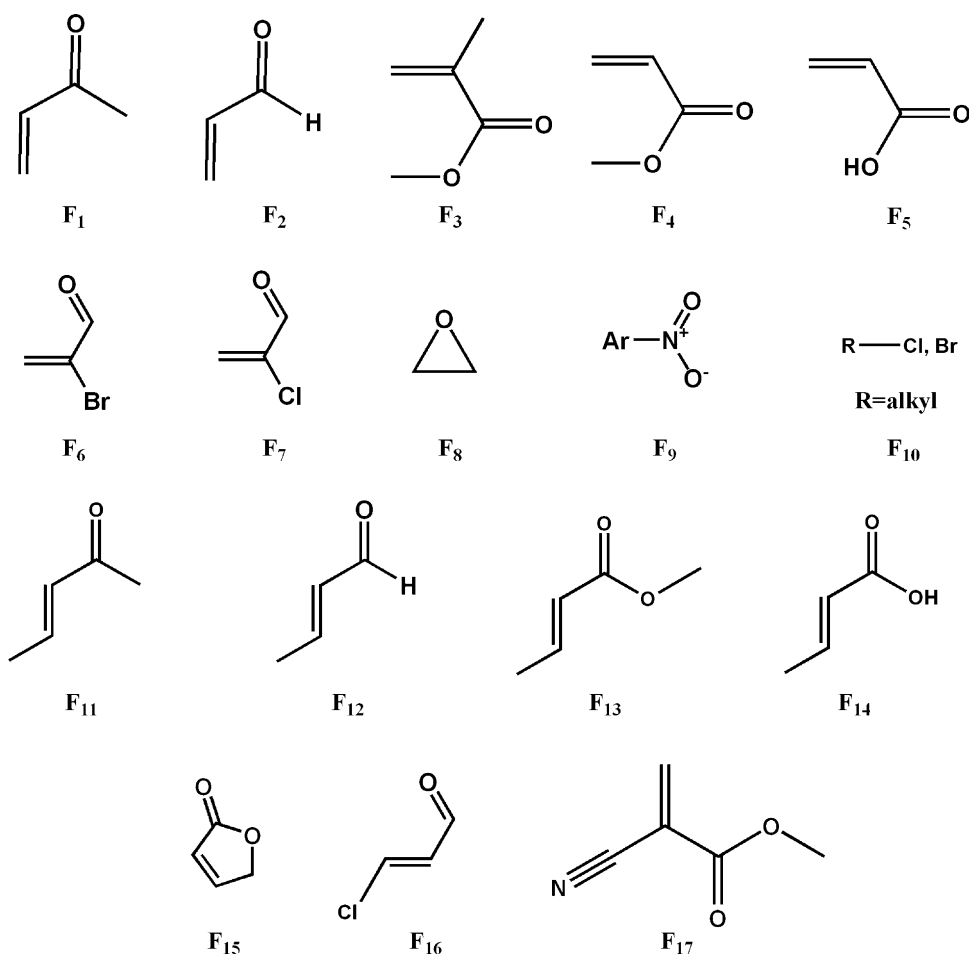


Fig. 3. Selected molecular fragments (substructures) for which their contributions to either the AMES or the MCGM mutagenicity were calculated according to the TOPS-MODE models obtained here (Eqs. (7) and eq:MLAorto).

toxic, to find new structural alerts or to a rapid screening among a long list of substances.

Regarding AMES mutagenicity, firstly, a comparison between fragments F_1 and F_2 shows that, the ketonic fragment (F_1) contributes less than the aldehyde fragment (F_2), the F_1 contribution being even negative. This is in clear agreement with the analysis performed by Koleva et al. (2008), where the authors concluded that

aldehydes are more reactive than ketones due to the size and electronic effects of their substituents. Besides, one can easily see in Fig. 1 that, the aldehyde group has a greater electron-withdrawing effect on the double bond than the ketonic group which increases its reactivity in the Michael addition mechanism.

Secondly, the presence of halogens in position α of the double bond adjacent to the carbonyl group (fragments F_6 , F_7 and F_{16}) increase the mutagenicity of this family of compounds (Eder and Weinfurtner, 1994; Eder et al., 1990), due to the cross-linking potential with another DNA or protein nucleophilic centre (Van Beerendonk et al., 1992).

On the other hand, fragments F_8 to F_{10} relate to well recognized structural alerts for mutagenic AMES data, i.e.: epoxides (F_8), alkyl halides (F_{10}) and nitro aromatics (F_9). The first two functional groups are known alkylating agents while the latter one is mainly activated by means of nitroreduction and oxidative pathways involving several enzymes in different organisms (Purohit and Basu, 2000) to the N hydroxyl species, which are then transformed into reactive nitrogen esters or nitrenium ions and that in turn, may attack DNA forming adducts (Miller and Miller, 1983; Sasaki et al., 2002). In what concerns the positive contribution of the cyano acrylate group (fragment F_{17}), possibly it is due to an electron-withdrawing effect of the cyano moiety which increases the Michael addition reactivity of the double bond (Aptula and Roberts, 2006), as acrylates have a negative contribution (fragment F_4).

So, without doubt, the main mechanism of action for this family of compounds is the Michael type addition mechanism since our

Table 7

Contributions of the different structural fragments to the AMES and MCGM mutagenicity according to the TOPS-MODE models obtained here.

Fragment	Ames contribution	MCGM contribution
F_1	-0.112	0.196
F_2	1.673	2.075
F_3	-0.824	1.465
F_4	-0.775	2.981
F_5	-1.494	2.769
F_6	2.225	-
F_7	2.234	2.784
F_8	0.091	-
F_9	3.332	-
F_{10}	1.364	-
F_{11}	-0.143	-1.932
F_{12}	1.470	-0.060
F_{13}	-0.464	2.119
F_{14}	-1.449	0.648
F_{15}	1.710	3.866
F_{16}	2.126	-
F_{17}	0.371	-

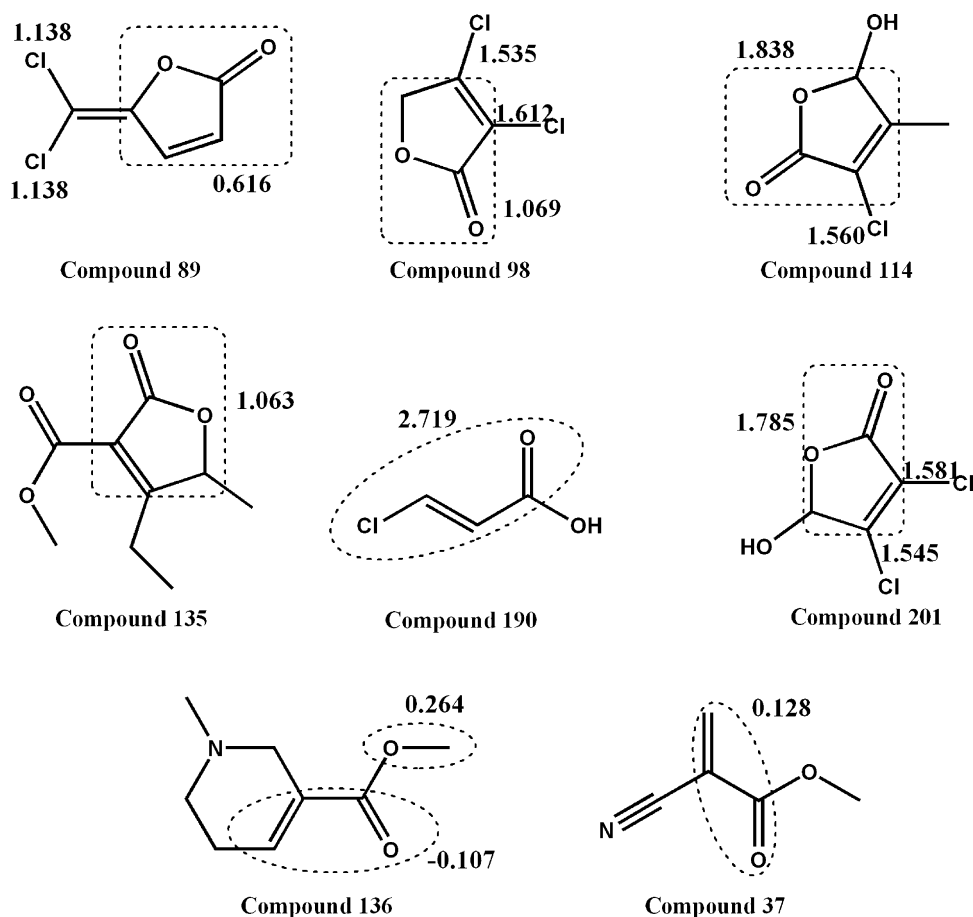


Fig. 4. Structure representation of some TOXTREE false negative compounds correctly predicted by the TOPS-MODE model together with their bond contributions (Eq. (7)).

analysis revealed that substituents in the α or β -carbon atoms have a strong influence in mutagenicity just as for Michael acceptors. Although it is known that the Michael acceptors are soft electrophiles, it does not mean that they are unreactive toward hard nucleophiles like DNA (Aptula and Roberts, 2006).

Another important feature that we can draw from both of our models (Eqs. (7) and (8)) is the mutagenicity of the furan-2(5H)-one ring (fragment F₁₅). This moiety together with halogenated α , β -unsaturated carbonyl compounds (fragments F₆, F₇ and F₁₆) are present in known mutagenic substances such as 3-chloro-4-(dichloromethyl) 5-hydroxy 2(5H) furanone (compound **213**) and 3-chloro-4(chloromethyl)-5-hydroxy-2(5H)-furanone (compound **82**) (McDonald and Komulainen, 2005), including in others not as well known, like compounds **89**, **98**, **114**, **135**, **190** and **201**. Here it should be emphasised that, the TOXTREE software does not recognize any structural alerts in the latter, classifying thus them as

false negatives. Fig. 4 displays some TOXTREE false negatives for the Ames test mutagenicity, which were correctly predicted by our model together with the computed TOPS MODE fragment contributions.

A close inspection of Fig. 4, shows that fragments F₇, F₁₅ and F₁₆ have positive contributions to the mutagenic activity and so, this may be due to the presence of the furan 2(5H)-one ring or to chlorine at the double bond adjacent to the carbonyl group, as it is known that the presence of halogens or halogenated alkyl groups at the furan double bond increases mutagenic potency (Lalonde et al., 1991; McDonald and Komulainen, 2005). In relation to compound **89**, the presence of allylic chlorines contributing positively also has a role. Notice however that our TOPS MODE model predicts a false positive, namely butenolide (compound **188**), which also contains fragment F₁₅. Thus, the modulating factors for this substructure have to be studied further.

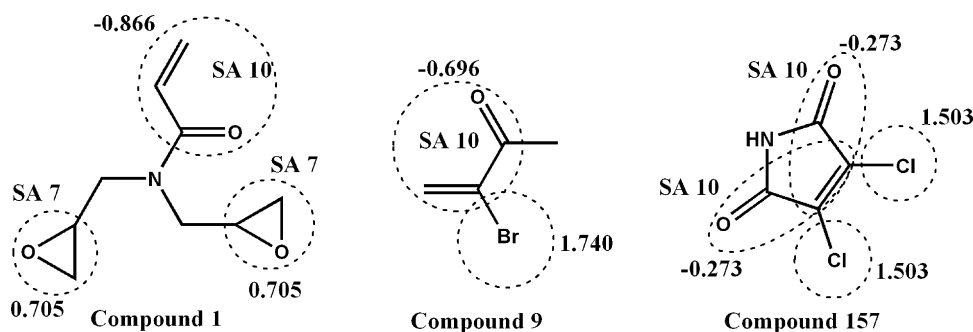


Fig. 5. Structure representation of some compounds of the AMES training set along with their TOPS-MODE bond contributions.

Another type of examples of TOXTREE false negatives, correctly predicted by our model, should be pointed out. For instance, our model recognises that the presence of a methyl group in compound **136** (see Fig. 4) yields a positive contribution (0.264) to its mutagenicity. For this compound, the mechanism of mutagenic action in bacteria is probably the same as the one in rats where it has been shown that the loss of the methyl group (Boyland and Nery, 1969) may bind with nucleic acid and protein (Nery, 1971). With regard to compound **37** (Fig. 4), its mutagenic mode of action is still not known (Richard, 2001), but the presence of fragment F₁₇ (referred to above) can be responsible for the mutagenicity of its α , β -unsaturated carbonyl moiety.

Fig. 5 depicts three mutagenic substances for which the TOXTREE software identifies structural alerts, but also the TOPS-MODE is able of discriminating their differences in terms of fragment contributions.

For instance, the values of the fragment contributions in compound **1** show that the presence of α , β -unsaturated carbonyl moiety is not responsible for its mutagenicity, but instead the presence of the epoxide because similar substances without oxirane moiety such as N,N-diethylacrylamide and N,N-dimethylacrylamide are not mutagenic to Ames test (Hashimoto and Tani, 1985).

As to compounds **9** and **157**, the structural alert detected by TOXTREE corresponds to the α , β -unsaturated carbonyl moiety while our model detected, as shown in Fig. 5, a negative contribution for this substructure, and relates their activity possibly due to the presence of allyls halogens corresponding to fragments F₆, F₇ and F₁₆, which could act as cross-linking agents (Van Beerendonk et al., 1992; Lynch and Crovetti, 1972; Smith, 1987).

Moreover, TOPS-MODE classifies correctly the majority of false positives obtained by the TOXTREE software, further detecting negative bond contributions for the structural alerts that allow, in a quantitative way, to properly interpret the possible cause of their non-mutagenicity (Table 8).

As for such compounds, there is no mechanism that explains their non-mutagenicity in *Salmonella typhimurium* strains with or without metabolic activation; we will try to set up hypothesis based on the contributions' results and on the bio-transformations produced by other organisms.

For example, in compound **2**, despite from having several hydroxyl groups which makes it more hydrophilic and therefore less mutagenic, the epoxide group displays no mutagenicity most likely due to a metabolic reduction of this group as seen in gastrointestinal microbes (Hedman and Pettersson, 1997).

The non-mutagenicity of compounds **3** and **5** are probably due to the large size of these molecules, but nevertheless TOPS-MODE identifies negative bond contributions for the TOXTREE structural alerts.

As to compound **42**, it is known that it is metabolized in humans mainly by several mechanisms, i.e.: reduction and subsequent oxidation of the hydroxymethylene group, a hydroxylation in position 6 and a carbonyl reduction (Fragkaki et al., 2009). These changes are consistent with the negative contributions obtained, and based on this a similar hydroxylation in position 11 could thus happen.

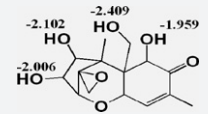
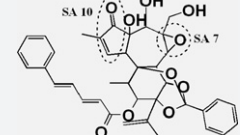
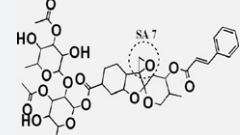
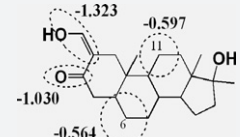
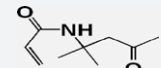
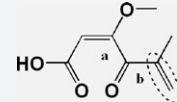
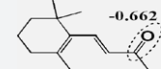
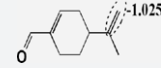
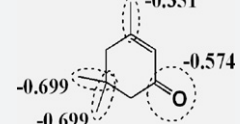
The presence of an acryl amide group, as seen previously for N,N-diglycidylacrylamide (compound **1**), is not responsible for the mutagenicity in the Ames test, which can be further observed by the values obtained for the SA of compound **72**.

For the following compound (**144**), a detoxification mechanism in mice through its conjugation with glutathione in the double bond has been presented (Chan et al., 1982, 1984), which has a large negative contribution (see Table 8), and so it could act similarly in *Salmonella typhimurium*.

The lack of mutagenicity for compound **149**, as seen in rabbits (Lalko et al., 2007), may be due to a hydroxylation of the carbonyl group (very negative bond contribution).

Table 8

Structure representation of some TOXTREE false positive compounds correctly predicted by the TOPS-MODE model together with their bond contributions (Eq. (7)).

Compound no.	Chemical representation	SA10 ^a	SA7 ^a
2		-0.969	-0.34
3		-1.381	-0.019
5		-	-0.032
42		0.054	-
72		-0.609	-
144		(a) 0.298 (b) -1.682	-
149		-0.764	-
154		1.593	-
155		-0.375	-

^a SA10 and SA7 are codes of the structural alerts defined by Benigni and Bossa (2008) corresponded to α , β -unsaturated carbonyl and oxirane moieties, respectively.

Compound **154**, although it is an aldehyde, is not mutagenic in the Ames test. This compound has a detoxification metabolism in *Euglena gracilis* Z (Noma et al., 1991) due to the oxidation of aldehyde and reduction of the double bond. Perhaps the latter is the one that predominates in *Salmonella typhimurium* because of the negative contribution that such double bond has in this compound. Compound **154** also has an aliphatic cycle around the double bond which affords an electron-donating effect that decreases its Michael addition reactivity (Aptula and Roberts, 2006).

The mechanism of detoxification in rats and rabbits of compound **155** is by methyl carboxylation or the reduction of the carbonyl group (Dutertre-Catella et al., 1978), and as can be seen both bonds have negative contributions, but maybe in *Salmonella typhimurium*, it is more important the metabolic pathway by a possible hydroxylation of methyl in position 5 as it has been observed too in *Aspergillus niger* (Joe et al., 1989).

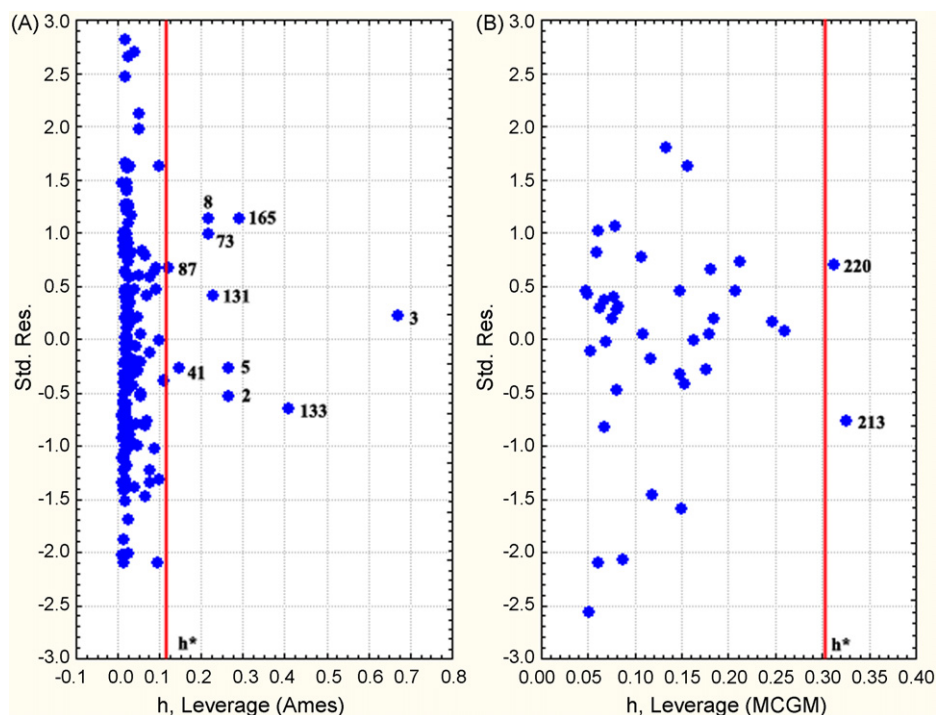


Fig. 6. Williams plot based on Eq. (7) (A) and (8) (B), i.e., plot of standardised residuals versus leverage values with a warning leverage of $h^* = 0.120$ and $h^* = 0.307$, respectively.

For the other endpoint studied, MCGM, there are no structural alerts identified in the literature. For this endpoint, the presence of the α , β -unsaturated carbonyl moiety produces mutagenicity either ketonic, aldehydic, acrylic or methacrylic (fragments F₁ to F₅ in Fig. 3 and Table 7). One can see also an increase in the contribution of fragment F₄ compared to fragment F₃. Electro-donating substituents such as methyl groups in position α reduce the reactivity of this moiety by Michael type mechanism (Aptula and Roberts, 2006). These findings, among other comparative analysis of the mutagenic potency of various acrylate and methyl acrylate derivatives, lead to the hypothesis that acrylate were more active mutagens than methylacrylates (Dearfield et al., 1989).

Moreover, when comparing fragments F₁, F₂, F₄ and F₅ to F₁₁, F₁₂, F₁₃ and F₁₄, respectively, a similar conclusion for this endpoint to that obtained by our group in a previous work (Pérez-Garrido et al., in press) is attained. That is to say, that alkyl substituents in position β at the double bond adjacent to the carbonyl group decrease the mutagenic character of the substance by reducing the positive charge at the terminal carbon (Aptula and Roberts, 2006), and the latter is the preferred site of nucleophilic attack (Feron et al., 1991; Dearfield et al., 1991) by a Michael type addition mechanism to the sulfhydryl of glutathione (GSH) or by an enzymatic reaction catalyzed by GSH transferase (Ciaccio et al., 1998; Schultz et al., 2005). But also, GSH when depleted down to < 20% (Glaab et al., 2001) is a prerequisite for α , β -unsaturated carbonyl-mediated generation of ROS (Radical Oxygen Species) and might initiate lipid peroxidation and other processes, leading to enhanced cytotoxic/genotoxic cell damage (Janowski et al., 2003). Hence, the presence of a terminal double bond without electron donating substituent makes these compounds more mutagenic. Based on this, we can say that, following the Michael addition mechanism, the presence of electron-withdrawing substituents in the double bond (i.e. fragment F₇) increase the mutagenicity of the substance (Aptula and Roberts, 2006; Schultz et al., 2005).

Compared with the results of the Ames test, the reactivity as Michael acceptors is more pronounced in MCGM, judging by the higher variation of its contribution values. As mentioned above,

this is most likely because Michael acceptors are soft electrophiles and as such reactivity is higher towards soft nucleophiles like GSH, then such appears to be the main mechanism step producing DNA damage in mammalian cells for these substances (Glaab et al., 2001; Janowski et al., 2003).

3.4. Applicability domain

It would be very interesting to have a predictive model for the vast majority of chemicals, especially for those who have not been tested yet and thus, with unknown mutagenicity, in particular taking into account that the European Union is launching the REACH standard. Since this is usually not possible, one should define the applicability domain of the QSAR model, that is, the range within which it bears a new compound. For that purpose, we built a Williams plot using the leverage values calculated for each compound. As seen in Fig. 6, most of the compounds of the test set are within the applicability domain covered by ± 3 times the standard residual (σ) and the leverage threshold h^* ($= 0.120$ and $= 0.307$ for AMES and MCGM, respectively), save for compounds 2, 3, 5, 8, 42, 73, 87, 131, 133 and 165 (AMES) and for compounds 213 and 220 (MCGM). Even so, the latter should not be considered outliers but influential chemicals (Eriksson et al., 2003).

Nevertheless, all evaluations pertaining to the external set were performed by taking into account the applicability domain of our QSAR model. So, if a chemical belonging to the test set had a leverage value greater than h^* , we consider that this means that the prediction is the result of substantial extrapolation and therefore may be unreliable (Netzeva et al., 2005).

4. Conclusions

Herein, we have examined the ability of the TOPS-MODE approach to provide discriminant models for probing the mutagenicity of the α , β -unsaturated carbonyl compounds over two endpoints: the Ames and mammalian cell mutation gene tests.

With regard to the QSAR modelling, the combination of LDA in conjunction with the TOPS-MODE structure representation was found to produce final classification models with high sensitivity, specificity and accuracy. The predictive power of such QSAR models was proved to even exceed that of state-of-art expert systems, such as the TOXTREE software, for this family of compounds. Furthermore, due to the ability of TOPS-MODE to express the activity at a local level, we could obtain a series of structural alerts for each compound and both endpoints under study. Among such alerts, the halogenated α , β -unsaturated carbonyls and the 2-furanone ring should be studied further in what concerns their Ames mutagenicity. As regards the mammalian cell gene mutation end-point, the presence of a terminal double bond with electron-withdrawing or without electron-donating substituents turns the compounds more mutagenic probably because they act through a Michael type addition mechanism. For both endpoints, we note that the predominant mechanism is Michael type addition by forming adducts either with DNA or with GSH. Moreover, by carefully analysing the fragment contributions obtained with the TOPS MODE approach, we were able to propose possible mutagenic mechanisms for a number of false negatives and false positives compounds settled on by the expert system TOXTREE in relation to the Ames data. In addition, the TOPS-MODE approach was able to quantify the influence of several molecular environments (modulating factors) to the structural alerts that describe the mutagenicity of the α , β -unsaturated carbonyl moiety. Overall, that structural information and the QSAR models per se can definitely aid in future improvements of software or experts systems based on SAs.

Conflict of interest

None.

Acknowledgements

The authors acknowledge to MODESLAB 1.0 software owners for delivering a free copy of such program. AMH acknowledges the Portuguese Fundação para a Ciência e a Tecnologia (FCT - Lisboa) (SFRH/BD/22692/2005) for financial support. APG acknowledges to Dr. Antonio Pérez-Garrido for his valuable collaboration in the completion of this work.

Appendix A. Supplementary Data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.tox.2009.11.023.

References

- Aptula, A.O., Roberts, D.W., 2006. Mechanistic applicability domains for non-animal-based toxicological end points: general principles and application to reactive toxicity. *Chem. Res. Toxicol.* 19, 1097–1105.
- Belsito, D., Bickers, D., Bruze, M., Calow, P., Greim, H., Hanifin, J.M., Rogers, A.E., Saurat, J.H., Sipes, I.G., Tagami, H., 2007. A toxicologic and dermatologic assessment of related esters and alcohols of cinnamic acid and cinnamyl alcohol when used as fragrance ingredients. *Food Chem. Toxicol.* 45, S1–S23.
- Benigni, R., Bossa, C., 2008. Structure alerts for carcinogenicity, and the Salmonella assay system: a novel insight through the chemical relational databases technology. *Mutat. Res.* 659, 248–261.
- Benigni, R., Bossa, C., Jeliakova, N.G., Netzeva, T.I., Worth, A.P. (Eds.), 2008. The Benigni/Bossa rulebase for mutagenicity and carcinogenicity—a module of Toxtree. EUR 23241 EN, EUR-Scientific and Technical Report Series, Office for the Official Publications of the European Communities, Luxembourg.
- Benigni, R., Conti, L., Crebelli, R., Rodomonte, A., Vari, M.R., 2005. Simple and α , β -unsaturated aldehydes: Correct prediction of genotoxicity activity through structure–activity relationship models. *Environ. Mol. Mutagen.* 46, 268–280.
- Benigni, R., Passerini, L., Gallo, G., Giorgi, F., Cotta-Ramusino, M., 1998. QSAR models for discriminating between mutagenic and nonmutagenic aromatic and heteroaromatic amines. *Environ. Mol. Mutagen.* 32, 75–83.
- Boelens, M.H., Gemert, L.J., 1987. Organoleptic properties of aliphatic aldehydes. *Perfumer Flavorist* 12, 31–43.
- Boylard, E., Nery, R., 1969. Mercapturic acid formation during metabolism of arecoline and arecaidine in the rat. *Biochem. J.* 113, 123–130.
- Chan, P.K., Hayes, W.A., Siraj, M.Y., 1982. Excretion of conjugated metabolites of the mycotoxin penicillic acid in male mice. *Toxicol. Appl. Pharm.* 66, 259–326.
- Chan, P.K., Hayes, W.A., Siraj, M.Y., Meydrech, E.F., 1984. Pharmacokinetics of the mycotoxin penicillic acid in male mice: Absorption, distribution, excretion, and kinetics. *Toxicol. Appl. Pharm.* 73, 195–203.
- Ciaccio, P.J., Gicquel, E., O'Neill, P.J., Scribner, H.E., Vandenberghe, Y.L., 1998. Investigation of the positive response of ethyl acrylate in the mouse lymphoma genotoxicity assay. *Toxicol. Sci.* 46, 324–332.
- Cohen, J., 1960. A coefficient of agreement for nominal scales. *J. Educ. Psychol. Meas.* 20, 37–46.
- Dearfield, K.L., Harrington-Brock, K., Doerr, C.L., Rabinowitz, J.R., Moore, M.M., 1991. Genotoxicity in mouse lymphoma cells of chemicals capable of Michael addition. *Mutagenesis* 6, 519–525.
- Dearfield, K.L., Millis, C.S., Harrington-Brock, K., Doerr, C.L., Moore, M.M., 1989. Analysis of genotoxicity of nine acrylate/methacrylate compounds in 15178y mouse lymphoma cells. *Mutagenesis* 4, 381–393.
- Duchowicz, P.R., Castro, E.A., Fernndez, F.M., 2006. Alternative algorithm for the search of an optimal set of descriptors in QSAR-QSPR studies. *MATCH Commun. Math. Comput. Chem.* 55, 179–192.
- Duterte-Catella, H., Nguyen, P.L., Dang Quoc, Q., Truhaut, R., 1978. Transformations métaboliques de la triméthyl-3,5,5-cyclohexène-2, one-1 (isophorone). *Toxicol. Eur. Res.* 1, 209–216.
- Eder, E., Hoffman, C., Bastian, H., Deininger, C., Scheckenbach, S., 1990. Molecular mechanisms of DNA damage initiated by α , β -unsaturated carbonyl compounds as criteria for genotoxicity and mutagenicity. *Environ. Health Perspect.* 88, 99–106.
- Eder, E., Weinfurter, E., 1994. Mutagenic and carcinogenic risk of oxygen containing chlorinated c-3 hydrocarbons: Putative secondary products of c-3 chlorohydrocarbons and chlorination of water. *Chemosphere* 29, 2455–2466.
- Eriksson, L., Jaworska, J., Worth, A.P., Cronin, M.T.D., McDowell, R.M., Gramatica, P., 2003. Methods for reliability and uncertainty assessment and for applicability evaluations of classification- and regression-based QSARs. *Environ. Health Perspect.* 111, 1361–1375.
- Estrada, E., 1995. Edge adjacency relationships and a novel topological index related to molecular volume. *J. Chem. Inf. Comput. Sci.* 35, 31–33.
- Estrada, E., 1996. Spectral moments of the edge adjacency matrix in molecular graphs. 1. definition and applications to the prediction of physical properties of alkanes. *J. Chem. Inf. Comput. Sci.* 36, 844–849.
- Estrada, E., 1997. Spectral moments of the edge-adjacency matrix of molecular graphs. 2. molecules containing heteroatoms and QSAR applications. *J. Chem. Inf. Comput. Sci.* 37, 320–328.
- Estrada, E., 2008. How the parts organize in the whole? A top-down view of molecular descriptors and properties for QSAR and drug design. *Mini-Rev. Med. Chem.* 8, 213–221.
- Estrada, E., Molina, E., 2006. Automatic extraction of structural alerts for predicting chromosome aberrations of organic compounds. *J. Mol. Graph. Model.* 25, 275–288.
- Estrada, E., Molina, E., Uriarte, E., 2001. Quantitative structure–toxicity relationships using tops-mode. 2. Neurotoxicity of a non-congeneric series of solvents. *SAR QSAR Environ. Res.* 12, 445–459.
- Estrada, E., Patlewicz, G., Gutierrez, Y., 2004. From knowledge generation to knowledge archive. A general strategy using TOPS-MODE with DEREK to formulate new alerts for skin sensitization. *J. Chem. Inf. Comput. Sci.* 44, 688–698.
- Estrada, E., Patlewicz, G., Chamberlain, M., Basketter, D., Larbey, S., 2003a. Computer-aided knowledge generation for understanding skin sensitization mechanisms: the tops-mode approach. *Chem. Res. Toxicol.* 16, 1226–1235.
- Estrada, E., Uriarte, E., 2001. Quantitative structure–toxicity relationships using tops-mode. 1. Nitrobenzene toxicity to tetrahymena pyriformis. *SAR QSAR Environ. Res.* 12, 309–324.
- Estrada, E., Uriarte, E., Gutierrez, Y., González, H., 2003b. Quantitative structure–toxicity relationships using tops-mode. 3. Structural factors influencing the permeability of commercial solvents through living human skin. *SAR QSAR Environ. Res.* 14, 145–163.
- Feron, V.J., Til, H.P., de Vrijer, F., Woutersen, R.A., Cassee, F.R., van Bladeren, P.J., 1991. Aldehydes: occurrence, carcinogenic potential, mechanism of action and risk assessment. *Mutat. Res.* 259, 363–385.
- Fragkaki, A.G., Angelis, Y.S., Tsantili-Kakoulidou, A., Koupparis, M., Georgakopoulou, C., 2009. Schemes of metabolic patterns of anabolic androgenic steroids for the estimation of metabolites of designer steroids in human urine. *J. Steroid Biochem. Mol. Biol.* 115, 44–61.
- Frank, J., 2002. STATISTICA, 8th ed. Statsoft, Inc.
- García-Lorenzo, A., Tojo, E., Teijeira, M., Rodríguez-Bercoval, F.J., González, M.P., Martínez-Zorzano, V.S., 2008. Cytotoxicity of selected imidazolium-derived ionic liquids in the human caco-2 cell line. sub-structural toxicological interpretation through a QSAR study. *Green Chem.* 10, 508–516.
- Glaab, V., Collins, A.R., Eisenbrand, G., Janzowski, C., 2001. DNA damaging potential and glutathione depletion of 2-cyclohexene-1-one in mammalian cells, compared to food relevant 2-alkenals. *Mutat. Res.* 497, 185–197.
- González, M.P., Dias, L.C., Helguera, A.M., 2004a. A topological sub-structural approach to the mutagenic activity in dental monomers. 2. Cycloaliphatic epoxides. *Polymer* 45, 5353–5359.
- González, M.P., Helguera, A.M., Cabrera, M.A., 2005a. Quantitative structure–activity relationship to predict toxicological properties of benzene derivative compounds. *Bioorg. Med. Chem.* 13, 1775–1781.

- González, M.P., Helguera, A.M., Collado, I.G., 2006. A topological substructural molecular design to predict soil sorption coefficients for pesticides. *Mol. Divers.* 10, 109–118.
- González, M.P., Helguera, A.M., Molina, R.R., García, J.F., 2004b. A topological substructural approach of the mutagenic activity in dental monomers. 1. Aromatic epoxides. *Polymer* 45, 2773–2779.
- González, M.P., Teran, M.C.T., Fall, Y., Dias, L.C., Helguera, A.M., 2005b. A topological sub-structural approach to the mutagenic activity in dental monomers. 3. Heterogeneous set of compounds. *Polymer* 46, 2783–2790.
- Gramatica, P., 2007. Principles of QSAR models validation: internal and external. *QSAR Comb. Sci.*, 1–9.
- Gutierrez, Y., Estrada, E., 2002. Modes lab, version 1.0.
- Hashimoto, K., Tani, H., 1985. Mutagenicity of acrylamide and its analogues in salmonella typhimurium. *Mutat. Res.* 158, 129–133.
- Hedman, R., Pettersson, H., 1997. Transformation of nivalenol by gastrointestinal microbes. *Arch. Anim. Nutr.* 50, 321–329.
- Helguera, A.M., Cabrera, M.A., Combes, R.D., González, M.P., 2006. Quantitative structure activity relationship for the computational prediction of nitrocompounds carcinogenicity. *Toxicology* 220, 51–62.
- Helguera, A.M., Cabrera, M.A., González, M.P., Molina, R.R., González, H.D., 2005. A topological sub-structural approach applied to the computational prediction of rodent carcinogenicity. *Bioorg. Med. Chem.* 13, 2477–2488.
- Helguera, A.M., Cordeiro, M.N.D.S., Cabrera, M.A., Combes, R.D., González, M.P., 2008a. Quantitative structure carcinogenicity relationship for detecting structural alerts in nitroso-compounds. species: rat; sex: male; route of administration: water. *Toxicol. Appl. Pharmacol.* 231, 197–207.
- Helguera, A.M., González, M.P., Briones, J.B., 2004. Tops-mode approach to predict mutagenicity in dental monomers. *Polymer* 45, 2045–2050.
- Helguera, A.M., González, M.P., Cordeiro, M.N.D.S., Cabrera, M.A., 2007. Quantitative structure carcinogenicity relationship for detecting structural alerts in nitroso-compounds. *Toxicol. Appl. Pharmacol.* 221, 189–202.
- Helguera, A.M., González, M.P., Cordeiro, M.N.D.S., Cabrera, M.A., 2008b. Quantitative structure carcinogenicity relationship for detecting structural alerts in nitroso-compounds. species: rat; sex: female; route of administration: gavage. *Chem. Res. Toxicol.* 21, 633–642.
- Holder, A.J., Ye, L., 2009. Quantum mechanical quantitative structure–activity relationships to avoid mutagenicity. *Dent. Mater.* 25, 20–25.
- Janzowski, C., Glaab, V., Mueller, C., Straesser, U., Kamp, H.G., Eisenbrand, G., 2003. α , β -unsaturated carbonyl compounds: induction of oxidative dna damage in mammalian cells. *Mutagenesis* 18, 465–470.
- Joe, Y.A., Goo, Y.-M., Lee, Y.Y., 1989. Microbiological oxidation of isophorone to 4-hydroxyisophorone and chemical transformation of 4-hydroxyisophorone to 2,3,5-trimethyl-p-benzoquinone. *Arch. Pharm. Res.* 12, 73–78.
- Kazius, J., McGuire, R., Bursi, R., 2005. Derivation and validation of toxicophores for mutagenicity prediction. *J. Med. Chem.* 48, 312–320.
- Klein, D., Randić, M., Babic, D., Lucic, B., Nikolic, S., Trinajstić, N., 1997. Hierarchical orthogonalization of descriptors. *Int. J. Quantum Chem.* 63, 215–222.
- Koleva, Y.K., Madden, J.C., Cronin, M.T.D., 2008. Formation of categories from structure–activity relationships to allow read-across for risk assessment: toxicity of α , β -unsaturated carbonyl compounds. *Chem. Res. Toxicol.* 21, 2300–2312.
- Kubinyi, H., 1994a. Variable selection in QSAR studies. 1. An evolutionary algorithm. *Quant. Struct. Act. Relat.* 13, 285.
- Kubinyi, H., 1994b. Variable selection in QSAR studies. 2. A highly efficient combination of systematic search and evolution. *Quant. Struct. Act. Relat.* 13, 393.
- Lalko, J., Lapczynski, A., McGinty, D., Bhatia, S., Letizia, C.S., Api, A.M., 2007. Fragrance material review on beta-ionone. *Food Chem. Toxicol.* 45, S241–S247.
- Lalonde, R.T., Cook, G.P., Perakyla, H., Bu, L., 1991. Structure–activity–relationships of bacterial mutagens related to 3-chloro-4-(dichloromethyl)-5-hydroxy-2(5h)-furanone an emphasis on the effect of stepwise removal of chlorine from the dichloromethyl group. *Chem. Res. Toxicol.* 4, 540–545.
- Landis, J.R., Koch, G.G., 1977. The measurement of observer agreement for categorical data. *Biometrics* 33, 159–174.
- Lucic, B., Nikolic, S., Trinajstić, N., Juric, D., 1995. The structure–property models can be improved using the orthogonalized descriptors. *J. Chem. Inf. Comput. Sci.* 35, 532–538.
- Lynch, D.M., Crovetto, A.J., 1972. Reactions of dichloromaleimides with alcohols, phenols, and thiols. *J. Heterocycl. Chem.* 9, 1027–1032.
- McDonald, T.A., Komulainen, H., 2005. Carcinogenicity of the chlorination disinfection by-product mx. *J. Environ. Sci. Health, Part C* 23, 163–214.
- Miller, I.A., Miller, E.C., 1983. Some historical aspects of n-aryl carcinogens and their metabolic activation. *Environ. Health Perspect.* 49, 3–12.
- Nery, R., 1971. The metabolic interconversion of arecoline and arecoline-1-oxide in the rat. *Biochem. J.* 122, 503–508.
- Netzeva, T.I., Worth, A.P., Aldenberg, T., Benigni, R., Cronin, M.T.D., Gramatica, P., Jaworska, J.S., Kahn, S., Klopman, P., Marchant, C.A., Myatt, G., Nikolova-Jeliazkova, N., Patlewicz, G.Y., Perkins, R., Roberts, D.W., Schultz, T.W., Stanton, D.T., van de Sandt, J.J.H., Tong, W., Veith, G., Yang, C., 2005. Current status of methods for defining the applicability domain of (quantitative) structure–activity relationships. *ATLA* 33, 155–173.
- Noma, Y., Takahashi, H., Asakawa, Y., 1991. Biotransformation of terpene aldehydes by euglena gracilis z. *Phytochemistry* 30, 1147–1151.
- OECD, 2007. Guidance document on the validation of (quantitative) structure–activity relationships [(Q)SAR] models, 49, OECD series on testing and assessment. Tech. Rep., Organisation for Economic Co-operation and Development, Paris, France.
- Pérez-Garrido, A., González, M.P., Escudero, A.G., 2008. Halogenated derivatives QSAR model using spectral moments to predict haloacetic acids (haa) mutagenicity. *Bioorg. Med. Chem.* 16, 5720–5732.
- Pérez-Garrido, A., Helguera, A.M., Girón-Rodríguez, F., Cordeiro, M.N.D.S., in press. QSAR models to predict mutagenicity of acrylates, methacrylates and α , β -unsaturated carbonyl compounds. *Dent. Mater.*
- Provost, F., Fawcett, T., 2001. Robust classification for imprecise environments. *Mach. Learn.* 42, 3.
- Purohit, V., Basu, A.K., 2000. Mutagenicity of nitroaromatic compounds. *Chem. Res. Toxicol.* 13, 673–692.
- Randić, M., 1991a. Correlation of enthalpy of octanes with orthogonal connectivity indices. *J. Mol. Struct. (Theochem.)* 233, 45–59.
- Randić, M., 1991b. Orthogonal molecular descriptors. *N. J. Chem.* 15, 517–525.
- Randić, M., 1991c. Resolution of ambiguities in structure–property studies by use of orthogonal descriptors. *J. Chem. Inf. Comput. Sci.* 31, 311–320.
- Richard, C., 2001. Concise international chemical assessment document 36: methyl cyanoacrylate and ethyl cyanoacrylate (draft). Tech. Rep., United Nations Environment Programme, International Labour Organization, and the World Health Organization, Geneva.
- Sasaki, J.C., Feller, R.S., Colvin, M.E., 2002. Metabolic oxidation of carcinogenic arylamines by p450 monooxygenases: theoretical support for one-electron transfer mechanism. *Mutat. Res.* 506/507, 79–89.
- Schultz, T.W., Yarbrough, J.W., Johnson, E.L., 2005. Structure–activity relationships for reactivity of carbonyl-containing compounds with glutathione. *SAR QSAR Environ. Res.* 16, 313–322.
- Smith, T.L., 1987. Method for detecting molecules containing amine or thiol groups. *U.S. Pat.* 4 680 272.
- Sosted, H., Basketter, D.A., Estrada, E., Johansen, J.D., Patlewicz, G.Y., 2004. Ranking of hair dye substances according to predicted sensitization potency: quantitative structure–activity relationships. *Contact Dermat.* 51, 241–254.
- Van Beerendonk, G.J.M., Nivard, M.J.M., Vogel, E.W., Nelson, S.D., Meerman, J.H.N., 1992. Formation of thymidine adducts and cross-linking potential of 2-bromoacrolein, a reactive metabolite of tris(2,3-dibromopropyl)phosphate. *Mutagenesis* 7, 19–24.
- van Noort, R., Brown, D., Causton, B.E., Combe, E.C., Fletcher, A.M., Lloyd, C.H., McCabe, J.F., Piddock, V., Sherriff, M., Strang, R., Waters, N.E., Watts, D.C., Williams, K., 1990. Dental materials: 1989 literature review. *J. Dent.* 18, 327–352.
- Vighi, M., Gramatica, P., Consolaro, F., Todeschini, R., 2001. Qsar and chemometrics approaches for setting water quality objectives for dangerous chemicals. *Ecotoxicol. Environ. Saf.* 49, 206–220.
- Weininger, D., 1988. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* 28, 31–36.
- Yang, C., Hasselgren, C.H., Boyer, S., Arvidson, K., Aveston, S., Dierkes, P., Benigni, R., Benz, R.D., Contrera, J., Kruhlik, N.L., Matthews, E.J., Han, X., Jaworska, J., Kemper, R.A., Rathman, J.F., Richard, A.M., 2008. Understanding genetic toxicity through data mining: The process of building knowledge by integrating multiple genetic toxicity databases. *Toxicol. Mech. Methods* 18 (2–3), 277–295.
- Yourtee, D., Holder, A.J., Smith, R., Morrill, J.A., Kostoryz, E., Brockmann, W., Glaros, A., Chappelow, C., Eick, D., 2001. Quantum mechanical quantitative structure activity relationships to avoid mutagenicity in dental monomers. *J. Biomater. Sci. Polym. Ed.* 12, 89–105.

Bibliografía

- [1] Lichtenstein P, Holm NV, Verkasalo PK, Iliadou A, Kaprio J, Koskenvuo M, et al. Environmental and heritable factors in the causation of cancer—analyses of cohorts of twins from Sweden, Denmark, and Finland. *N Engl J Med* 2000;343:78–85.
- [2] Arcos JC, Argus MF. *Chemical Induction of Cancer. Modulation and Combination Effects*. Boston: Birkhauser, 1995.
- [3] Philp RB. *Ecosystems and Human Health: Toxicology and Environmental Hazards*. Boca Raton, FL: Lewis Publishers, 2001.
- [4] Brusick D. *Principles of Genetic Toxicology*. 1987: Plenum Press, 2001.
- [5] Kosaka H, Wishnok JS, Miwa M, Leaf CD, Tannenbaum SR. Nitrosation by stimulated macrophages, inhibitors, enhancers, and substrates. *Carcinogenesis* 1989; 10:563–566.
- [6] Wink DA, Kasprzak KS, Maragos CM, Elespuru RK, Misra M, Dunams TM, et al. DNA deaminating ability and genotoxicity of nitric oxide and its progenitors. *Science* 1991;254:1001–1003.
- [7] Lo TL. Hard soft acids bases (HSAB) principle in organic chemistry. *Chem Rev* 1975;75:1–20.

-
- [8] Gates KS. An Overview of Chemical Processes That Damage Cellular DNA: Spontaneous Hydrolysis, Alkylation, and Reactions with Radicals. *Chem Res Toxicol* 2009;22:1747–1760.
- [9] Rajski SR, Williams RM. DNA cross-linking agents as antitumor drugs. *Chem Rev* 1998;98:2723–2795.
- [10] Noll DM, Mason TM, Miller PS. Formation and repair of interstrand crosslinks in DNA. *Chem Rev* 2006;106:277–301.
- [11] Schärer OD. DNA interstrand crosslinks: Natural and druginduced DNA adducts that induce unique cellular responses. *Chem-BioChem* 2005;6:27–32.
- [12] Jeffrey A. DNA modification by chemical carcinogens. *Pharmacol Ther* 1985; 28:237–272.
- [13] Cadet J, Delatour T, Douki T, Gasparutto D, Pouget J, Ravanat J, et al. Hydroxyl radicals and DNA base damage. *Mutat Res* 1996;424:9–21.
- [14] Reha D, Kabelác M, Ryjáček F, Sponer J, Sponer JE, Elstner M, et al. Intercalators. 1. Nature of stacking interactions between intercalators (ethidium, daunomycin, ellipticine, and 4',6-diaminide-2-phenylindole) and DNA base pairs. Ab initio quantum chemical, density functional theory, and empirical potential study. *J Am Chem Soc* 2003;124:3366–33762.
- [15] OECD. Test No. 471: Bacterial Reverse Mutation Test, OECD Guidelines for the Testing of Chemicals. Inf. téc., Organisation for Economic Co-operation and Development, Paris, France, 1997.
- [16] OECD. Test No. 476: In vitro Mammalian Cell Gene Mutation Test, OECD Guidelines for the Testing of Chemicals. Inf. téc., Organisation for Economic Co-operation and Development, Paris, France, 1997.

-
- [17] OECD. Test No. 475: Mammalian Bone Marrow Chromosome Aberration Test, OECD Guidelines for the Testing of Chemicals. Inf. téc., Organisation for Economic Co-operation and Development, Paris, France, 1997.
- [18] OECD. Test No. 473: In vitro Mammalian Chromosome Aberration Test, OECD Guidelines for the Testing of Chemicals. Inf. téc., Organisation for Economic Co-operation and Development, Paris, France, 1997.
- [19] OECD. Test No. 474: Mammalian Erythrocyte Micronucleus Test, OECD Guidelines for the Testing of Chemicals. Inf. téc., Organisation for Economic Co-operation and Development, Paris, France, 1997.
- [20] Ames BN, McCann J, Yamasaki E. Methods for Detecting Carcinogens and Mutagens with the Salmonella/Mammalian-Microsome Mutagenicity Test. *Mutat Res* 1975;31:347–364.
- [21] Maron DM, Ames BN. Revised Methods for the Salmonella Mutagenicity Test. *Mutat Res* 1983;113:173–215.
- [22] Gatehouse D, Haworth S, Cebula T, Gocke E, Kier L, Matsushima T, et al. Recommendations for the Performance of Bacterial Mutation Assays. *Mutat Res* 1994; 312:217–233.
- [23] Natarajan AT, Tates AD, van Buul PPW, Meijers M, de Vogel N. Cytogenetic Effects of Mutagens/Carcinogens after Activation in a Microsomal System In Vitro, I. Induction of Chromosome Aberrations and Sister Chromatid Exchanges by Diethylnitrosamine (DEN) and Dimethylnitrosamine (DMN) in CHO Cells in the Presence of Rat-Liver Microsomes. *Mutat Res* 1976;37:83–90.
- [24] Matsuoka A, Hayashi M, Ishidate MJ. Chromosomal Aberration Tests on 29 Chemicals Combined with S9 Mix In vitro. *Mutat Res* 1979;66:277–290.

-
- [25] Tamura G, Gold C, Ferro-Luzzi A, N AB. Fecalase: A Model for Activation of Dietary Glycosides to Mutagens by Intestinal Flora. *Proc Natl Acad Sci USA* 1979;77:4961–4965.
- [26] Matsushima T, Sawamura M, Hara K, Sugimura T. *In vitro* Metabolic Activation in Mutagenesis Testing. North Holland: Elsevier, 85–88.
- [27] Elliott BM, Combes RD, Elcombe CR, Gatehouse DG, Gibson GG, Mackay JM, et al. Alternatives to Aroclor 1254-induced S9 in *in vitro* Genotoxicity Assays. *Mutagenesis* 1992;7:175–177.
- [28] Galloway SM, Aardema MJ, Ishidate MJ, Ivett JL, Kirkland DJ, Morita T, et al. Report from Working Group on *in Vitro* Tests for Chromosomal Aberrations. *Mutat Res* 1994;312:241–261.
- [29] Commission E. Official J Eur Community 1986;L358(1).
- [30] Environment Directorate Joint Meeting of the Chemicals Committee and the Working Party on Chemicals, pesticides and Biotechnology. OECD Series on Testing and Assessment. Number 49. The report from the Expert Group on (Quantitative) Structure-Activity Relationships [(Q)SARs] on the Principles for the Validation of (Q)SARs. Paris, France: OECD, 2004.
- [31] OECD. The Report from the Expert Group on (Quantitative) Structure Activity Relationship ([Q]SARs) on the Principles for the Validation of (Q)SARs, 49, OECD Series on Testing and Assessment. Inf. téc., Organisation for Economic Co-operation and Development, Paris, 2004.
- [32] OECD. Guidance Document on the Validation of (Quantitative) Structure-Activity Relationships [(Q)SAR] Models, 49, OECD Series on Testing and Assessment. Inf. téc., Organisation for Economic Co-operation and Development, Paris, France, 2007.

-
- [33] Rosenkranz HS, Klopman G. Structural Alerts to Genotoxicity- The Interaction of Human and Artificial-Intelligence. *Mutagenesis* 1990;5:333.
- [34] Klopman G. Artificial-Intelligence Approach to Structure Activity Studies- Computer Automated Structure Evaluation of Biological-Activity of Organic-Molecules. *J Am Chem Soc* 1984;106:7315.
- [35] Klopman G, Rosenkranz HS. Approaches to SAR in Carcinogenesis and Mutagenesis- Prediction of Carcinogenicity/Mutagenicity Using MULTI-CASE. *Mutat Res* 1994;305:33-46.
- [36] Cunningham AR, Klopman G, Rosenkranz HS. Identification of structural features and associated mechanisms of action for carcinogens in rats. *Mutat Res* 1998; 405:9.
- [37] Cunningham AR, Rosenkranz HS. Estimating the extent of the health hazard posed by high-production volume chemicals. *Environ Health Perspect* 2001;109:953.
- [38] Rosenkranz HS. Quantitative Structure-Activity Relationship (QSAR) Models of Chemical Mutagens and Carcinogens. Boca Raton: CRC Press, 175.
- [39] Klopman G. A Hierarchical Computer Automated Structure Evaluation Program .1. *Quant Struct-Act Relat* 1992;11:176-184.
- [40] Klopman G, Rosenkranz H. Testing by Artificial-Intelligence-Computational Alternatives to the Determination of Mutagenicity. *Mutat Res* 1992;272:59.
- [41] Rosenkranz H, Cunningham A, Zhang Y, Claycamp H, Macina O, Sussman N, et al. Development, characterization and application of predictive-toxicology models. *SAR QSAR Environ Res* 1999;10:277-298.
- [42] Enslein K, Gombar V, Blake B. Use of SAR in Computer-Assisted Prediction of Carcinogenicity and Mutagenicity of Chemicals by the TOPKAT Program. *Mutat Res* 1994;305:47-61.

- [43] Enslein K. An overview of structure-activity-relationships as an alternative to testing in animals for carcinogenicity, mutagenicity, dermal and eye irritation, and acute oral toxicity. *Toxicol Ind Health* 1988;4:479–498.
- [44] Stouch TR, Jurs PC. Computer-assisted studies of molecular-structure and genotoxic activity by pattern-recognition techniques. *Environ Health Perspect* 1985; 61:329.
- [45] Jurs PC, Chou JT, Yuan MJ. Computer-Assisted Structure-Activity Studies of Chemical Carcinogens. A Heterogeneous Data Set. *J Med Chem* 1979;22:476.
- [46] Serra JR, Thompson ED, Jurs PC. Development of binary classification of structural chromosome aberrations for a diverse set of organic compounds from molecular structure. *Chem Res Toxicol* 2003;16:153.
- [47] Sanderson D, Earnshaw C. Computer-prediction of possible toxic action from chemical-structure - the DEREK system. *Hum Exp Toxicol* 1991;10:261–273.
- [48] Ridings JE, Barratt MD, Cary R, Earnshaw GG, Eggington CE, Ellis MK, et al. *Toxicology* 1996;106:267.
- [49] Greene N, Judson P, Langowski C, Marchant C. SAR QSAR *Environ Res* 1999; 10:299–314.
- [50] Benigni R, Bossa C. Structure alerts for carcinogenicity, and the *Salmonella* assay system: A novel insight through the chemical relational databases technology. *Mutat Res* 2008;659:248–261.
- [51] Benigni R, Bossa C, Jeliaskova NG, Netzeva TI, Worth AP, editores. The Benigni/Bossa rulebase for mutagenicity and carcinogenicity-a module of Toxtree. Luxembourg: EUR 23241 EN, EUR-Scientific and Technical Report Series, Office for the Official Publications of the European Communities, 2008.

- [52] Benigni R, Bossa C, Netzeva TI, Worth AP, editores. Collection and evaluation of (Q)SAR models for mutagenicity and carcinogenicity. Luxembourg: EUR 22772 EN, EUR-Scientific and Technical Report Series, Office for the Official Publications of the European Communities, 2007.
- [53] Benigni R. Structure-activity Relationship studies of chemical mutagens and carcinogens: Mechanistic investigations and prediction approaches. *Chem Rev* 2005; 105:1767.
- [54] Benigni R, Bossa C, Netzeva T, Worth A. Collection and evaluation of (Q)SAR models for mutagenicity and carcinogenicity. *Inf. téc., Institute for Health and Consumer Protection Toxicology and Chemical Substances Unit European Chemicals Bureau*, 2004.
- [55] Hatch FT, Colvin ME. Quantitative structure-activity (QSAR) relationships of mutagenic aromatic and heterocyclic amines. *Mutat Res* 1997;376:87–96.
- [56] Benigni R, Passerini L, Gallo G, Giorgi F, Cotta-Ramusino M. QSAR models for discriminating between mutagenic and nonmutagenic aromatic and heteroaromatic amines. *Environ Mol Mutagen* 1998;32:75–83.
- [57] Maran U, Karelson M, Katritzky AR. A comprehensive QSAR treatment of the genotoxicity of heteroaromatic and aromatic amines. *Quant Struct-Act Relat* 1999; 18:3.
- [58] Basak S, Gute B, Grunwald G. Assessment of the mutagenicity of aromatic amines from theoretical structural, parameters: A hierarchical approach. *SAR AND QSAR IN ENVIRONMENTAL RESEARCH* 1999;10(2-3):117–129.
- [59] Benigni R, Giuliani A, Franke R, Gruska A. Quantitative structure-activity relationships of mutagenic and carcinogenic aromatic amines. *Chem Rev* 2000; 100:3697–3714.

- [60] Franke R, Gruska A, Giuliani A, Benigni R. Prediction of rodent carcinogenicity of aromatic amines: a quantitative structure-activity relationships model. *Carcinogenesis* 2001;22:1561.
- [61] Glende C, Schmitt H, Erdinger L, Engelhardt G, Boche G. Transformation of mutagenic aromatic amines into non-mutagenic species by alkyl substituents Part I: Alkylation *ortho* to the amino function. *Mutat Res* 2001;498:19.
- [62] Glende C, Klein M, Schmitt H, Erdinger L, Boche G. Transformation of mutagenic aromatic amines into non-mutagenic species by alkyl substituents Part II: Alkylation far away from the amino function. *Mutat Res* 2002;515:15–38.
- [63] Basak S, Gute B, Mills D, Hawkins D. Quantitative molecular similarity methods in the property/toxicity estimation of chemicals: a comparison of arbitrary versus tailored similarity spaces. *JOURNAL OF MOLECULAR STRUCTURE-THEOCHEM* 2003;622(1-2):127–145.
- [64] Gramatica P, Consonni V, Pavan M. Prediction of aromatic amines mutagenicity from theoretical molecular descriptors. *SAR AND QSAR IN ENVIRONMENTAL RESEARCH* 2003;14(4):237–250.
- [65] Cash G, Anderson B, Mayo K, Bogaczyk S, Tunkel J. Predicting genotoxicity of aromatic and heteroaromatic amines using electrotopological state indices. *MUTATION RESEARCH-GENETIC TOXICOLOGY AND ENVIRONMENTAL MUTAGENESIS* 2005;585(1-2):170–183.
- [66] Bhat K, Hayik S, Sztandera L, Bock C. Mutagenicity of aromatic and heteroaromatic amines and related compounds: A QSAR investigation. *QSAR & COMBINATORIAL SCIENCE* 2005;24(7):831–843.
- [67] Felton JS, Knize MG, Wu RW, Colvin ME, Hatch FT, Malfatti MA. Mutagenic potency of food-derived heterocyclic amines. *MUTATION RESEARCH-*

- FUNDAMENTAL AND MOLECULAR MECHANISMS OF MUTAGENESIS
2007;616(1-2):90–94.
- [68] Benigni R, Bossa C, Netzeva T, Rodomonte A, Tsakovska I. Mechanistic QSAR of aromatic amines: New models for discriminating between homocyclic mutagens and nonmutagens, and validation of models for carcinogens. ENVIRONMENTAL AND MOLECULAR MUTAGENESIS 2007;48(9):754–771.
- [69] Singh J, Shaika B, Agrawal VK, Khadikar PV. Modeling of mutagenicity of aromatic and heteroaromatic amines in Salmonella typhimurium TA98: Role of hydrophobicity and topological indices. JOURNAL OF THE INDIAN CHEMICAL SOCIETY 2008;85(5):517–535.
- [70] Borosky GL. Carcinogenic carbocyclic and heterocyclic aromatic amines: A DFT study concerning their mutagenic potency. JOURNAL OF MOLECULAR GRAPHICS & MODELLING 2008;27(4):459–465.
- [71] Toropov AA, Toropova AP, Benfenati E. QSAR Modelling for Mutagenic Potency of Heteroaromatic Amines by Optimal SMILES-based Descriptors. CHEMICAL BIOLOGY & DRUG DESIGN 2009;73(4):482.
- [72] Biagi LG, Hrelia P, Guerra MG, Paolini M, Barbaro AM, Cantelli-Forti G. Structure-activity relationship of nitroimidazo (2,1-b) thiazoles in the salmonella mutagenicity assay. Arch Toxicol 1986;9:425.
- [73] Maynard AT, Pedersen LG, Posner HS, Mckinney JD. An abinitio study of the relationship between nitroarene mutagenicity and electron-affinity. Mol Pharmacol 1986;29:629.
- [74] Decompadre RL, Shusterman AJ, Hansch C. The role of hydrophobicity in the ames test - the correlation of the mutagenicity of nitropolycyclic hydrocarbons with partition-coefficients and molecular-orbital indexes. Int J Quantum Chem 1988;34:91.

- [75] Debnath AK, de Compadre RLL, Shusterman A, Hansch C. Quantitative structure-activity relationship investigation of the role of hydrophobicity in regulating mutagenicity in the ames test .2. Mutagenicity of aromatic and heteroaromatic nitro-compounds in salmonella-typhimurium TA100. *Environ Mol Mutagen* 1992;19:53.
- [76] Klein M, Voigtmann U, Haack T, Erdinger L, Boche G. From mutagenic to non-mutagenic nitroarenes: effect of bulky alkyl substituents on the mutagenic activity of 4-nitrobiphenyl in *Salmonella typhimurium* - Part I. Substituents ortho to the nitro group and in 2'-position. *Mutat Res* 2000;467:55.
- [77] Klein M, Erdinger L, Boche G. From mutagenic to non-mutagenic nitroarenes: effect of bulky alkyl substituents on the mutagenic activity of nitroaromatics in *Salmonella typhimurium* - Part II. Substituents far away from the nitro group. *Mutat Res* 2000;467:69.
- [78] Akin EW, Hoff JC, Lippy EC, Karelson M, Suzuki T, Solov'ev VP, et al. *Environ Health Perspect* 1982;46:7.
- [79] Nestmann ER, Lebel GL, Williams DT, Kowbel DJ. Mutagenicity of organic extracts from Canadian drinking water in the *Salmonella*/mammalian microsome assay. *Environ Mutagenesis* 1979;1:337-345.
- [80] Maruoka S S nd Yamanaka. Production of mutagenic substances by chlorination waters. *Mutat Res* 1980;79:381-386.
- [81] Maruoka S S nd Yamanaka. Mutagenic potencial of laboratory chlorinated river water. *Sci Total Environ* 1983;29:143-154.
- [82] Wilcox P, Williamson S. *Environ Health Perspect* 1986;69:141-149.
- [83] Morris RD, Audet AM, Angelino IF, Chalmers TC, Nosteller F. *Am J Public Health* 1992;82:955.

- [84] Koivusalo M, Jaakkola JJ, Vartiainen T, Hakulinen T, Karjalainen S, Pukkala E, et al. Am J Public Health 1994;84:1223.
- [85] Bull RJ, Birnbaum LS, Cantor KP, Rose JB, Butterworth BE, Pegram R, et al. Fundam Appl Toxicol 1995;28:155.
- [86] Waegemaekers THJM, Bensink MPM. Non-mutagenicity of 27 aliphatic acrylate esters in the Salmonella-microsome test. Mutat Res 1984;137:95–102.
- [87] Schweikl H, Schmalz G, Rackebrandt K. The mutagenic activity of unpolymers of resin monomers in *Salmonella typhimurium* and V79 cells. Mutat Res 1998;415:119–130.
- [88] Schweikl H, Schmalz G. Triethylene glycol dimethacrylate induces large deletions in the hprt gene of V79 cells. Mutat Res 1999;438:71–78.
- [89] Nieuwenhuijsen MJ, Toledano MB, Elliott P. J Expo Anal Environ Epidemiol 2000;10:586.
- [90] Vol 52: Chlorinated Drinking-Water; Chlorination By-products; Some other Halogenated Compounds; Cobalt and Cobalt Compounds. Lyon: International Agency for Research on Cancer, 1991.
- [91] Bull RJ, Sanchez IM, Nelson MA, Larson JL, Lansing AJ. Liver tumor induction in B6C3F1 mice by dichloroacetate and trichloroacetate. Toxicology 1990;63:341–359.
- [92] DeAngelo AB, Daniel FB, Stober JA, Olson GR. The carcinogenicity of dichloroacetic acid in the male B6C3F1 mouse. Fundam Appl Toxicol 1991;16:337–347.
- [93] DeAngelo AB, Daniel FB, Most BM, Olson GR. The carcinogenicity of dichloroacetic acid in the male Fischer 344 rat. Toxicology 1996;114:207–221.
- [94] Herbert V, Gardner A, Colman N. Mutagenicity of dichloroacetate, an ingredient of some formulations of pangamic acid (trade name "vitamin B₁₅"). Am J Clin Nutr 1980;33:1179–1182.

- [95] Nestman ER, Chu I, Kowbel DJ, Matula TI. Short-lived mutagen in *Salmonella* produced by reaction of trichloroacetic acid and dimethyl sulphoxide. *Can J Genet Cytol* 1980;22:35–40.
- [96] DeMarini DM, Perry E, Shelton ML. Dichloroacetic acid and related compounds: introduction of prophage in *E. coli* and mutagenicity and mutation spectra in *Salmonella* TA100. *Mutagenesis* 1994;9:429–437.
- [97] Giller S, Le Curieux F, Erb F, Marzin D. Comparative genotoxicity of halogenated acetic acids found in drinking water. *Mutagenesis* 1997;12:321–328.
- [98] Kargalioglu Y, McMillan BJ, Minear RA, Plewa MJ. Analysis of the cytotoxicity and mutagenicity of drinking water disinfection byproducts in *Salmonella typhimurium*. *Teratog Carcinog Mutagen* 2002;22:113–128.
- [99] Kundu B, Richardson SD, Swartz PD, Matthews PP, Richard AM, DeMarini DM. Mutagenicity in *Salmonella* of trihalomethanes: a recently recognized class of disinfection byproducts in drinking water. *Mutat Res* 2004;562:39–65.
- [100] Plewa MJ, Wagner ED, Richardson SD, Thruston J A D, Woo YT, McKague AB. Chemical and biological characterization of newly discovered iodoacid drinking water disinfection byproducts. *Environ Sci Technol* 2004;38:4713–4722.
- [101] Plewa MJ, Kargalioglu Y, Vankerk D, Minear RA, Wagner ED. Mammalian cell cytotoxicity and genotoxicity analysis of drinking water disinfection byproducts. *Environ Mol Mutagen* 2002;40:134–142.
- [102] Plewa MJ, Wagner ED, Jazwierska P, Richardson SD, Chen PH, McKague AB. Halonitromethane drinking water disinfection byproducts: chemical characterization and mammalian cell cytotoxicity and genotoxicity. *Environ Sci Technol* 2004;38:62–68.
- [103] Richardson SD, Thruston J A D, Rav-Acha C, Groisman L, Popilevsky I, Juraev O, et al. Tribromopyrrole, brominated acids, and other disinfection byproducts

- produced by disinfection of drinking water rich in bromide. *Environ Sci Technol* 2003;37:3782–3793.
- [104] Richard AM, Hunter III ES. Quantitative Structure-Activity Relationships for the Developmental Toxicity of Haloacetic Acids in Mammalian Whole Embryo Culture. *Teratology* 1996;53:352–360.
- [105] Richard AM. Structure-based methods for predicting mutagenicity and carcinogenicity: are we there yet?. *Mutat Res* 1998;400:493–507.
- [106] Woo YT, Lai DY, Arcos JC, Argus MF, Cimino MC, DeVito S, et al. Peroxisomes: Biology and Importance in Toxicology and Medicine. *J Environ Sci Health* 1997; 15:139.
- [107] Venkatapathy R, Bruce R, Moudgal C. Impact of Addition of Halogens to the α -Carbon of Acetic Acids on Mutagenicity and Developmental Toxicity Endpoints, 2004. Presented at the EPA Science Forum, Mandarin Oriental Hotel, Washington, DC; <http://www.epa.gov/ord/scienceforum/2004/poster-ord-NtoZ.htm>.
- [108] March J. *Advanced Organic Chemistry: Reactions, Mechanisms and Structure*. New York: John Wiley & Sons, 1992.
- [109] Worth AP, Van Leeuwen CJ, Hartung T. The prospects for using (Q)SARs in a changing political environment-high expectations and a key role for the European Commission's joint research centre. *SAR QSAR Environ Res* 2004;5:31–43.
- [110] Jaworska JS, Comber M, Auer C, Van Leeuwen CJ. Summary of a workshop on regulatory acceptance of (Q)SARs for human health and environmental endpoints. *Environ Health Persp* 2004;111:1358–1360.
- [111] Feron VJ, Til HP, de Vrijer F, Woutersen RA, Cassee FR, van Bladeren PJ. Aldehydes: Occurrence, carcinogenic potential, mechanism of action and risk assessment. *Mutat Res* 1991;259:363–385.

- [112] Boelens MH, Gemert LJ. Organoleptic properties of aliphatic aldehydes. *Perfumer Flavorist* 1987;12:31–43.
- [113] van Noort R, Brown D, Causton BE, Combe EC, Fletcher AM, Lloyd CH, et al. Dental materials: 1989 literature review. *J Dent* 1990;18:327–352.
- [114] Asmussen E. Factors affecting the quantity of remaining double bonds in restorative resin polymers. *Scand J Dent Res* 1982;90:490–496.
- [115] Imazato S, McCabe JF, Tarumi H, Ehara A, Ebisu S. Degree of conversion of composites measured by DTA and FTIR. *Dent Mater* 2001;17:178–183.
- [116] Aptula AO, Roberts DW. Mechanistic applicability domains for non-animal-based toxicological end points: General principles and application to reactive toxicity. *Chem Res Toxicol* 2006;19:1097–1105.
- [117] Yourtee D, Holder AJ, Smith R, Morrill JA, Kostoryz E, Brockmann W, et al. Quantum mechanical quantitative structure activity relationships to avoid mutagenicity in dental monomers. *J Biomater Sci Polymer Edn* 2001;12:89–105.
- [118] Holder AJ, Ye L. Quantum mechanical quantitative structure-activity relationships to avoid mutagenicity. *Dent Mater* 2009;25:20–25.
- [119] Morales AH, Pérez MAC, González MP, Ruiz RM, Díaz HG. A topological substructural approach applied to the computational prediction of rodent carcinogenicity. *Bioorg Med Chem Lett* 2005;13:2477–2488.
- [120] González MP, Teran MCT, Fall Y, Dias LC, Helguera AM. A topological substructural approach to the mutagenic activity in dental monomers. 3. Heterogeneous set of compounds. *Bioorg Med Chem* 2005;46:2783–2790.
- [121] Benigni R, Conti L, Crebelli R, Rodomonte A, Vari MR. Simple and α , β -Unsaturated aldehydes: Correct prediction of genotoxicity activity through structure-activity relationship models. *Environ Mol Mutagen* 2005;46:268–280.

- [122] Koleva KY, Madden JC, Cronin MTD. Formation of Categories from Structure-Activity Relationships To Allow Read-Across for Risk Assessment: Toxicity of α,β -Unsaturated Carbonyl Compounds. *Chem Res Toxicol* 2008;21:2300–2312.
- [123] Yang C, Hasselgren CH, Boyer S, Arvidson K, Aveston S, Dierkes P, et al. Understanding genetic toxicity through data mining: The process of building knowledge by integrating multiple genetic toxicity databases. *Toxicol Mech Meth* 2008;18(2-3):277–295.
- [124] Saenger W, Jacob J, Gessler K, Steiner T, Daniel S, Sanbe H, et al. Structures of the common cyclodextrins and their larger analogues - Beyond the doughnut. *Chem Rev* 1998;98:1787–1802.
- [125] Loftsson T, Duchêne D. Historical perspective. Cyclodextrins and their pharmaceutical applications. *Int J Pharm* 2007;329:1–11.
- [126] Szente L, Szejtli J. Cyclodextrins as food ingredients. *Trends in Food Science & Technology* 2004;15:137–142.
- [127] Suzuki T. A Nonlinear Group Contribution Method for Predicting the Free Energies of Inclusion Complexation of Organic Molecules with α - and β -Cyclodextrins. *J Chem Inf Comput Sci* 2001;41:1266–1273.
- [128] Pérez F, Jaime C, Sánchez-Ruiz X. MM2 Calculations on Cyclodextrins: Multimodel Inclusion Complexes. *J Org Chem* 1995;60:3840–3845.
- [129] Matsui Y, Nishioka T, Fujita T. Quantitative Structure-Reactivity Analysis of the Inclusion Mechanism by Cyclodextrins. *Top Curr Chem* 1985;128:61–89.
- [130] Davis DM, Savage JR. Correlation Analysis of the Host-Guest Interaction of α -Cyclodextrin and Substituted Benzenes. *J Chem Res(S)* 1993;:94–95.

- [131] Park JH, Nah TH. Binding Forces Contributing to the Complexation of Organic Molecules with β -Cyclodextrin in Aqueous Solution. *J Chem Soc Perkin Trans* 1994;2:1359–1362.
- [132] Klein CT, Polheim D, Viernstein H, Wolschann P. A Method for Predicting the Free Energies of Complexation between β -Cyclodextrin and Guest Molecules. *J Inclusion Phenom Macrocyclic Chem* 2000;36:409–423.
- [133] Liu L, Guo QX. Wavelet Neural Network and its Application to the Inclusion of β -Cyclodextrin with Benzene Derivatives. *J Chem Inf Comput Sci* 1999;39:133–138.
- [134] Suzuki T, Ishida M, Fabian WMF. Classical QSAR and Comparative Molecular Field Analyses of the Host-Guest Interaction of Organic Molecules with Cyclodextrins. *J Comput-Aided Mol Des* 2000;14:669–678.
- [135] Cramer I R D, Patterson DE, Bunce JD. Comparative Molecular Field Analysis (CoMFA). 1. Effect of Shape on Binding of Steroids to Carrier Proteins. *J Am Chem Soc* 1988;110:5959–5967.
- [136] Katritzky AR, Fara DC, Yang HF, Karelson M, Suzuki T, Solov'ev VP, et al. Quantitative Structure-Property Relationship Modeling of β -Cyclodextrin Complexation Free Energies. *J Chem Inf Comput Sci* 2004;44:529–541.
- [137] Kola I, Landis J. Can the pharmaceutical industry reduce attrition rates? *Nat Rev, Drug Discov* 2004;2:711–715.
- [138] Loftsson T, Brewster M, Masson M. Role of cyclodextrins in improving oral drug delivery. *Am J Drug Deliv* 2004;2:261–275.
- [139] Davis ME, Brewster M. Cyclodextrin-based pharmaceuticals: past, present and future. *Nat Rev, Drug Discov* 2004;3:1023–1035.

- [140] Avdeef A, Bendels S, Tsinman O, Tsinman K, Kansy M. Solubility excipient classification gradient maps. *Pharm Res* 2007;24:530–545.
- [141] Kim C, Park J. Solubility enhancers for oral drug delivery. *Am J Drug Deliv* 2004;2:113–130.
- [142] Loftsson T, Jarho P, Masson M, Jarvinen T. Cyclodextrins in drug delivery. *Expert Opin Drug Deliv* 2005;2:335–351.
- [143] Liu R. Water-insoluble drug formulation. Englewood: CO Interpharm Press, 2000.
- [144] Irie T, Uekama K. Pharmaceutical applications of cyclodextrins. III. Toxicological issues and safety evaluation. *J Pharm Sci* 1997;86:147–162.
- [145] Szejtli J, Szente L. Elimination of bitter, disgusting tastes of drugs and foods by cyclodextrins. *Eur J Pharm Biopharm* 2005;61:115–125.
- [146] Lantz A, Rodriguez M, Wetterer S, Armstrong D. Estimation of association constants between oral malodor components and various native and derivatized cyclodextrins. *Anal Chim Acta* 2006;557:184–190.
- [147] Uekama K. Cyclodextrins in drug delivery. *Adv Drug Deliv Rev* 1999;36:1–2.
- [148] Duchêne D, editor. *Cyclodextrins and their Industrial Uses*. Paris: Editions de Santé Paris, 1987.
- [149] Horvath G, Premkumar T, Boztas A, Lee E, Jon S, Geckeler KE. Supramolecular nanoencapsulation as a tool: Solubilization of the anticancer drug trans-dichloro(dipyridine)platinum(II) by complexation with beta-cyclodextrin. *Mol Pharm* 2008;5:358–363.
- [150] *Handbook of pharmaceutical excipients*, second edition. London: Pharmaceutical Press, 186–190.

- [151] Parfitt K. Cyclodextrins-The complete drug reference. London: Martindale Pharmaceutical Press, 1999.
- [152] Horský J, Pitha J. Hydroxypropyl cyclodextrins: Potential synergism with carcinogens. *J Pharm Sci* 1996;85:96–100.
- [153] Westerberg G, Wiklund L. β -cyclodextrin reduces bioavailability of orally administered [H-3]benzo[a]pyrene in the rat. *J Pharm Sci* 2005;94:114–119.
- [154] Ono N, Hirayama F, Arima H, Uekama K, Rytting JH. Model analysis for oral absorption of a drug/cyclodextrin complex involving competitive inclusion complexes. *J Incl Phenom Macroc Chem* 2002;44:93–96.
- [155] Zheng Y, Zuo Z, Chow AHL. Lack of effect of β -cyclodextrin and its water-soluble derivatives on in vitro drug transport across rat intestinal epithelium. *Int Pharm J* 2006;309:123–128.
- [156] Lipkowitz KB. Applications of Computational Chemistry to the Study of Cyclodextrins. *Chem Rev* 1998;98:1829–1873.
- [157] Liu J, Pan D, Tseng YF, Hopfinger A. 4D-QSAR analysis of a series of antifungal P450 inhibitors and 3D-pharmacophore comparisons as a function of alignment. *J Chem Inf Comp Sci* 2003;43(6):2170–2179.
- [158] Hansch C, Leo A. Exploring QSAR. 1. Fundamentals and Applications in Chemistry and Biology. Washington, DC.: American Chemical Society, 1995.
- [159] Free SM, Wilson JW. Mathematical Contribution to Structure-Activity Studies. *J Med Chem* 1964;7:395.
- [160] Estrada E. How the Parts Organize in the Whole? A Top-Down View of Molecular Descriptors and Properties for QSAR and Drug Design. *Mini Rev Med Chem* 2008;8:213–221.

- [161] Pérez-Garrido A, González MP, Escudero AG. Halogenated derivatives QSAR model using spectral moments to predict haloacetic acids (HAA) mutagenicity. *Bioorg Med Chem* 2008;16:5720–5732.
- [162] Pérez-Garrido A, Helguera AM, Caravaca G, Cordeiro MNDS, Escudero AG. A TOPological Substructural MOlecular Design approach for predicting mutagenesis end-points of α , β -unsaturated carbonyl compounds. *Toxicology* 2010;268:64–77.
- [163] Pérez-Garrido A, Helguera AM, Girón-Rodríguez F, Cordeiro MNDS. QSAR models to predict mutagenicity of acrylates, methacrylates and α , β -unsaturated carbonyl compounds. *Dental material* 2010;26:397–415.
- [164] Pérez-Garrido A, Helguera AM, Cordeiro MNDS, Abellán A, Escudero AG. Convenient QSAR model for predicting the complexation of structurally diverse compounds with β -cyclodextrins. *Bioorg Med Chem* 2009;17:896–904.
- [165] Pérez-Garrido A, Helguera AM, Cordeiro MNDS, Escudero AG. QSPR Modelling With the Topological Substructural Molecular Design Approach: β -Cyclodextrin Complexation. *J Pharm Sci* 2009;98:4557–4576.
- [166] LaLonde RT, Leo H, Perakyla H, Dence CW, Farrell RP. Associations of the bacterial mutagenicity of halogenated 2(5h)-furanones with their MNDO-PM3 computed properties and mode of reactivity with sodium-borohydride. *Chem Res Toxicol* 1992;5:392.
- [167] Guengerich FP. Mechanisms of Mutagenicity of DNA Adducts Derived from Alkyl and Vinyl Halides. *Jpn J Toxicol Environ Health* 1997;43:69–82.
- [168] Woo YT, Lai D, Arcos JC, Argus MF. Chemical Induction of Cancer, Structural Bases and Biological Mechanism, Vol IIIB. Aliphatic and Polyhalogenated Carcinogens. Orlando, Florida: Academic Press, 1985.

- [169] Woo YT, Lai DY, McLain JL, Ko Manibusan M, Dellarco V. Use Mechanism-Based Structure-Activity Relationships Analysis in Carcinogenic Potential Ranking for Drinking Water Disinfection By-Products. *Environ Health Perspect* 2002;110:75.
- [170] Simon P, Epe B, Mützel P, Schiffmann D, Wild D, Ottenwälder H, et al. Synthesis and genotoxicity of acetoxyoxirane, the epoxide of vinyl acetate. *J Biochem Toxicol* 1997;1:43–55.
- [171] Castelain PH, Criado B, Cornet M, Laib R, Rogiers V, Kirsch-Volders M. Comparative mutagenicity of structurally related aliphatic epoxides in a modified Salmonella/microsome assay. *Mutagenesis* 1993;8:387–393.
- [172] Eder E, Henschler D, Neudecker T. Mutagenic properties of allylic and α, β -unsaturated compounds: consideration of alkylating mechanism. *Xenobiotica* 1982;12:831–848.
- [173] Eder E, Weinfurtner E. Mutagenic and Carcinogenic Risk of Oxygen Containing Chlorinated C-3 Hydrocarbons: Putative Secondary Products of C-3 Chlorohydrocarbons and Chlorination of Water. *Chemosphere* 1994;29:2455–2466.
- [174] Van Beerendonk GJM, Nivard MJM, Vogel EW, Nelson SD, Meerman JHN. Formation of thymidine adducts and cross-linking potential of 2-bromoacrolein, a reactive metabolite of tris(2,3-dibromopropyl)phosphate. *Mutagenesis* 1992;7:19–24.
- [175] McGregor DB, Cruzan G, Callander RD, May K, Banton M. The mutagenicity testing of tertiary-butyl alcohol, tertiary-butyl acetate and methyl tertiary-butyl ether in *Salmonella typhimurium*. *Mutat Res* 2005;565:181.
- [176] Stolzenberg SJ, Hine CH. Mutagenicity of halogenated and oxygenated three-carbon compounds. *J Toxicol Environ Health* 1979;5:1149–1158.
- [177] Simmon VF, Kauhanen K, Tardiff RG. Progress in genetic toxicology. Amsterdam: Elsevier/North Holland Press, 249–268.

- [178] Heck JD, Vollmuth TA, Cifone MA, Jagannath DR, Myhr B, Curren RD. An evaluation of food flavoring ingredients in a genetic toxicity screening battery. *The Toxicologist* 1989;9:257.
- [179] Philipose B, Singh R, Khan KA, Giri AK. Comparative mutagenic and genotoxic effects of three propionic acid derivatives ibuprofen, ketoprofen and naproxen. *Mutat Res* 1997;393:123–131.
- [180] Provost F, Fawcett T. Robust Classification for Imprecise Environments. *Mach Learn* 2001;42:3.
- [181] Purohit V, Basu AK. Mutagenicity of nitroaromatic compounds. *Chem Res Toxicol* 2000;13:673–692.
- [182] Miller KJ. Additivity methods in molecular polarizability. *J Am Chem Soc* 1990; 112:8533–8542.
- [183] Sasaki JC, Feller RS, Colvin ME. Metabolic oxidation of carcinogenic arylamines by P450 monooxygenases: theoretical support for one-electron transfer mechanism. *Mutat Res* 2002;506/507:79–89.
- [184] Eder E, Hoffman C, Bastian H, Deininger C, Scheckenbach S. Molecular mechanisms of DNA damage initiated by α, β -Unsaturated carbonyl compounds as criteria for genotoxicity and mutagenicity. *Environ Health Perspect* 1990;88:99–106.
- [185] O'Brien SE, de Groot MJ. Greater than the sum of its parts: combining models for useful ADMET prediction. *J Med Chem* 2005;48:1287–1291.
- [186] Dearfield KL, Harrington-Brock K, Doerr CL, Rabinowitz JR, Moore MM. Genotoxicity in mouse lymphoma cells of chemicals capable of Michael addition. *Mutagenesis* 1991;6:519–525.

- [187] Ciaccio PJ, Gicquel E, O'Neill PJ, Scribner HE, Vandenberghe YL. Investigation of the positive response of ethyl acrylate in the mouse lymphoma genotoxicity assay. *Toxicol Sci* 1998;46:324–332.
- [188] Schultz T, Cronin M, Netzeva T. The present status of QSAR in toxicology. *J Mol Struc-Theochem* 2003;622:23–38.
- [189] Glaab V, Collins AR, Eisenbrand G, Janzowski C. DNA damaging potential and glutathione depletion of 2-cyclohexene-1-one in mammalian cells, compared to food relevant 2-alkenals. *Mutat Res* 2001;497:185–197.
- [190] Janzowski C, Glaab V, Mueller C, Straesser U, Kamp HG, Eisenbrand G. α , β -Unsaturated carbonyl compounds: induction of oxidative DNA damage in mammalian cells. *Mutagenesis* 2003;18:465–470.
- [191] Schultz TW, Yarbrough JW, Johnson EL. Structure-activity relationships for reactivity of carbonyl-containing compounds with glutathione. *SAR and QSAR in Environmental Research* 2005;16:313–322.
- [192] Eriksson L, Jaworska J, Worth AP, Cronin MTD, McDowell RM, Gramatica P. Methods for Reliability and Uncertainty Assessment and for applicability Evaluations of Classification- and Regression-Based QSARs. *Environmental Health Perspectives* 2003;111:1361.
- [193] Netzeva TI, Worth AP, Aldenberg T, Benigni R, Cronin MTD, Gramatica P, et al. Current Status of Methods for Defining the Applicability Domain of (Quantitative) Structure-Activity Relationships. *ATLA* 2005;33:155–173.
- [194] Gramatica P. Principles of QSAR models validation: internal and external. *QSAR Comb Sci* 2007;00:1–9.
- [195] Vighi M, Gramatica P, Consolaro F, Todeschini R. QSAR and chemometrics approaches for setting water quality objectives for dangerous chemicals. *Ecotoxicol Environ Saf* 2001;49:206–220.

-
- [196] Cal K, Centkowska K. Use of cyclodextrins in topical formulations: Practical aspects. *Eur J Pharm Biopharm* 2008;68:467–478.
- [197] Fortin D, Vargas M. The spectrum of composites: New techniques and materials. *J AM Dent Assoc* 2000;131:26–37.
- [198] Fong H, Dickens S, Flaim G. Evaluation of dental restorative composites containing polyhedral oligomeric silsesquioxane methacrylate. *Dent Mater* 2005; 21:520–529.