Universidad Central "Marta Abreu" de Las Villas
Facultad Matemática, Física y Computación

Tesis en opción al Título de Máster en Bioinformática y Biología Computacional

EVALUACIÓN DE LA INFLUENCIA DE LOS ACANTILADOS DE ACTIVIDAD (*ACTIVITY CLIFFS*) EN LA MODELACIÓN QSAR.

AUTOR: Lic. José Luís Velázquez Libera.

TUTORES: Dr. Maykel Cruz Monteagudo.

Dr. Yunierkis Pérez Castillo.

Santa Clara, 2015

FRASES O PENSAMIENTOS

“Nunca consideres el estudio como una obligación, sino como la oportunidad para penetrar en el bello y maravilloso mundo del saber.”

“Hay una fuerza motriz más poderosa que el vapor, la electricidad y la energía atómica: la voluntad.”

“La imaginación es más importante que el conocimiento.”

“La mente es como un paracaídas… Sólo funciona si la tenemos abierta”

Albert Einstein.

DEDICATORIA

A mis padres, que son lo más grande que tengo en mi vida, que han sabido guiar mis pasos hasta hacerme el hombre soy y que han sacrificado todo por mí.

A mis abuelos, que en paz descansen, y que siempre llevaré en mi corazón.

A mi familia de Guantánamo: a mis tíos y a mis primos, que aunque los veo con poca frecuencia, los llevo en mi corazón y los tengo presente en todo momento.

A mi esposa que ha sido lo mejor que ha pasado en mi vida.

A mi suegra, que día a día en los últimos años, ha sido otra madre más para mí.

A mis amigos de toda la vida y que cada día están más lejos de mí.

A todo joven cubano que lucha por una vida mejor y más digna.

Con todo mi corazón.

AGRADECIMIENTOS

A mis padres por todo su apoyo, confianza, dedicación y sacrificio; por darme la vida y el ser; por guiar mis pasos y principios sin los cuales no sería hoy quien soy.

A mi esposa por su paciencia, amor y constante apoyo.

A mis tutores, Maykel y Chiqui, por todo su apoyo e infinita paciencia, y por su amistad sincera.

A mis profesores de la maestría, por su esfuerzo, por su tiempo, y por todo lo aprendido.

A la profesora Gladita por todo su apoyo y paciencia.

A los miembros del Grupo de Investigación de Simulaciones Moleculares y Diseño de Fármacos del Centro de Bioactivos Químicos, por su ayuda y por tenderme su mano desde los primeros momentos que llegué aquí.

A mis compañeros de la maestría, en especial a Rotceh, con quien he hecho largos viajes mes por mes desde Las Tunas.

A las tías de la beca, por toda su ayuda.

DECLARACIÓN JURADA DEL AUTOR

El que suscribe, José Luís Velázquez Libera, hago constar que el presente trabajo para optar por el Título de Máster en Bioinformática y Biología Computacional ha sido realizado en el Centro de Bioactivos Químicos (CBQ) de la Universidad Central "Marta Abreu" de Las Villas como parte de la culminación de los estudios de la Maestría en Bioinformática y Biología Computacional, autorizando a que el mismo sea utilizado por la institución, para los fines que estime conveniente, tanto de forma parcial como total y que además no podrá ser presentado en eventos ni publicado sin la autorización de la Universidad.

_____

Firma del Autor

RESUMEN

La modelación QSAR es un ejemplo de herramienta quimioinformática cuyo uso se ha extendido a diferentes esferas del desarrollo de la sociedad. El principal supuesto de las aproximaciones en modelos QSAR es la continuidad del espacio de las Relaciones Estructura-Actividad (SAR), la cual se puede ver afectada por la presencia de los *activity cliffs*. Estudios recientes han mostrado los efectos negativos de la presencia de los *activity cliffs* sobre la capacidad predictiva de los modelos QSAR. Sin embargo, no se reportan estudios en los que se evalúe el efecto de eliminarlos de los conjuntos de datos previamente a la modelación.

El objetivo del presente trabajo fue evaluar el efecto de la eliminación de los *activity cliffs* sobre la capacidad predictiva de modelos QSAR basados en algoritmos de aprendizaje automatizado. Con este propósito se diseñó e implementó un procedimiento para identificar los *activity cliffs*, y eliminar los más influyentes de los conjuntos de datos. Se utilizaron nueve algoritmos de aprendizaje automatizado en la modelación de los cinco conjuntos de datos seleccionados. Se evaluó el desempeño de los modelos QSAR obtenidos a partir de los conjuntos de datos "sin *activity cliffs*" respecto a los obtenidos para los conjuntos de datos originales.

Durante el proceso de evaluación se pudo comprobar que la eliminación de los *activity cliffs* no condujo a cambios estadísticamente significativos de la continuidad de las SAR. Sin embargo, si se apreciaron mejoras estadísticamente significativas en la modelabilidad de los conjuntos de entrenamiento; específicamente los procesados empleando el algoritmo que realiza agregación de las matrices de similitud por media geométrica. Por otra parte, eliminar los *activity cliffs* permitió mejoras estadísticamente significativas en el proceso de entrenamiento y validación de los modelos, no siendo así en la clasificación de los subconjuntos de validación externa, donde de manera general no hubo cambios estadísticamente significativos. No obstante, se mejoró la clasificación de la clase peor clasificada por los modelos obtenidos de los subconjuntos de entrenamiento originales. Este último resultado fue estadísticamente significativo para el algoritmo de eliminación de *activity cliffs* que no realiza fusión de matrices de similitud, lo que muestra una tendencia a balancear la clasificación.

ABSTRACT

The QSAR modeling is an example of cheminformatics tool which use has spread to different areas of development of society. The main assumption of the approximations in QSAR models is the continuity of space of Structure-Activity Relationships (SAR), the presence of *activity cliffs* may be affected it. Recent studies have shown the negative effects of the presence of the *activity cliffs* on the predictive ability of QSAR models. However, there are not reports showing if the removal of *activity cliffs* from a data sets is beneficial, detrimental or non-significant.

The goal of this study was to evaluate the effect of removing the *activity cliffs* on the predictive ability of QSAR models based on machine learning algorithms. For this purpose, we designed and implemented a procedure to identify the *activity cliffs*, and eliminate the most influential from data sets. We used nine machine-learning algorithms in modeling the five selected data sets. In addition, we evaluated the performance of QSAR models obtained from data sets "without *activity cliffs*" compared to those obtained for the original data sets.

During the evaluation process, we found that removal of the *activity cliffs* did not lead to statistically significant changes in the continuity of the SAR. However, we did noticed statistically significant improvements in modelability of training sets; specifically processed using the algorithm that performs aggregation of similarity matrices by geometric mean. Moreover, eliminating *activity cliffs* allowed statistically significant improvements in the training process and validation of the models, not the case in the classification of subsets of external validation, where generally there were no statistically significant changes. However, the classification of the worst class classified by the models achieved from training subsets, improved. The latter result was statistically significant for the removal of *activity cliffs* algorithm that does not perform fusion of similarity matrices, showing a tendency to balance the classification.

ÍNDICE

INTRODUCCIÓN

En los últimos años el empleo de herramientas quimioinformáticas se ha extendido a diferentes esferas del desarrollo de la sociedad. Estas han permitido una reducción considerable del consumo de recursos y de tiempo gracias a su capacidad de transformar datos en conocimiento, y con esto optimizar la toma de decisiones.

La modelación QSAR (por sus siglas en inglés, *Quantitative Structure-Activity Relationships*) es un ejemplo de herramienta quimioinformática cuyo uso se ha extendido. La industria farmacéutica constituye un área de investigación donde la aplicación de modelos QSAR en el filtrado de grandes bases de datos de moléculas ha sido fundamental en la optimización de compuestos líderes.

Todo modelo QSAR se basa en el principio de que compuestos similares, deberían presentar propiedades similares también. En esencia, un modelo QSAR correlaciona las estructuras moleculares con una propiedad dada. Puesto que las representaciones estructurales por si solas no brindan información utilizable para la modelación, estas son codificadas por una serie de atributos, comúnmente conocidos como descriptores moleculares.

A pesar de la ilimitada disponibilidad de descriptores moleculares y el incremento de la eficiencia de las técnicas de aprendizaje automatizado, su capacidad predictiva es aún limitada en la implementación de modelos QSAR. Todavía son significativos los errores de predicción de actividad entre moléculas similares, incluso en casos donde la predictividad global es elevada [2].

El principal supuesto de las aproximaciones en modelos QSAR es la continuidad del espacio de las Relaciones Estructura-Actividad (SAR por sus siglas en inglés, *Structure-Activity Relationships*). En este contexto, el paisaje estructura-actividad luce como colinas onduladas, y por ello, cambios graduales en la estructura conducirían a cambios graduales en la actividad. Sin embargo, descripciones cuantitativas de varios conjuntos de compuestos activos han mostrado que la mayoría de las SAR globales son heterogéneas por naturaleza [2]. Por lo que, sus paisajes estructura-actividad contienen tanto regiones ligeramente inclinadas como acantilados agudos.

La continuidad en las SAR brinda las bases fundamentales para el análisis QSAR y las resultantes predicciones de actividad de compuestos [2], mientras que la presencia

de discontinuidad queda fuera del dominio de aplicación del paradigma de QSAR, limitando la capacidad predictiva de estos modelos. Hasta el momento, no se reporta ningún trabajo que se haya dirigido a reducir la discontinuidad de las SAR sobre un conjunto de datos y consecuentemente restaurar tanto como sea posible el principio fundamental de los métodos QSAR y los basados en similitud.

En los inicios de la modelación QSAR los métodos de aprendizaje estadístico jugaron un rol fundamental. Sin embargo, en la actualidad los algoritmos de aprendizaje automatizado constituyen las herramientas más extensamente empleadas. Los dos propósitos fundamentales para los cuales el aprendizaje automatizado es aplicado en esta área son: la clasificación y la generalización de los datos. Estos métodos utilizan el conocimiento sobre las SAR para guiar el proceso a favor de producir clasificaciones y generalizaciones con un elevado significado conceptual [3].

Los métodos de clasificación del aprendizaje automatizado construyen los modelos sobre la base de generalidades presentes en los datos. Los acantilados de estructura-actividad (*activity cliffs*) representan excepciones de estas generalidades. Por lo que, la influencia contraproducente de un par de moléculas que forman acantilados de actividad es evidente.

Una posible solución podría ser la eliminación de tales compuestos “problemáticos” responsables de la discontinuidad de las SAR, y consecuentemente restaurar la continuidad de estas. El principal supuesto de esta solución es que el aprendizaje a partir de un conjunto de entrenamiento libre del ruido inducido por esos ejemplos “problemáticos”, debería producir modelos más predictivos. Sin embargo, el proceso de entrenamiento podría verse afectado por la pérdida de la información presente en los pares de compuestos que forman acantilados de actividad (*activity cliffs*). Desde este punto de vista, eliminar estos puntos podría perjudicar severamente las capacidades predictivas de los modelos.

Por todo lo antes expuesto, se plantea el siguiente **problema científico**:

La presencia de *activity cliffs* en los conjuntos de datos empleados en la modelación QSAR puede influir negativamente en la calidad de los modelos obtenidos.

Para dar respuesta al problema científico planteado, se propone el siguiente **objetivo general**:

Evaluar el efecto de la eliminación de los *activity cliffs* sobre la capacidad predictiva de modelos QSAR basados en algoritmos de aprendizaje automatizado.

Para alcanzar el objetivo planteado en la presente investigación se definieron los siguientes **objetivos específicos:**

- Analizar los conceptos y definiciones fundamentales relacionados con los *activity cliffs* y la modelación QSAR.
- Seleccionar las herramientas y los métodos para el pre-procesamiento de las estructuras moleculares.
- Diseñar e implementar un procedimiento para identificar los *activity cliffs* y eliminar los más influyentes.
- Seleccionar los conjuntos de datos y los métodos de aprendizaje automatizado para llevar a cabo la modelación.
- Evaluar el desempeño de los modelos QSAR obtenidos a partir de los conjuntos de datos "sin *activity cliffs*", respecto a los obtenidos para los conjuntos de datos originales.

Los resultados derivados de la ejecución de los objetivos permitirán evaluar la veracidad de la siguiente **hipótesis**:

"La eliminación de los *activity cliffs* traerá consigo una mejora significativa en la capacidad predictiva de los modelos QSAR basados en algoritmos de aprendizaje automatizado."

El resto del documento se encuentra estructurado de la siguiente manera:

En el Capítulo 1, titulado "Marco teórico", se exponen aspectos fundamentales referentes al concepto de *activity cliffs*. Se ilustran evidencias en la literatura sobre los efectos negativos de la presencia de estos sobre la modelación QSAR, así como la analogía con las ISM (por sus siglas en inglés, *Instances that Should be Misclassified*). Se hace énfasis en el pre-procesamiento estructural de los conjuntos de datos, así como los diferentes métodos reportados en la literatura que permiten una partición adecuada de estos. Por último, se caracterizan las familias de métodos de selección

de atributos y los principales algoritmos de aprendizaje automatizado que han sido empleados en la modelación QSAR.

En el Capítulo 2, titulado “Procedimientos y métodos”, se describen: el procedimiento propuesto para la modelación QSAR basada en la eliminación de los *activity cliffs*, los conjuntos de datos seleccionados para probar la hipótesis planteada y los detalles del pre-procesamiento de estos, así como, la descripción del proceso de modelación.

En el Capítulo 3, titulado “Resultados y discusión” se exponen y discuten los resultados de la experimentación realizada. Se analizan los resultados de la eliminación de los *activity cliffs* desde varios puntos de vista. Entre estos: la modelabilidad y continuidad de las Relaciones Estructura-Actividad de los subconjuntos de entrenamiento, y los resultados de la clasificación basados en nueve métodos de clasificación empleados.

Finalmente se presentan las conclusiones de la tesis, así como algunas recomendaciones que abren futuras líneas de investigación, se relaciona la bibliografía y se muestran algunos anexos.

CAPÍTULO 1. MARCO TEÓRICO.

En el presente capítulo se exponen los conceptos e ideas fundamentales alrededor del fenómeno de los *activity cliffs* desde dos puntos de vista, el de los químicos medicinales y el de los quimioinformáticos practicantes de la modelación QSAR. Además se abordan los aspectos más relevantes referentes al pre-procesamiento de los conjuntos de datos y un resumen de los principales algoritmos de aprendizaje automatizado de uso frecuente en la modelación QSAR.

1.1. Los *activity cliffs* en la modelación QSAR.

En esencia, un modelo QSAR consiste en correlacionar los aspectos estructurales de las moléculas con su actividad biológica. En esta aproximación, las estructuras son codificadas numéricamente mediante los descriptores moleculares. Luego estos descriptores son sometidos a un proceso de selección, donde se espera que queden los que posean mayor valor informativo. Por último, los descriptores seleccionados son empleados en la construcción de los modelos, en los cuales se expresará la actividad biológica (o propiedad modelada) como una función matemática de estos descriptores.

La modelación QSAR ha constituido una poderosa herramienta en el proceso de diseño de fármacos [4, 5]. En un mundo en el cual el desarrollo de nuevas tecnologías en las diferentes ramas de la ciencia ha acrecentado los volúmenes de datos disponibles para la extracción de conocimiento, la aplicación de modelos QSAR en el filtrado de grandes bases de datos de moléculas constituye un considerable ahorro de tiempo y recursos [6, 7].

A pesar de sus ventajas, todavía existen grandes limitaciones en el empleo de los modelos QSAR. Aun siguiendo cuidadosas metodologías o “buenas prácticas” reportadas en la literatura para la construcción y evaluación de modelos más fiables [1, 8], en ocasiones no se logra obtener la capacidad predictiva deseada. En su artículo “*On Outliers and Activity Cliffs - Why QSAR Often Disappoints*” [2], Gerald M. Maggiora alertó sobre las influencias negativas de diferentes factores en la obtención de modelos QSAR. Entre los factores mencionados, resaltaba la naturaleza de los paisajes de

estructura-actividad basados en una representación estructural dada, y a consecuencia de estos, la presencia de los *activity cliffs*.

Los *activity cliffs* son pares de compuestos que presentan elevada similitud estructural pero grandes diferencias en los valores de actividad biológica o propiedad modelada [9]. Para los químicos medicinales la existencia y aplicación de estos es bastante clara [10, 11]. Todo químico medicinal experimentado está continuamente buscando pares de moléculas con elevada similitud estructural pero valores de actividad muy diferentes [12, 13]. Sin embargo, esfuerzos enfocados hacia detectarlos en conjuntos de datos analizados empleando métodos computacionales, han generado la interrogante referente a si existen o son justamente resultado de los métodos computacionales empleados [14, 15]. En los últimos años ha sido considerable el número de publicaciones enfocadas en definir o aplicar el concepto de *activity cliffs* [13]. Sin embargo, los efectos adversos de estos en la aplicación de los modelos QSAR y de otras aproximaciones basadas en similitud han sido poco estudiados.

1.1.1. Los paisajes de estructura-actividad y los *activity cliffs*.

Las definiciones más aceptadas sobre los *activity cliffs* se basan en el concepto de *paisajes de estructura-actividad*. Estos “paisajes” y los *activity cliffs* son portadores de una enorme riqueza informativa en lo que a las Relaciones Estructura-Actividad (SAR, por sus siglas en inglés: *Structure-Activity Relationships*) respecta [16]. Un paisaje de estructura-actividad representa una hipersuperficie en un espacio químico biológicamente relevante, similar a los mapas geográficos, donde los valores de potencia de los compuestos son agregados como una tercera dimensión a una proyección bidimensional del espacio químico [17]. En los paisajes de estructura-actividad, regiones suaves son asociadas con SAR continuas y representan áreas donde cambios graduales en la estructura química inducen cambios moderados en la actividad biológica. En cambio, regiones rugosas son asociadas con SAR discontinuas, donde pequeñas modificaciones en la estructura química cambian drásticamente la actividad biológica [18]. Las formas extremas de discontinuidad de las SAR son denominadas *activity cliffs*, los cuales están formados por pares de compuestos de elevada similitud estructural y con grandes diferencias en potencia [16].

Mientras las SAR discontinuas y los *activity cliffs* constituyen la base para la optimización de compuestos líderes, paisajes de estructura-actividad con regiones suaves y SAR continuas son prerrequisitos para la aplicación exitosa de modelos QSAR y métodos basados en similitud o simplemente como herramientas predictivas [16]. Estas aproximaciones cuantitativas se basan en el Principio de Similitud-Propiedad (SPP, por sus siglas en inglés: *Similarity Property Principle*) [19], el cual establece que: moléculas similares deberían tener valores similares de sus propiedades (en este caso actividad), eso, asumiendo la presencia de las SAR continuas. En contraste, en regiones rugosas y SAR discontinuas, la aplicación de métodos basados en similitud carece de significado [16].

En la literatura han sido introducidas varias funciones de análisis numérico para cuantificar la discontinuidad de las SAR e identificar los *activity cliffs*. Entre estas resaltan: el Índice de Relaciones Estructura-Actividad (SARI, por sus siglas en inglés: *Structure-Activity Relationships Index*) [20] y el Índice de Paisajes de Estructura-Actividad (SALI, por sus siglas en inglés: *Structure-Activity Landscape Index*) [21].

La aproximación SALI es particularmente adecuada para detectar *activity cliffs* en conjuntos de datos. Sin embargo la magnitud de estos no es determinada por esta métrica dado que sus valores son comparados en una escala relativa. Como consecuencia de esto, los *activity cliffs* detectados a un cierto valor de corte podrían ser irrelevantes (*cliffs superficiales o pseudo-cliffs*).

Respecto a lo anterior, Stumpfe y Bajorath [12], resaltaron la necesidad de utilizar un criterio discreto para definir *activity cliffs* incluyendo el criterio de similitud aplicado, las magnitudes de potencia y las diferencias de dichas magnitudes para pares de compuestos. Estos expertos recomendaron considerar un par de compuestos *activity cliffs* solo si: (i) un criterio de similitud preestablecido se cumple; (ii) uno de los dos compuestos presenta potencia en el orden de los nanomol (**nM**) y; (iii) al menos uno es 100 veces más potente que el otro.

Aun así, estas definiciones discretas sobre los *activity cliffs* tienen importantes limitaciones. Por mencionar algunas: el tipo (*IC50 o* K_i) y la calidad de las mediciones experimentales, las representaciones moleculares, y las métricas de similitud

empleadas; pueden influir considerablemente en el análisis y detección de los *activity cliffs* [16]. Por lo tanto, *activity cliffs* identificados en un espacio químico y biológico dado pueden no conservarse en un espacio de referencia diferente [22].

Los químicos medicinales cuestionan los *activity cliffs* definidos utilizando aproximaciones de similitud, debido a su limitada interpretabilidad química [12, 23]. Para tratar ese problema, Hu y colaboradores en [24] utilizaron el formalismo de correspondencia de pares moleculares (MMP, por sus siglas en inglés: *Matched Molecular Pairs*). Un MMP se define como un par de compuestos que solamente difieren en un sitio tal como un anillo o un grupo radical (representado en la química orgánica con la letra R). Por lo tanto, para clasificar un par de moléculas como un MMP-*cliff*, la diferencia de potencia requerida sigue siendo esencialmente la misma, como las que se aplican en las definiciones basadas en similitud. De hecho, la diferencia en tamaño entre los fragmentos intercambiados y su tamaño se restringen a un predefinido número de átomos diferentes del hidrógeno que garantizan el nivel de similitud estructural esperado para un *activity cliff* [24].

1.1.2. *Activity cliffs*: consideraciones.

Como se había argumentado anteriormente, los *activity cliffs* pueden ser artilugios debido a una descripción estructural que no es relevante para un conjunto de datos específico en el contexto del problema biológico en cuestión [14]. Una descripción ilustrativa de este punto de vista es ofrecida por Horvath en un extenso artículo que familiariza a los químicos experimentales con QSAR [15]. Este autor menciona que dos moléculas codificadas por un conjunto de descriptores que no sea capaz de capturar las diferencias estructurales entre estas, las mostrará como muy cercanas en el espacio químico. De esta manera, no podrían ser explicadas grandes diferencias de actividad entre dichas moléculas, con lo que se generaría un *activity cliff*. Por lo tanto, no sería una tarea sencilla diferenciar directamente dicha situación de la presencia de verdaderos *activity cliffs*.

Dado que los *activity cliffs* se podrían producir como consecuencia de la inadecuada selección de descriptores, es tentador tratar de removerlos de un conjunto de datos mediante la búsqueda de descriptores apropiados que suavicen los *paisajes de*

estructura-actividad. Sin embargo, identificar dichos descriptores no ocurre directamente puesto que esa aproximación involucra métodos efectivos de selección de rasgos o de mapeo, y/o la selección de un adecuado criterio de evaluación [16].

Herramientas quimioinformáticas en la detección de los *activity cliffs*.

Encontrar los descriptores "apropiados" para suavizar los paisajes de estructura-actividad es un problema dependiente del conjunto de datos. Por consiguiente, la búsqueda de una representación molecular universal no constituye una forma realista de enfrentar el problema. Una solución potencial para encontrar descriptores adecuados para un conjunto de datos dado, podría ser la integración de métodos de selección de atributos con funciones de análisis numérico para cuantificar la discontinuidad de los espacios de las SAR tales como: SARI [20] y/o SALI [21]. De esta manera, el subconjunto de descriptores relevantes es buscado tratando de maximizar/minimizar la continuidad/discontinuidad del espacio de las SAR presente en el conjunto de entrenamiento.

La razón fundamental del uso de funciones de análisis numérico para cuantificar la discontinuidad de las SAR como funciones objetivo para la selección de atributos radica en tomar ventaja de dos factores bien conocidos en el análisis de similitud:

(i) la dependencia de los paisajes de actividad respecto a la representación y;
(ii) las características distintivas de los descriptores escalares comúnmente empleados en la modelación de QSAR para indirectamente representar las interacciones ligando-receptor.

Por esta vía, se dirige la búsqueda a través de la selección de características estructurales que permitan obtener espacios de las SAR continuos. Aplicaciones potenciales de QSAR y los análisis basados en similitud sobre tales espacios continuos de referencia deberían ser más significativos y predictivos que aquellos basados en espacios heterogéneos de referencia [16].

Un ejemplo del potencial que ofrecen las funciones de análisis como criterio de evaluación para cuantificar la discontinuidad de los espacios de las SAR, lo constituye un importante trabajo de Guha y Van Drie [25]. Este fue dirigido a evaluar la capacidad de capturar los paisajes de estructura actividad por un protocolo de modelación

mediante las curvas de SALI. A diferencia de aplicaciones previas de la aproximación SALI, en dicho trabajo es demostrado que esas curvas SALI pueden ser aplicadas para medir cuantitativamente la capacidad de un modelo dado de capturar los *activity cliffs* que son inherentes en un espacio de las SAR.

Alternativamente a lo antes mencionado, se puede dividir el conjunto de entrenamiento en clústeres estructurales para construir modelos locales o multidominios más confiables.

Otra solución en la búsqueda del mejor conjunto de descriptores es el desarrollo de representaciones moleculares mejoradas. Dichas representaciones serían capaces de codificar información relevante y tendrían la habilidad de generar espacios de descriptores dinámicos o estáticos robustos. Una vez que un espacio de descriptores robusto ha sido encontrado para un conjunto de datos específico, este puede ser considerado como dinámico si prueba ser lo suficientemente robusto como para explicar las SAR de un nuevo conjunto de datos aumentado con nuevos datos cribados. De otra manera, el espacio de descriptores encontrados sería robusto pero estático. La introducción de ISIDA-PLF (por sus siglas en inglés: *In SIlico Design and Data Analysis Property-Labeled Fragment Descriptors*) es un buen ejemplo apuntando en esa dirección [26].

Similitud molecular global vs local.

Los químicos medicinales y los quimioinformáticos ocasionalmente, tienen opiniones divergentes sobre la utilidad de los *activity cliffs*. Estas divergencias están referidas a si son características deseables o no en un conjunto de datos, y están arraigadas sobre la aplicación de diferentes conceptos de similitud molecular y los respectivos métodos de similitud molecular requeridos para cada tarea. Los quimioinformáticos, y más específicamente, los practicantes de la modelación de QSAR y de los métodos basados en farmacóforos están acostumbrados a trabajar con similitudes `locales´. Estas similitudes se expresan comúnmente mediante descriptores moleculares escalares los cuales codifican, por ejemplo, aspectos topológicos, constitucionales o funcionales de la estructura molecular. Por otra parte, químicos medicinales suelen vislumbrar la similitud molecular empleando representaciones moleculares globales u

holísticas [16], las cuales son frecuentemente codificadas usando *fingerprints* tales como las claves MACCS [27] o los *fingerprints* de conectividad extendida (ECFP, por sus siglas en inglés, *Extended Connectivity Fingerprints*) [28]). Probablemente, esta es la razón por la que químicos medicinales aceptan fácilmente la existencia de *activity cliffs* mientras que no todo quimioinformático lo hace. Este razonamiento ha sido discutido en una reciente publicación de Stumpfe y Bajorath [29]. Estos autores, esencialmente afirman que cualquier método de similitud holístico debe reconocer dos análogos cercanamente relacionados por ser similares, incluso si uno es activo y el otro no lo es, debido a la violación crítica de las interacciones receptor-ligando. Esto significa que métodos basados en el SPP son realmente ligando-céntricos y no tienen en cuenta el criterio de interacción. Aunque ambas vistas de similitud (global y local) son empleadas en búsquedas basadas en similitud, las representaciones moleculares holísticas son las más frecuentemente utilizadas hasta el momento [16].

Dependencia de la representación molecular.

Es aceptado que la representación molecular es el parámetro más crítico para definir *activity cliffs*. Como se discutió anteriormente, aproximaciones para tratar esta dependencia pueden consistir en utilizar MMPs o sustituciones discretas de grupos funcionales R alrededor de un esqueleto o plantilla.

Otra aproximación propuesta por Medina-Franco y colaboradores en [22], consiste en utilizar múltiples representaciones para derivar a un modelo de paisajes de estructura-actividad de consenso. Usando esta estrategia, la aproximación SALI [21] fue extendida para computar valores consenso obtenidos sobre múltiples, diversas y ortogonales representaciones 2D y 3D. Esta aproximación conduce a la identificación de pares de moléculas concurrentemente clasificados como *activity cliffs* por múltiples representaciones que codifican información determinante topológica, conformacional y basada en farmacóforos. Una estrategia similar de fusión de datos combinando múltiples representaciones 2D y 3D puede ser extendida a otros métodos de paisajes de estructura-actividad tales como la aproximación SARI propuesta por Bajorath en [20] para derivar a otro consenso.

1.1.3. Los *activity cliffs* en la química medicinal.

En la química medicinal, el concepto *activity cliff* puede ser convenientemente empleado en la optimización de compuestos líderes. En esta área de investigación es altamente relevante identificar pequeñas modificaciones estructurales asociadas con cambios en potencia significativos. Un estudio dirigido a capturar la progresión de las SAR teniendo en cuenta varios caminos, unos independientes y otros dependientes de los *activity cliffs* [30], evidencia el potencial de explotar este concepto en proyectos de optimización de compuestos líderes. El estudio mostró que los compuestos más potentes fueron identificados mediante los caminos dependientes de los *activity cliffs* en comparación con los caminos independientes de estos, lo que muestra evidencias de las ventajas de explotar la información latente en estos.

Adicionalmente, el concepto de *activity cliffs* puede ser implementado sobre diferentes tipos de análisis computacionales basados en similitud puesto que se basa en dos tipos de relaciones de similitud: similitudes estructurales y de potencia. Sin embargo, en aplicaciones computacionales específicas que se basan en SAR continuas, los *activity cliffs* imponen una limitación al representar excepciones del SPP.

1.1.4. Analogía entre los *activity cliffs, outliers*, ruido e instancias que podrían ser mal clasificadas.

Encontrar un paralelismo para el fenómeno de los *activity cliffs* en el área del aprendizaje automatizado, y establecer un fundamento para el posible efecto negativo de estos sobre las capacidades predictivas de los modelos no son tareas triviales. Sin embargo, Smith y Martínez [31] introdujeron PRISM (por sus siglas en inglés, *Preprocessing Instances that Should be Misclassified*), un nuevo método de filtrado que identifica “instancias que deberían ser mal clasificadas” (ISM, por sus siglas en inglés, *Instances that Should be Misclassified*) empleando heurísticas que predicen cuan probable es que una instancia será mal clasificada. El concepto básico detrás de PRISM, ISM, parece ser el análogo más cercano al concepto *activity cliff* en el área del aprendizaje automatizado. Una instancia es reconocida como ISM si, sujeto a la información provista en el conjunto de entrenamiento, la etiqueta asignada por el algoritmo de aprendizaje a dicha instancia es probablemente la más correcta, incluso

si su etiqueta actual es diferente. A diferencia de los *outliers* tradicionales y los ruidos de clase, las ISM exhiben un alto grado de solapamiento de las clases. Por lo que, una ISM está cerca en el espacio de características a otras instancias de diferente clase [16].

Pese a que las ISM han sido definidas en un contexto de clasificación, una buena analogía puede ser establecida entre estas y los *activity cliffs* [16]. En el aprendizaje automatizado, las ISM son instancias similares con diferentes etiquetas a otras instancias vecinas en una región del espacio de características; en la modelación de paisajes de estructura-actividad, los *activity cliffs* representan moléculas cercanas en el espacio químico con gran diferencia en actividad.

En este contexto, los *activity cliffs* pueden ser vistos como casos especiales de ISM. La mayor diferencia radica sobre la consideración explícita del grado de similitud entre instancias. Además del solapamiento de clases, un alto grado de similitud es requerido para etiquetar un par de instancias con clases opuestas como un *activity cliff* [16].

El algoritmo PRISM propone la eliminación de las ISM previo a la modelación, a diferencia de la modelación QSAR tradicional, donde los *outliers* frecuentemente son removidos después de que los modelos han sido construidos. Por otra parte, los métodos de detección de *outliers* aspiran a encontrar anomalías en los datos, mientras que los métodos de reducción de ruido intentan identificar y remover instancias mal etiquetadas. Sin embargo, la detección de *outliers* y ruido, y la eliminación de estos es difícil porque actualmente no existe una definición universal de que es un *outlier* o de si una instancia es ruido o no [16]. Así, las ISM deben ser diferenciadas de los *outliers* y el ruido recurriendo a sus características básicas como se consideran por sus respectivos métodos de detección. Cruz-Monteagudo y colaboradores [16] compararon estos tres conceptos en términos de aprendizaje automatizado y espacios de estructura-actividad, la cual se muestra, en la Tabla 1.1.

Principales aspectos que caracterizan los *outliers*, el ruido y las ISM, considerando sus respectivos métodos de detección.

	Valor Extremo	Etiquetado Incorrecto	Solapamiento de clases
Outlier	+	-	*
Ruido	-	+	*
ISM	+/-	-	+

Para identificar cada tipo de excepción de datos, los métodos correspondientes toman en consideración la presencia (+) o la ausencia (-) de: valores extremos en los datos, etiquetado incorrecto de instancias o el solapamiento entre clases. El último aspecto no es tenido en cuenta (*) por los métodos de detección de *outliers* y ruido. Basándose en estas tres características:

(i) los *outliers* pueden ser definidos como excepciones en los datos representadas por valores extremos en los espacios de descriptores y/o propiedades no atribuibles al etiquetado incorrecto de las clases.
(ii) el ruido de clase representa excepciones en los datos atribuibles al etiquetado incorrecto de las clases y;
(iii) las ISM son excepciones en los datos no atribuibles al etiquetado incorrecto de las clases caracterizadas por un alto grado de solapamiento entre las clases.

Tabla 1.1. Comparación entre outliers, ruido e ISMs realizada por Cruz-Monteagudo y colaboradores [16] en términos de aprendizaje automatizado y *espacios de estructura-actividad*.

Smith y Martínez [31] compararon PRISM con tres métodos de detección de *outliers* y una técnica de reducción de ruido existentes, utilizando 48 conjuntos de datos y 9 algoritmos de aprendizaje. Eliminando instancias identificadas por PRISM previamente al entrenamiento, consiguieron el más alto desempeño en la clasificación comparado con el entrenamiento de los algoritmos de aprendizaje automatizado con los conjuntos de datos originales así como con los *outliers* removidos por los otros métodos. En vez de centrarse en clasificar correctamente las ISM, y arbitrariamente ajustar el límite de la clasificación, remover las ISMs previo al entrenamiento permitió a los algoritmos de aprendizaje automatizado enfocarse sobre las instancias que podían ser correctamente clasificadas. En otras palabras, eliminar las ISMs permitió al proceso de aprendizaje enfocarse sobre los patrones observados en vez de memorizar muestras que no los seguían. Esta aproximación puede ser extendida a la modelación de paisajes de estructura-actividad eliminando *activity cliffs* para suavizar dichos paisajes.

En adición al algoritmo PRISM, otros métodos de detección de *outliers* y/o ruido han sido introducidos recientemente, demostrando su habilidad para mejorar la exactitud de los modelos de aprendizaje automatizado [32, 33]. Por ejemplo, Byeon introdujo una técnica novedosa para mejorar la calidad de los datos de entrenamiento con una variable dependiente del ruido para la clasificación binaria [32]. La aproximación denominada GAPS (por sus siglas en inglés, *Genetic Algorithm and Prototype Selection*) usa un algoritmo genético para crear un conjunto de posibles instancias ruidosas y una selección de prototipo para identificar el conjunto actual de estas instancias. Los autores compararon el desempeño de GAPS con métodos de filtrado implementados en WEKA [34] en dos conjuntos de datos sintéticos provenientes del repositorio de la "*University of California-Irvine"* (UCI) (http://archive.ics.uci.edu/ml/datasets.html). Mientras los conjuntos de datos mejorados con WEKA mostraron niveles similares de exactitud en tareas de clasificación sobre conjuntos de datos ruidosos, la aproximación GAPS, como promedio, redujo el error en clasificación sobre estos en un 26% para todos los niveles de ruido. Más recientemente, basado en el factor de ruido de la línea límite, Yang y Gao [33] aplicaron técnicas de limpieza de datos para remover ruido de la línea límite de clasificación. En ese trabajo fueron comparados tres métodos de sub-muestreo para seleccionar ejemplos representativos de la clase mayoritaria y eliminar las muestras distantes las cuales son innecesarias para obtener una frontera de clasificación. Los resultados experimentales sobre los conjuntos de datos procesados mostraron que el método propuesto puede efectivamente mejorar la exactitud en clasificación de clases minoritarias mientras que logró mejor desempeño en la clasificación global.

Ya que eliminar instancias problemáticas es una práctica bien justificada en el aprendizaje automatizado para mejorar la exactitud de las predicciones de modelos [31-33, 35], es razonable la idea de remover los *activity cliffs* para obtener modelos QSAR más predictivos. Obviamente, el inconveniente de este procedimiento es la inevitable pérdida de información potencialmente crítica sobre las SAR. Sin embargo, este es el precio que ha de ser pagado en favor de la capacidad predictiva de los modelos basados en aprendizaje automatizado.

1.1.5. *Activity cliffs* y QSAR: un posible tratamiento.

A pesar del hecho de que es bastante aceptada la idea de que la presencia los *activity cliffs* va en detrimento de la modelación QSAR, no se reportan en la literatura trabajos dirigidos a reducir la discontinuidad de las SAR en un conjunto de datos mediante la remoción de los *activity cliffs* [16].

De manera similar, no existen reportes mostrando que remover los *activity cliffs* de conjuntos de datos sea beneficioso, perjudicial o no significativo en la obtención de modelos de las SAR [16]. Lejos de esto, la mayoría de las aplicaciones quimioinformáticas se centran en la descripción de las SAR y la identificación de los *activity cliffs* [13, 36]. Como se mencionó anteriormente, para propósitos prácticos (por ejemplo, para químicos medicinales) los *activity cliffs* pueden proveer información vital para entender las SAR y guiar los esfuerzos en la optimización de compuestos líderes [12, 13].

Restauración de la continuidad de los espacios de las Relaciones Estructura-Actividad identificando y removiendo los generadores de *activity cliffs*.

El concepto más cercano a una ISM en la modelación de *paisajes de estructura-actividad* es el de ACG (por sus siglas en inglés, *Activity Cliffs Generators*), el cual se puede definir como una estructura molecular que presenta una alta probabilidad de formar *activity cliffs* con moléculas probadas en el mismo ensayo biológico [37]. En analogía con las ISMs, se pueden identificar y remover los ACG.

Idealmente, los conceptos *activity cliffs* por consenso [22] o MMP-*cliffs* [24] deberían ser aplicados en la determinación de ACG. Los ACG identificados por consenso usando varias representaciones, deberían comportarse como tales, independientemente del espacio de referencia empleado o, al menos, para la mayoría de los posibles espacios de referencia. Como se discutió previamente, los *fingerprints* son representaciones ligando-céntricas basadas en la definición de similitud global y de esta manera, las más idóneas para codificar las similitudes globales de las estructuras. Por otra parte, los descriptores moleculares codifican aspectos locales de la estructura molecular y pueden tomar en cuenta indirectamente información sobre

las interacciones ligando-receptor, y así, quizás ser más apropiados para la modelación eficiente de las SAR [29].

Una vez removidos los ACG previamente identificados, el propósito es que el conjunto de entrenamiento original cumpla con el supuesto hecho por el algoritmo a emplear para la construcción del modelo, lo cual es además la principal premisa de la modelación QSAR. Adicionalmente, los procedimientos de curación [1] y balanceo [38, 39] deberían ser también aplicados para corresponderse con los objetivos del paradigma de la modelación de QSAR [2] y del modelo de aprendizaje automatizado [40].

En el proceso de modelación de QSAR clásico, la eliminación de *outliers* es un procedimiento a *posteriori* dirigido a mejorar el desempeño de los modelos, y depende además del espacio de referencia. En contraste, los procedimientos de detección y eliminación de los ACG son realizados para optimizar el conjunto de entrenamiento para la modelación basada en aprendizaje automatizado previamente a cualquier esfuerzo de modelación y también es dependiente del espacio de referencia [16].

La esencia de esta solución es remover del proceso de entrenamiento esos compuestos responsables de la discontinuidad de las SAR, y consecuentemente restaurar la continuidad de estas requerida para derivar en modelos QSAR más confiables y predictivos.

1.2. Pre-procesamiento de los conjuntos de datos. “Buenas prácticas”.

En todo proceso de modelación, el pre-procesamiento de los datos desempeña un rol fundamental. Un modelo, por definición, es una representación simplificada de la realidad, con un dominio de aplicación limitado y sujeto a ciertos márgenes de error. En base a estas posibles limitaciones, la idea de construir modelos basados en datos imprecisos y ruidosos no constituye un camino razonable para brindar soluciones potenciales a problemas de la vida real. En este contexto, se hace necesario definir metodologías o caminos a seguir para adecuar los datos a los requerimientos de los métodos de modelación existentes.

La modelación QSAR tiene sus particularidades que la hacen muy sensible a esta etapa del proceso de extracción del conocimiento (KDD por sus siglas en inglés:

Knowledge Data Discovery). Los datos de partida están constituidos principalmente por una representación estructural y por valores de actividad biológica (variable dependiente) para cada compuesto/instancia en específico. En el caso de las representaciones estructurales, frecuentemente estas pueden estar en código *smiles* (representación de la estructura química en forma de texto plano) o en alguna otra representación en forma de grafo, las cuales por si solas no tienen sentido matemático alguno para la modelación, sin embargo, desde el punto de vista químico sí.

Debido a la gran variedad de posibles representaciones para un compuesto dado y que a partir de estas se calculan los descriptores moleculares (variables independientes o atributos), se hace necesario un meticuloso proceso de curación de los conjuntos de datos. Las formas de representación estructural deben ser estandarizadas, así como eliminadas las diversas formas de ruido en dichas representaciones y removidos los duplicados estructurales presentes.

Una vez realizado el proceso anterior, se deben tener en cuenta las dimensiones y el balanceo de los conjuntos de datos, ya que la mayoría de los métodos de aprendizaje automatizado son sensibles a estos dos aspectos. Por otra parte, obtener los conjuntos de entrenamiento y prueba a partir del conjunto de datos original es otro asunto de vital importancia. No es trivial particionar un conjunto de datos de partida para obtener subconjuntos de entrenamiento y prueba representativos del espacio químico en cuestión. Por último, una vez calculados los valores de los descriptores moleculares seleccionados para la modelación, debe ser llevada a cabo la selección de los atributos más relevantes.

1.2.1. Proceso de curación de los conjuntos de datos.

Este no es un proceso de simple ejecución, por el contrario, consta de varias etapas, de la aplicación de herramientas muy avanzadas y específicas para cada una de ellas, así como del criterio entrenado de un especialista.

Una de las asunciones más importantes de cualquier estudio basado en QSAR es la exactitud de los datos generados por los estudios experimentales. La presencia de datos imprecisos o incorrectos en los conjuntos de modelación es considerada la mayor preocupación en la construcción de modelos computacionales. Esto es

particularmente importante donde los valores de la actividad están esparcidos o la variación de potencia es muy limitada [41]. Varios autores han hecho especial énfasis sobre esta problemática y en la importancia de curar los datos previo a la modelación [1, 8, 42]. Entre estos resalta el artículo de Fourches y colaboradores [1], donde se propone una detallada metodología a seguir para curar conjuntos de datos basada en "buenas prácticas" así como los software disponibles para realizar dicha tarea en cada una de sus etapas. Por su relevancia en el tema, a continuación se detallarán las etapas de dicha metodología.

Eliminación de compuestos inorgánicos y mezclas.

La mayoría del software disponible para la modelación de QSAR no es capaz de extraer información de las estructuras de moléculas inorgánicas. De hecho, gran parte de los descriptores moleculares sólo pueden ser computados para compuestos orgánicos, lo cual constituye una limitación obvia de estas herramientas. Generalmente la fracción de compuestos inorgánicos en la mayoría de los conjuntos de datos disponibles, y especialmente aquellos con relevancia para el descubrimiento de fármacos es muy pequeña. No obstante algunos conjuntos de datos generados automáticamente con la ayuda de herramientas de minería de texto que extraen datos de la literatura o fuentes electrónicas, pueden contener un número significativo de compuestos inorgánicos de conocido efecto biológico. Por lo anterior, es aconsejable remover todos los compuestos inorgánicos antes de calcular los descriptores moleculares [1].

Para la eliminación de esta familia de compuestos, una alternativa podría ser calcular la fórmula empírica de cada compuesto, y luego remover todos los que en cuya composición no aparezca el carbono. Para esto existen varias opciones, por ejemplo, se podría emplear el programa *cxcalc* [43] de *ChemAxon* o utilizar el *pluging JChem for Excel* [44] de *ChemAxon* para procesar las estructuras utilizando la función *JCAtomCount* (la cual calcula la cantidad de átomos de un elemento dado en la fórmula química) implementada en dicho software.

Por otra parte, el tratamiento de mezclas no es una tarea tan sencilla como pueda parecer. En estos casos se suele proceder reteniendo los compuestos con la mayor

masa molecular o la mayor cantidad de átomos, lo cual no necesariamente deba ser la mejor solución al problema. Realmente la mejor opción es eliminar esos casos antes de calcular los descriptores moleculares [1]. Sin embargo, en algunos casos es razonable creer que la actividad biológica determinada experimentalmente puede ser causada por la molécula de mayor talla y no por la mezcla en sí. Estas soluciones son admisibles solamente en el caso de mezclas formadas por moléculas orgánicas de tallas relativamente grandes y pequeñas moléculas inorgánicas tales como: agua, clorhidratos, etc. [1].

En estos casos se pueden presentar diferentes situaciones:

1) Todos los compuestos en la mezcla son (o parecen ser) idénticos, tal puede ser el caso de mezclas racémicas utilizando representaciones bidimensionales de las moléculas. En este caso, solo una molécula puede ser retenida, y las otras simplemente eliminadas. Es necesario aclarar que este tratamiento solo es válido en casos de estudios QSAR bidimensionales cuando la mezcla racémica posee la misma actividad biológica (propiedad) que sus correspondientes enantiómeros.
2) La mezcla contiene un compuesto orgánico de gran talla y otros compuestos de poca talla, que pueden ser inorgánicos u orgánicos. Como se mencionó anteriormente, es necesario valorar, y solamente en determinados casos retener el compuestos de mayor talla.
3) Se tiene una mezcla consistente en varios compuestos orgánicos de masa molecular similar: estos son los casos más complicados y usualmente, se recomienda la eliminación de estos casos, a menos que se conozca el ingrediente activo [1]. En estas situaciones es necesario la intervención directa de un especialista.

Para los casos más simples se puede emplear el *Standardizer* [45] de *ChemAxon*, cuyo uso es muy simple e intuitivo. En este software se pueden seleccionar diferentes variantes de eliminación de solventes y pequeñas moléculas, las cuales se ponen en cola y después se procesan una a la vez para cada caso.

Conversión estructural y eliminación de sales.

Primeramente se convierten las representaciones que están en código *smiles* a grafos moleculares bidimensionales. Una herramienta bastante confiable en esta tarea es el paquete *Marvin* [46] de *ChemAxon*, aunque también se puede utilizar el mencionado *plugin JChem for Excel* [44].

Una vez realizada la conversión, se continúa con la eliminación de las sales. Muchos fármacos pueden ser encontrados en forma de sales y debido a que las propiedades de estas pueden ser muy diferentes de sus correspondientes moléculas neutras, es preferible eliminarlas. La eliminación de algunos tipos de contraiones metálicos, así como la neutralización de carbocationes y/o carbaniones resultantes, es otra práctica que en ocasiones puede ser aceptable. De hecho, similar a los compuestos inorgánicos, las sales no son procesadas por la mayoría del software que calcula descriptores, y su presencia puede generar numerosos errores en el cálculo de descriptores. La neutralización de las moléculas orgánicas cargadas es más cuestionable. Es bastante raro el hecho de conocer precisamente las condiciones experimentales bajo las cuales los compuestos han sido probados o las condiciones físico-químicas dentro de las células en las cuales un compuesto es activo. En los casos, cuando el pH de la solución y su composición exacta son conocidos, puede ser posible evaluar si un compuesto debería estar cargado [1]. Para realizar estas predicciones se puede emplear el programa *cxcalc* [43]. Cuando no se pueden realizar estimaciones confiables de las cargas de las moléculas, o si los descriptores a emplear son insensibles a la carga, es simplemente recomendado neutralizar todas las estructuras de los compuestos. Estas tareas pueden ser realizadas completamente empleando el *Standardizer* [45] o el OpenBabel [47], con los cuales se pueden identificar sales, eliminar contraiones, y luego neutralizar los compuestos orgánicos restantes.

Normalización de quimiotipos específicos y tratamiento de formas tautoméricas.

Es frecuente encontrar diferentes representaciones del mismo grupo funcional en un mismo conjunto de datos. Esto puede conducir a un gran número de inconsistencias, puesto que para dos representaciones diferentes de la misma molécula se obtendrían

diferentes valores de algunos de los descriptores calculados, y por lo tanto, estas no serían reconocidas como idénticas. Es importante representar cada grupo funcional de la misma forma, para evitar este tipo de problemas. Un ejemplo de esto se puede observar en la Figura 1.1, donde Fourches y colaboradores en [1] mostraron cinco representaciones diferentes de la misma molécula.

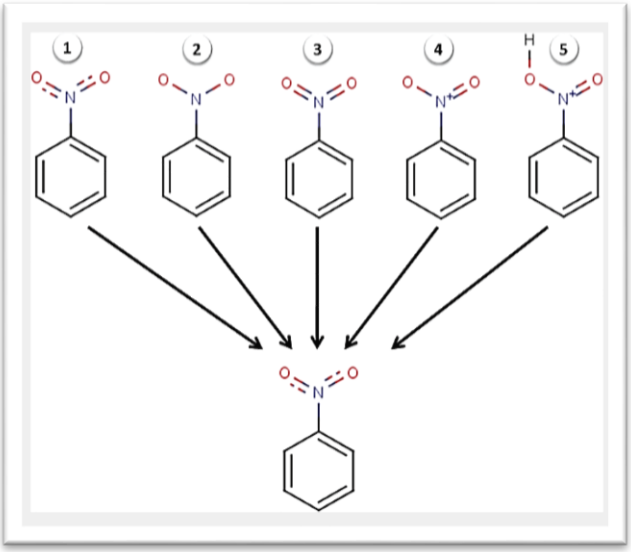

Figura 1.1. Cinco formas diferentes de representar el grupo nitro que conducen a cinco formas diferentes de representar la misma molécula.[1]

Obviamente esta tarea es prácticamente imposible realizarla manualmente, para esto se puede emplear el *Standardizer* [45]. Este software permite crear un fichero de configuración de formato *XML* con todas las posibles e imaginables transformaciones, las cuales se pueden llevar a cabo después de forma automática.

Eliminación de duplicados.

En las bases de datos de moléculas es frecuente encontrar casos duplicados, incluso con diferentes valores de actividad biológica, o estereoisómeros cuyas representaciones bidimensionales no reflejan sus diferencias en cuanto a centros quirales se trata. Esto constituye un serio problema en la modelación de QSAR, hasta en casos donde es reducido el porciento de duplicados estructurales puede verse considerablemente afectada la capacidad predictiva de los modelos [1]. Los duplicados pueden afectar la frecuencia observada de determinados quimiotipos en el conjunto

de datos, y de esta manera afectar el aprendizaje de los modelos a obtener. Por lo tanto, la eliminación de los duplicados estructurales constituye una de las etapas cruciales en el proceso de curación de los conjuntos de datos.

Para llevar a cabo esta tarea, primero es necesario identificarlos. En caso de tener las estructuras representadas en código *smiles*, se hace necesario que estas representaciones sean las *smiles* canónicas, de lo contrario podrían tenerse diferentes representaciones de la misma molécula sin percatarse de dicha coincidencia [1]. Si por el contrario se tiene otra representación (por ejemplo: bidimensional como en la Figura 1.1), entonces es recomendable emplear el software EdiSDF [48].

Antes de eliminar los duplicados se deben tener en cuenta los valores de actividad biológica de cada uno de estos. Si ambos casos coinciden en sus valores de actividad, entonces se elimina cualquiera de los dos, dejando el otro. De lo contrario, es más complicado el análisis a realizar. Entre las posibles situaciones tenemos:

(i) Error humano en todas sus variantes (ejemplo: error al pasar los datos a la base de datos, error en las mediciones, etc…).
(ii) Las mediciones se realizaron empleando variantes de los protocolos, o en diferentes laboratorios con diferentes condiciones experimentales.
(iii) Ambas mediciones son correctas y la diferencia radica en que son diferentes sales del mismo fármaco, cuyos contraiones fueron removidos en pasos anteriores o el compuesto neutro y una sal.
(iv) Ambas mediciones son correctas y la diferencia radica en que son estereoisómeros que difieren en sus valores de actividad biológica.

En los casos 1 y 2 se recomienda investigar la fuente de las mediciones en busca de mayor información sobre los datos. Hay que señalar que todas las mediciones de actividad en el conjunto de datos deben ser realizadas en las mismas condiciones o bajo los mismos protocolos, e idealmente en el mismo laboratorio (de ser posible).

En el caso 3 se recomienda quedarse con el compuesto neutro, de estar presente, o determinar la media de las actividades de las sales, solo si dichos valores son similares, de no ser así, deben ser eliminadas.

En el caso 4 pueden tenerse en cuenta dos variantes. La primera es retener el isómero más activo y eliminar el otro, y la segunda sería eliminarlos a ambos.

Chequeo manual.

Esta es la etapa final del proceso de curación y la que requiere de mayor cuidado. Implica, de ser posible, la inspección de todas las estructuras en el conjunto de datos, o al menos una muestra representativa de estas [1]. Es posible que aún después del empleo de poderosas herramientas especialmente diseñadas para las tareas anteriormente mencionadas (normalización de quimiotipos específicos, eliminación de duplicados, etc…) permanezcan errores considerables en las estructuras. Por el momento es imposible suplantar por completo el papel de un especialista en estas tareas. Todo software hace lo que se le programa para hacer, y si obviamos alguna posibilidad en el proceso de estandarización de las estructuras, esto podría ser fuente de errores en estas. En la Figura 1.2 se muestran otras variantes de las estructuras representadas en la Figura 1.1, que pudieran ser pasadas por alto.

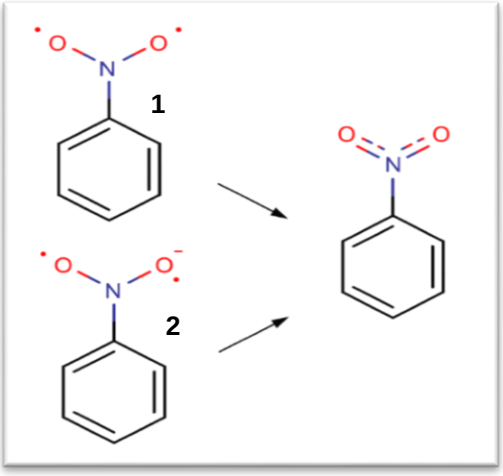

Figura 1.2. Dos variantes de estructuras no tenidas en cuenta en la **Figura 1.1**, que podrían ser encontradas en conjuntos de datos y que constituyen probables fuentes de error a la hora de calcular descriptores moleculares.

Las estructuras etiquetadas en la Figura 1.2 constituyen formas erróneas de representación de la estructura del nitrobenceno. Ambas, sumadas a las de la Figura 1.1, constituyen ejemplos del relativamente grande número de transformaciones que pueden ser necesarias para estandarizar las estructuras presentes en un conjunto de

datos. Hay que tener en cuenta que existen otros grupos funcionales, o combinaciones de ellos que pueden ser representados de muchas otras formas. Es por eso que esta etapa es tan necesaria. Un especialista podría identificar y corregir errores ahí donde un software con altas prestaciones pudiera fallar.

1.2.2. Tamaño de los conjuntos de datos y balance de sus clases.

El tamaño de los conjuntos de datos y el balance de las clases son puntos clave en la modelación de QSAR debido a que, como se mencionó anteriormente, la mayoría de los métodos de aprendizaje automatizado son muy sensibles a estos dos factores. Tropsha en su publicación [8] argumentaba que, por motivos prácticos, el número de compuestos en un conjunto de datos no debía ser muy grande, pero tampoco muy pequeño. Además planteaba que el límite superior lo impone el método de modelación empleado, y que en caso de ser muy grande el número de compuestos, se podrían poner en práctica varias estrategias tales como:

(i) seleccionar un subconjunto diverso de compuestos;
(ii) agrupar el conjunto de datos y construir modelos por separado para cada subconjunto y;
(iii) en caso de existir pocas clases o categorías, excluir varios compuestos del proceso de modelación.

Por otra parte, el límite inferior del número de compuestos, es también definido por una serie de factores [8]. Los esquemas de validación en la modelación de QSAR habitualmente establecen que el conjunto de datos sea dividido en tres partes: entrenamiento, prueba y validación externa. Dado que esto puede disminuir considerablemente el tamaño del conjunto de entrenamiento, el cual se emplea en la construcción de los modelos, si este resulta ser demasiado reducido, puede verse afectada considerablemente la predictibilidad de estos [8]. Al limitarse a ciertos niveles el número de compuestos en el conjunto de entrenamiento, el modelo a construir podría pasar por alto patrones significativos en el espacio químico.

En términos de números, Tropsha sugiere [8] que en casos en que la actividad esté definida por valores continuos, el número de compuestos en el conjunto de entrenamiento no debe ser menor que 20 y no menor a 10 para cada uno de los

conjuntos restantes (conjuntos de prueba y validación). En casos de clasificación, recomienda que deben haber no menos de 10 compuestos de cada clase para el conjunto de entrenamiento y 5 por cada clase para cada uno de los restantes conjuntos.

En cuanto al balance de las clases, lo ideal obviamente sería que el número de compuestos en todas las clases fuese aproximadamente el mismo. En el mundo real esto rara vez ocurre, generalmente es mayor el número de compuestos inactivos que el de activos [8]. Aproximaciones adicionales deben ser tenidas en cuenta, o de lo contrario, los modelos obtenidos clasificarían mejor los casos de la clase mayoritaria (compuestos inactivos). Según se explica en [8] los compuestos de la clase mayoritaria suelen ser más diversos. Estos cubren mayor área del espacio de descriptores que los activos, los cuales son muy similares los unos a los otros, hecho que dificulta o imposibilita la modelación directa empleando estos conjuntos de datos, al menos sin correr el riesgo de viciar los resultados.

Una potencial solución planteada en [8] consiste en computar la matriz de similitud entre los compuestos de las diferentes clases y eliminar los de la clase mayoritaria más discordantes de la clase minoritaria a partir de cierto valor de corte. Con este proceder se trata de priorizar aquellos patrones que diferencian las clases en los casos donde es más complejo.

En la literatura se reportan otras formas de tratar el balance de las clases, entre estas se pueden mencionar: el *over-sampling* [39, 49] de la clase minoritaria y el *under-sampling* [50, 51] de la clase mayoritaria. El primer caso consiste en generar artificialmente nuevos casos de la clase minoritaria, ya sea por replicación aleatoria de casos o generando a partir de otros criterios. En el segundo caso se reduce el número de instancias de la clase mayoritaria, donde esta eliminación puede realizarse aleatoriamente o por otras vías [51].

1.2.3. Criterios para la selección de los subconjuntos de entrenamiento, prueba y validación externa.

Como se mencionó anteriormente, en la validación de los modelos QSAR se suele dividir los conjuntos de datos de partida en subconjuntos de entrenamiento, de prueba

y de validación externa. En varias publicaciones [8, 52, 53] se ha alertado sobre la necesidad de validar adecuadamente los modelos obtenidos, y entre los principales aspectos mencionados se encuentra la adecuada partición de los conjuntos de datos. En cuanto a este aspecto, en [8] Tropsha recomienda como se debe proceder. Primero, la selección del subconjunto de validación externa se puede llevar a cabo mediante la selección aleatoria de hasta el 20% de las instancias del conjunto de datos de partida. El resto de las instancias quedan reservadas para la modelación. En el caso del subconjunto de modelación obtenido, varios autores [52-56] coinciden en que este debe ser dividido en subconjuntos de entrenamiento y prueba. Aquí, el objetivo fundamental es que ambos sean representativos de todo el espacio de descriptores del conjunto de datos original. De esta manera, cada punto del subconjunto de prueba estará cercano al menos a un punto del subconjunto de entrenamiento.

En [56-58] se reporta el uso de la selección aleatoria de los conjuntos de datos, la cual, según Golbraikh y colaboradores [54], no garantiza la representatividad de los subconjuntos de entrenamiento y prueba. Por otra parte, en [56] Ballabio y colaboradores argumentan que si el conjunto de datos de partida es lo suficientemente grande, de la selección aleatoria de instancias puede obtenerse un subconjunto de prueba que sea representativo de todo el espacio de descriptores original.

También han sido varios los trabajos que han empleado métodos de agrupamiento. En [58] se utilizó un método basado en la similitud de Tanimoto usando como valor de corte Tc=75% y *fingerprints* del tipo MACCS, en [59, 60] utilizaron otro conocido como K-means, mientras que en [61] se aplicó el algoritmo Kennard-Stones. Respecto a este tipo de métodos Golbraikh y colaboradores [54] plantearon que la dificultad de su uso radica en que diferentes clústeres presentan diferentes números de puntos, y por tanto diferentes densidades de puntos representativos. En este caso no se garantiza que cada punto del conjunto de prueba obtenido esté cerca de al menos uno del conjunto de entrenamiento. Por otra parte, señalaron que el algoritmo Kennard-Stones permite obtener conjuntos de entrenamiento y prueba que puedan satisfacer los criterios de representatividad antes mencionados.

Más prometedor aun, Golbraikh y colaboradores en [54] reportaron tres nuevos métodos basados en algoritmos de exclusión de esferas para la división racional de conjuntos de datos. Estos permitieron obtener mejores modelos en comparación con otros métodos empleados en la literatura, así como subconjuntos de datos más representativos.

1.2.4. Métodos de selección de atributos frecuentemente empleados.

Los métodos de filtrado, los métodos *wrapper* y la combinación de estos en métodos híbridos, son las tres categorías principales de técnicas de selección de atributos.

El filtrado reduce el número de descriptores a un subconjunto más pequeño basándose en un criterio específico (típicamente en el contenido informativo o correlaciones entre los atributos). Estos no aplican ningún método de aprendizaje en el proceso, por lo que se les considera métodos de selección de atributos no supervisados.

Los métodos de clasificación y regresión que utilizan una función objetivo basada en un criterio de optimización para seleccionar descriptores u otro tipo de atributos, son clasificados dentro de las técnicas de *wrapper*. Estas son más costosas computacionalmente que los métodos de filtrado, aunque su desempeño en la generalización es mejor [62].

Los métodos híbridos buscan tomar ventaja de las dos aproximaciones anteriores explotando los diferentes criterios de evaluación en diferentes etapas de la búsqueda. La mayoría de estos métodos son clasificados dentro de los métodos *wrapper*, debido a la poca diferencia con estos, aunque la diferencia radica en que antes de aplicar los métodos *wrapper* se aplican métodos de filtrado [62].

En [63] Dudek y colaboradores describen los principales métodos computacionales en el desarrollo de modelos QSAR, entre estos, describieron los métodos de selección de atributos. A continuación se comentan estos últimos tomando como referencia dicho artículo.

Métodos de filtrado.

Estos suelen ser empleados después de calcular los descriptores moleculares, como una primera aproximación de reducción de dimensionalidad y son independientes de los métodos de clasificación y/o regresión a utilizar.

(i) Métodos Basados en Correlación.

Los coeficientes de correlación de Pearson pueden servir como un filtro preliminar para descartar descriptores intercorrelacionados. Una forma de hacerlo podría ser creando clústeres de descriptores con coeficientes de correlación por encima de determinado valor de corte, reteniendo solamente uno de ellos, eligiendo así, aleatoriamente, un miembro de cada clúster. Otro procedimiento involucraría la estimación de correlaciones entre pares de descriptores y, si este valor excede determinado límite, para cada par, eliminar uno de los descriptores.

Por otra parte, ordenar los pares que se van a evaluar podría conducir a resultados significativamente diferentes. Para ello, primero se ordenan los descriptores mediante algún criterio, y luego iterativamente se busca el subconjunto de estos empezando por los pares que tienen los mejores valores en dicho ordenamiento. Un ejemplo de los criterios a aplicar en dicho ordenamiento podría ser el coeficiente de correlación entre la actividad y los descriptores.

(ii) Métodos basados en la Teoría de la Información.

El contenido informativo de un descriptor se define en términos de entropía tratándolo como una variable aleatoria. Existen varios indicadores relacionados con la información compartida por dos descriptores o por uno de estos y la actividad biológica. Un ejemplo lo constituye la información mutua o ganancia de información. Esta cuantifica la reducción de incertidumbre, o el contenido informativo, de la actividad biológica dados los valores de los descriptores y es utilizada para ordenarlos.

La aplicación del criterio basado en la teoría de la información es sencillo cuando ambos descriptores o la actividad biológica presentan valores categóricos. En caso de estar definidos por valores numéricos, deben ser aplicados algunos esquemas

de discretización. Es por eso que este criterio suele aplicarse sobre descriptores binarios.

(iii) Criterios estadísticos.

El coeficiente de Fisher, definido por la razón entre la varianza de la clase y la varianza interna de la clase, puede ser empleado para ordenar los descriptores, y luego, la correlación entre pares se aplicaría para reducir el número de estos atributos.

Otro método empleado en la evaluación de la calidad de los descriptores se basa en el estadístico de *Kolmogorov-Smirnov*. Este es un método rápido no relacionado con el conocimiento sobre la distribución subyacente y no requiriendo la conversión de los descriptores numéricos a valores categóricos. Para dos clases a ser predichas, este método mide la distancia máxima absoluta entre las funciones de distribución acumulativa de los descriptores para las clases de actividad individuales.

Métodos *Wrapper*.

Como fue anteriormente mencionado, estos métodos funcionan en conjunción con un algoritmo de clasificación o regresión (según sea el caso). La selección del mejor subconjunto de descriptores es guiada por el desempeño del algoritmo que correlaciona los descriptores con la actividad biológica para cada subconjunto evaluado.

(i) Algoritmos Genéticos (AG).

Los AG son métodos eficientes en la optimización de funciones. En la selección de descriptores, el error de predicción de los modelos construidos sobre un conjunto de atributos es optimizado. Estos simulan la evolución natural mediante la modelación de poblaciones dinámicas de soluciones. Los conjuntos de atributos seleccionados constituyen los cromosomas y estos a su vez las poblaciones de soluciones. Frecuentemente estas se codifican utilizando valores binarios para identificar cuales descriptores son seleccionados, y cuales se eliminan para cada miembro de la población generada. Cada cromosoma conduce a un modelo construido utilizando los atributos codificados. Para cada modelo, el error en la

predicción sobre el conjunto de datos de entrenamiento es cuantificado y sirve como una función de evaluación o ajuste. Durante el curso de la evolución, los cromosomas se someten al cruce y mutación. Promoviendo la supervivencia y reproducción de los cromosomas de mayor ajuste, el algoritmo minimiza de manera efectiva la función de error en las siguientes generaciones.

El desempeño de los AG depende de varios factores. Los parámetros que rigen el cruce, mutación y supervivencia de los cromosomas deberían ser cuidadosamente escogidos para permitir explorar el espacio de soluciones y prevenir una convergencia temprana hacia poblaciones homogéneas ocupando un mínimo local. La selección de una población inicial también es importante en la selección de atributos mediante este tipo de algoritmos.

Los AG han sido aplicados en la selección de atributos en QSAR con un amplio rango de métodos tales como: Redes Neuronales Artificiales, k-NN (por sus siglas en inglés: *k-Nearest Neighbor*) y *Random Forest*.

(ii) Recocido Simulado (RS).

El RS es un método estocástico para la optimización de funciones, que también ha sido aplicado en la modelación QSAR. Como en los AG, la función minimizada representa el error del modelo construido utilizando un subconjunto de descriptores. Un algoritmo RS opera iterativamente buscando un nuevo subconjunto de descriptores mediante la variación de determinado porcentaje de los atributos del mejor subconjunto. Luego, evalúa el nuevo subconjunto y hace la elección sobre si adoptar o no la nueva solución como la óptima. Esta decisión depende del desempeño de la nueva solución. Si disminuye el error de predicción respecto a la conocida como mejor solución, entonces la sustituye, de lo contrario, no se descarta por completo. Con una probabilidad dada, basada en la Distribución de Boltzman, la peor solución puede reemplazar la mejor, lo que permite al método escapar de mínimos locales de la función de error.

El poder del método RS proviene de alterar el término temperatura en la distribución de Boltzman. Al inicio, en las primeras etapas, cuando la solución no ha sido altamente optimizada y existe gran tendencia a encontrar mínimos locales, la

temperatura es alta. Durante el curso del algoritmo, la temperatura es disminuida y la aceptación de las peores soluciones es menos probable. Por esto, incluso si el mínimo obtenido no es el global, aun así, usualmente este es de gran calidad.

(iii) Selección Secuencial de Atributos Hacia Adelante.

A diferencia de los AG y RS este método opera de forma determinística, llevando a cabo una búsqueda ambiciosa a través de los subconjuntos de atributos. Como primera etapa, es seleccionado el atributo que conduce a la mejor predicción. Luego, secuencialmente, cada atributo es individualmente adicionado al correspondiente subconjunto de estos formado, y los errores de los modelos resultantes son cuantificados. El atributo que resulte ser el mejor en reducir el error de predicción es incorporado en el subconjunto antes mencionado. Así, de esta manera, en cada etapa el mejor atributo es adicionado, resultando en una secuencia de subconjuntos de atributos anidados. El algoritmo se detiene cuando un número específico de atributos es seleccionado.

La desventaja de este método es que si varios atributos son buenos prediciendo colectivamente, es posible que ninguno de estos sea elegido.

(iv) Eliminación Secuencial de Atributos Hacia Atrás.

Este es otro ejemplo de un método secuencial ambicioso. En contraste con la *Selección Hacia Adelante,* al inicio se emplea el conjunto de descriptores completo. Luego, en cada etapa, todos los subconjuntos de atributos resultantes de remover un solo atributo son analizados y el atributo cuya eliminación conduzca al modelo con mayor error es removido. El algoritmo se detiene cuando determinado número de atributos es eliminado.

Este método es más lento que el anterior y, sin embargo, frecuentemente conduce a mejores resultados.

Métodos Híbridos.

Como se había mencionado, los métodos híbridos combinan métodos de filtrado y *wrapper*. Primero se emplearía un método de filtrado, y luego, como complemento, un método *wrapper*, el cual sería computacionalmente más lento que el primero, pero más preciso. Un ejemplo de dicho proceder sería reducir primero aplicando una prueba

basada en correlación, seguida por un AG, RS o uno de los algoritmos de selección secuencial visto.

La selección de atributos puede estar implícita en algunos métodos de clasificación y/o regresión. Un ejemplo de esto lo constituyen los árboles de decisión, los que solamente utilizan un subconjunto de atributos en el proceso de decisión.

1.3. Algoritmos de aprendizaje automatizado en la modelación QSAR.

Una vez realizada la selección de atributos, la etapa final en la construcción de los modelos QSAR consiste en obtener la relación o función matemática entre los descriptores y la actividad biológica. En algunos casos se emplean métodos de modelación donde la actividad biológica es modelada como función lineal de los descriptores, mientras que en otros, dicha función resulta más compleja, obteniéndose relaciones no lineales.

En cuanto a la naturaleza de la variable dependiente, cuando la actividad biológica toma valores numéricos, estamos en presencia de un problema de regresión; mientras que si resulta estar definida por valores categóricos (por ejemplo: *activo o inactivo*), estaríamos enfrentándonos a un problema de clasificación.

A continuación se resumen y describen brevemente los principales algoritmos de aprendizaje automatizado utilizados en la obtención de modelos QSAR. Al igual que en el Subepígrafe 1.2.4, este epígrafe se basa fundamentalmente en la revisión [63], donde aparece esta información desglosada y ampliamente comentada y referenciada.

1.3.1. Modelos lineales.

Desde los inicios de la modelación QSAR, estos han constituido las bases de su análisis. En general los modelos lineales son fácilmente interpretables y suficientemente precisos para pequeños conjuntos de datos de compuestos similares, especialmente cuando los descriptores son cuidadosamente seleccionados para un tipo de actividad biológica específica.

Regresión Lineal Múltiple (RLM).

La RLM es uno de los métodos más utilizados en la modelación de QSAR en las últimas décadas. Mediante esta se modela la actividad biológica como una función

lineal de todos los descriptores, donde los coeficientes de la función son estimados a partir de instancias del conjunto de entrenamiento. Estos parámetros son estimados minimizando los errores entre la actividad predicha y la medida experimentalmente. La principal restricción del análisis por RLM es que, cuando la razón descriptores/compuestos es grande o existen descriptores multicolineales, los modelos obtenidos tienden al sobreajuste [64], lo que hace los resultados inestables.

Mínimos Cuadrados Parciales (PLS).

La regresión lineal por PLS (por sus siglas en inglés, *Partial Least Squares*) es un método adecuado para superar los problemas planteados para la RLM. En esta se asume que, a pesar del enorme número de descriptores, el proceso de modelación es gobernado por un número relativamente pequeño de variables independientes latentes, y trata de obtener indirectamente conocimiento sobre estas variables descomponiendo la matriz de descriptores de entrada en dos componentes, matriz *scores* y matriz *loadings.* Los scores son ortogonales y, mientras son capaces de capturar la información de los descriptores, permiten además el buen desempeño en la predicción de la actividad.

Análisis Discriminante Lineal (LDA).

El LDA (por sus siglas en inglés, *Linear Discriminant Analysis*) es un método de clasificación que crea una transformación lineal del espacio original, a un espacio que maximiza la separabilidad entre clases y minimiza la varianza interna de estas. Consiste en resolver un problema de valores propios basado en las matrices de covarianza inter e intra-clases, por lo cual el número de descriptores debe ser significativamente menor que el número de compuestos para lograr un buen desempeño de este método. Para evitar este problema, previamente, puede ejecutarse el Análisis de Componentes Principales (PCA) para reducir la dimensionalidad.

1.3.2. Modelos no lineales.

Los modelos no lineales extienden las SAR a funciones no lineales de los descriptores. Estos pueden ser más precisos, en especial para conjuntos de datos grandes y

diversos, pero al costo de ser menos interpretables. Por otra parte, modelos no lineales complejos pueden tender al sobreajuste de los datos.

Clasificador de Bayes (CB).

El CB se deriva de la regla de Bayes que relaciona la probabilidad a posteriori de una clase a su probabilidad global, la probabilidad de las observaciones y la probabilidad de una clase con respecto a las variables observadas. En la regla de Bayes, la clase que minimiza la probabilidad a posteriori se elige como el resultado de predicción. Sin embargo, en los problemas reales, las probabilidades no son conocidas y tienen que ser estimadas. Todavía, dado un número finito de ejemplos de entrenamiento, tal estimación no es trivial. Un método para abordar este problema es hacer una suposición de independencia de probabilidades de clase con respecto a los diferentes descriptores. Esto lleva al *Clasificador de Bayes Ingenuo (CBI)*. Para conjuntos de datos típicos, la estimación de probabilidades con respecto a variables individuales es factible. El inconveniente de este método es que la hipótesis de la independencia por lo general no se sostiene.

Método de los k vecinos más cercanos (k-NN).

El k-NN es un esquema de decisión simple que casi no requiere entrenamiento y es asintóticamente óptimo, ya que con el incremento del número de instancias del conjunto de datos de entrenamiento converge al error de predicción óptimo. Para un compuesto representado en el espacio de los descriptores, este método analiza los k compuestos más cercanos a este pertenecientes al conjunto de datos de entrenamiento y predice la clase que más representada esté en dicha vecindad. K-NN resulta sensible a la métrica de distancia escogida y al número de compuestos de entrenamiento disponibles. Por otra parte, el número de vecinos analizados puede ser optimizado para obtener los mejores resultados.

Redes Neuronales Artificiales (RNA).

Las RNA son métodos de predicción biológicamente inspirados, basados sobre la arquitectura de una red de neuronas. En la literatura se han analizado un amplio rango de modelos basados en ese paradigma, prevaleciendo dos tipos fundamentales: los

basados en perceptrón y los basados en funciones de base radial. Ambos entran en la categoría de las redes de alimentación hacia adelante, en las que durante la predicción, la información fluye sólo en la dirección de los descriptores de entrada, a través de un conjunto de capas, y hasta la salida de la red. A continuación se describen estos dos métodos:

(i) *Perceptrón Multicapa (MLP).*

Un modelo MLP (por sus siglas en inglés, *Multilayer Perceptron*) consiste en una red por capas de neuronas interconectadas. Cada perceptrón es capaz de hacer una combinación lineal de sus valores de entrada y, por medio de cierta función de transferencia, una salida binaria o continua. Cada entrada tiene un peso adaptativo especificando la importancia de esta. En el entrenamiento de un solo perceptrón, las entradas están constituidas por los descriptores moleculares (atributos), mientras que la salida debería predecir la actividad del compuesto (la clase). Con este fin, el perceptrón es entrenado mediante el ajuste de los pesos, para producir una combinación lineal de descriptores que predigan óptimamente la actividad. El proceso de ajuste se basa en la retroalimentación de comparar la salida predicha con la esperada. Los valores de los pesos son modificados con el objetivo de disminuir el error de la predicción.

Mientras un solo perceptrón conduce a un modelo lineal, una red consistente en múltiples capas de estos, con las salidas de una capa conectadas a las entradas de las neuronas de la que le sigue, conduce a una predicción no lineal. Una red multicapa está formada por una capa de entrada, consistente en los valores de los descriptores moleculares, una o más capas ocultas, las que procesan estos descriptores en representaciones internas y una capa de salida que utiliza estas representaciones para generar la predicción final.

De hecho, cualquier función puede ser aproximada utilizando una red neuronal lo suficientemente grande, aunque esto podría conducir al sobreajuste en el caso de emplear conjuntos de datos finitos. Es por esto que la elección del número de capas y de neuronas es esencial en la construcción de redes neuronales artificiales. También es importante la elección de la función mediante la cual se actualizan los pesos a partir del error en las predicciones. La más popular es la *delta rule*. En esta,

el cambio del peso es proporcional a la diferencia entre lo predicho y lo esperado, donde la constante de proporcionalidad determina la tasa de aprendizaje. Otro aspecto a tener en cuenta es la selección de la función de transferencia de la neurona, teniendo a la Función Sigmoidea como una de las más utilizadas. Finalmente, los valores de los pesos iniciales deben ser generados. Una forma de obtener estos pesos sería a través de la generación de números aleatorios de pequeña magnitud.

(ii) *Funciones de Base Radial (RBF).*

Estos modelos de redes neuronales RBF (por sus siglas en inglés, *Radial Basis Function*) consisten en una capa de entrada, una sola capa oculta y una capa de salida. A diferencia de las redes MLP, las neuronas en la capa oculta no computan sus salidas basadas en el producto de los pesos y los valores de entrada. Cada neurona en dicha capa es definida por su centro, el cual constituye un punto en el espacio de los descriptores. La salida de la neurona es calculada como la función de la distancia entre el compuesto de entrada y el punto constituyente de la neurona. Típicamente es empleada la Función Gaussiana, aunque otras funciones de distancia pueden ser aplicadas. La neurona de salida es del tipo perceptrón, teniendo la salida como una función de transferencia del producto de los valores de salida de las neuronas de RBF y sus pesos.

Varios parámetros deben ser ajustados durante el proceso de entrenamiento de una red neuronal de tipo RBF. Primero se deben crear neuronas del tipo RBF y sus centros, además de escalar sus distancias. En caso de la Función Gaussiana, estos parámetros corresponden a la media y la desviación estándar. La aproximación más simple es crear tantas neuronas como compuestos en el conjunto de entrenamiento y localizar los centros en las coordenadas de cada uno de los compuestos. De igual manera se podría agrupar el conjunto de entrenamiento en un número de grupos y utilizar los centros de estos grupos.

Árboles de Decisión (DT).

Cada DT (por sus siglas en inglés, *Decision Tree*) puede ser descrito como un conjunto de reglas predictivas basadas en la Lógica Booleana. Un modelo de clasificación

basado en DT consiste en una estructura de “árbol” consistente en nodos y enlaces. Los primeros son enlazados jerárquicamente, con varios nodos hijo bifurcándose de un mismo nodo padre. Un nodo sin nodos hijos se denomina hoja (en analogía con las hojas de los árboles en la naturaleza). En cada nodo se lleva a cabo una evaluación basada en un solo descriptor. Teniendo en cuenta el resultado de dicha evaluación, el algoritmo es dirigido a uno de los nodos hijo, y así repetidamente hasta llegar a un nodo hoja. La decisión final es basada en la actividad (clase) asociada con el nodo hoja.

El proceso de entrenamiento del modelo consiste en la adición incremental de nodos. Este comienza con la realización de varias pruebas al nodo raíz, donde es elegida la que separe óptimamente los compuestos en las clases de actividad apropiadas. Si una prueba permite discriminar completamente las clases, no se adicionarán más nodos. En cada etapa del proceso, cada nodo es examinado respecto al criterio de ser o no un nodo hoja. Siendo un caso positivo cuando los compuestos dirigidos a este desde el nodo raíz quedan dentro de una sola clase o al menos una clase forma una clara mayoría.

Un modelo de DT puede conducir al sobreajuste del conjunto de entrenamiento si durante el proceso de construcción este crece hasta que los nodos hoja consistan solamente de casos pertenecientes una sola clase. Es por esto que se detiene tempranamente el algoritmo una vez que los nodos son lo suficientemente puros.

Máquinas de Soporte Vectorial (SVM).

Las SVM (por sus siglas en inglés, *Support Vector Machines*) buscan crear un hiperplano de decisión que maximice la distancia (margen) de este a cada uno de los casos más cercanos de cada una de las clases. Este permite la formulación del entrenamiento de un clasificador como un problema de optimización constreñido.

En los casos más simples, compuestos de diferentes clases pueden ser separados por un hiperplano lineal, el cual es definido solamente por los compuestos más cercanos del conjunto de entrenamiento. Estos compuestos (o instancias) son conocidos como vectores de soporte, dándole el nombre al método.

En la mayoría de los casos no es posible la separación lineal de las clases, problema que se trata introduciendo variables de holgura. Estas variables son asociadas con los compuestos mal clasificados y, en conjunto con el margen, son sujetas a optimización.

Las SVM son fácilmente transformadas en clasificadores no lineales mediante el empleo de la *función kernel*. Dicha función introduce un mapeo implícito del espacio original de los descriptores a un espacio de altas o infinitas dimensiones. El hiperplano lineal en el espacio *kernel* puede ser altamente no lineal en el espacio original. Existen dos funciones *kernel* frecuentemente empleadas, estas son: el *kernel polinomial* y la *función kernel de base radial*.

Métodos Ensemble (ME).

Estos métodos consisten en combinar las salidas de varios clasificadores buscando mejorar el desempeño de la clasificación.

(i) *Bagging.*

Este método se centra en mejorar la calidad de predicción mediante la creación de un conjunto de modelos base construidos usando el mismo algoritmo, pero variando el conjunto de entrenamiento. Antes del entrenamiento de cada modelo base, el conjunto de datos original es sujeto a un muestreo con reemplazo, lo que conduce a réplicas de partida. Las decisiones de los modelos entrenados sobre cada réplica son promediadas para obtener el resultado final.

(ii) *Método del Subespacio Aleatorio (MSA).*

Este es otro ejemplo de esquema de ensemble que busca estabilizar el modelo base. En el entrenamiento de los modelos, todo el conjunto de entrenamiento es empleado. Sin embargo, para garantizar la diversidad, solo se emplean subconjuntos de descriptores escogidos aleatoriamente. El ejemplo más notable de MSA es Random Forest (RF), el cual usa árboles de decisión como modelos base.

(iii) *Boosting.*

Este es un esquema de ensemble más elaborado que los anteriores. Durante el entrenamiento de los modelos base, este utiliza todos los descriptores y compuestos en el conjunto de entrenamiento. Sin embargo, para cada compuesto se define un

peso. Inicialmente los pesos son uniformes. Después del entrenamiento de un modelo base, su error de predicción es evaluado y los pesos de los compuestos incorrectamente predichos son incrementados. Este enfoca el próximo modelo base sobre los casos previamente mal clasificados, incluso a costa de generar errores para aquellos correctamente clasificados anteriormente. Es por esto que, los compuestos cuya predicción de actividad sea más difícil obtienen más atención del modelo de ensemble. La ventaja de este método respecto a los otros de tipo ensemble radica en que permite el empleo de modelos bases relativamente simples.

1.4. Los *activity cliffs*, el aprendizaje automatizado, y la quimioinformática.

La continuidad de los espacios SAR provee la base fundamental para el análisis de QSAR. En contraste, la discontinuidad de dichos espacios va en detrimento de la habilidad predictiva de los modelos QSAR [2]. Mientras que los métodos de aprendizaje estadístico han jugado un rol protagónico en el desarrollo de la modelación de QSAR. Actualmente, los algoritmos de aprendizaje automatizado son las herramientas más extendidas en las aplicaciones quimioinformáticas [65, 66]. Al igual que para QSAR y las aproximaciones basadas en similitud, los métodos de aprendizaje automatizado se basan en la continuidad de los espacios de las SAR. Los dos propósitos generales por los cuales el aprendizaje automatizado es empleado en quimioinformática son la clasificación y la generalización de los datos, extrayendo por diferentes vías regularidades o patrones de estos. En el descubrimiento de fármacos, los algoritmos de aprendizaje automatizado emplean el conocimiento sobre las SAR para generar clasificaciones y generalizaciones que son conceptualmente significativas [67]. En este contexto los *activity cliffs* son excepciones o contradicciones al supuesto de que el espacio de las SAR del conjunto de datos es continuo. Así, si el mecanismo de clasificación en aprendizaje automatizado es comprendido como una función que mapea una descripción de un ejemplo (la estructura química codificada por descriptores moleculares) a su etiqueta (por ejemplo, un valor continuo o la pertenencia a una clase), el efecto negativo de los *activity cliffs* queda claro.

La mayoría de las técnicas de aprendizaje automatizado capturan las mayores tendencias y fallan en reconocer los *activity cliffs,* reduciendo la confiabilidad de las

potenciales predicciones. Incluso para técnicas avanzadas, tales como redes neuronales y máquinas de soporte vectorial, capaces de manipular relaciones no lineales, constituye una dificultad lidiar con estos pares de compuestos. Pero inclusive, si el modelo de aprendizaje automatizado tiene éxito capturando la mayor parte de los *activity cliffs* más relevantes, lo logra a un alto costo. Un modelo que aprende de un conjunto de entrenamiento incluyendo un número significativo de estos, tiende al sobreajuste [68].

Finalmente, aunque es altamente deseable esforzarse por un modelo de aprendizaje automatizado que explique eficientemente la variabilidad, es importante también estar consciente que es un tanto utópico encontrar dicho algoritmo. El presente conocimiento químico y biológico es todavía inmaduro, y además, sobre la base de información incompleta [16]. En este caso sería mejor afrontar dichas dificultades tratando dos problemáticas fundamentales: ¿cómo lidiar con datos quimioinformáticos y la ausencia de la habilidad predictiva de los modelos entrenados sobre este tipo de datos? y ¿cómo desarrollar modelos predictivos con SAR heterogéneas?

1.5. Conclusiones del capítulo

En este capítulo se discutieron aspectos fundamentales referentes al concepto de *activity cliffs*; definiéndose estos en el marco de la modelación QSAR y la química medicinal. También se ilustraron evidencias presentes en la literatura sobre los efectos negativos de la presencia de los *activity cliffs* sobre la modelación QSAR. Además se estableció una analogía entre estos y las ISM, cuya eliminación, según se reporta en la literatura, permitió mejorar significativamente el desempeño de gran variedad de algoritmos de aprendizaje automatizado. Por otra parte, no existen reportes mostrando que remover los *activity cliffs* de conjuntos de datos sea beneficioso, perjudicial o no significativo en la obtención de modelos de las SAR.

Aquí se abordó el pre-procesamiento estructural de los conjuntos de datos sobre la base de seguir las “buenas prácticas” para la modelación QSAR. Además, se expusieron y analizaron los diferentes procedimientos previamente reportados para realizar una partición adecuada de los conjuntos de datos. Por último, se presentó un resumen de las familias de métodos de selección de atributos y los principales

algoritmos de aprendizaje automatizado que han sido empleados en la modelación QSAR.

CAPÍTULO 2. PROCEDIMIENTOS Y MÉTODOS.

En este capítulo se expone el procedimiento propuesto para la eliminación de los *activity cliffs*. Este incluye un algoritmo para la eliminación de estos pares de compuestos, el cual es un aporte de esta investigación. Posteriormente, se expone una descripción detallada de los conjuntos de datos empleados y un resumen de los resultados de la etapa de pre-procesamiento.

2.1. Propuesta de procedimiento previo a la modelación QSAR. Detección y eliminación de los *activity cliffs* más influyentes.

La eliminación de los *activity cliffs* es una tarea altamente sensible a la calidad de las representaciones moleculares. Si no se garantiza una correcta y homogénea representación de los grupos funcionales en estas, la veracidad de los resultados podría verse comprometida. Falsos *activity cliffs* podrían detectarse en lugar de duplicados estructurales, cuyos grupos funcionales se encuentran representados de diferentes formas. Adicionalmente, los modelos podrían ser construidos sobre la base de información errónea. Para dar solución a esta problemática, se propone un procedimiento el cual contempla el pre-procesamiento estructural previo a la eliminación de los *activity cliffs*. Este puede ser dividido en tres etapas:

(i) La primera consiste en curar los conjuntos de datos mediante el pre-procesamiento estructural y la eliminación de duplicados.
(ii) La segunda etapa contempla la eliminación de los *activity cliffs* más influyentes.
(iii) Por último, la tercera etapa incluye el cálculo de los descriptores moleculares.

En los subepígrafes siguientes se describirán en detalle los pasos a desarrollar cada una de estas etapas.

2.1.1. Procedimiento de curación de los conjuntos de datos.

Para llevar a cabo esta etapa, el primer paso consiste en importar los conjuntos de datos a una hoja de *Excel* empleando para ello el *plugin JChem for Excel* de *ChemAxon*. Después, se procede a realizar un análisis elemental de las estructuras químicas utilizando la función *JCAtomCount,* la cual forma parte del *plugin* de

ChemAxon antes referido. Por esta vía es posible la detección y eliminación de compuestos inorgánicos, organometálicos y que contienen elementos traza o poco representados en el conjunto de datos. En los casos de las mezclas, sales, y la normalización de los diferentes quimiotipos, se procede como se recomienda en el Subepígrafe 1.2.1.

En la normalización de las estructuras, se toman en cuenta los elementos representados en el conjunto de datos, y en base a estos se escogen los quimiotipos a normalizar. El paso siguiente consiste en la generación del fichero de configuración *XML* con las posibles transformaciones, a partir de las cuales el *Standardizer* [45] normaliza las estructuras. Después se vuelve a hacer una inspección visual para detectar deficiencias en este proceso, y de ser necesario, repetirlo iterativamente.

Una vez manejados los casos de mezclas, sales, y estandarizadas las estructuras, se procede a predecir las cargas de las moléculas en condiciones de pH común a todas. Para esto se puede emplear el programa *cxcalc* [43] en la consola de Windows, con el comando: *"cxcalc* ***majormicrospecies -H*** *[valor de PH]* ***-f*** *sdf* ***-K*** *[nombre fichero entrada].sdf > [nombre fichero salida].sdf"*.

La detección y eliminación de los duplicados es el último paso en esta etapa y debe realizarse siguiendo las indicaciones dadas en el Subepígrafe 1.2.1. En este caso cabe señalar que es crucial la supervisión de un especialista con cierto nivel de experiencia o conocimiento en el área de las ciencias químicas.

2.1.2. Algoritmo de detección y eliminación de *activity cliffs*.

Para la detección de los *activity cliffs* es necesario establecer dos criterios básicos. El primero se relaciona con la similitud estructural y el segundo con la diferencia de potencia. Los *activity cliffs* son pares de compuestos muy similares estructuralmente (criterio de similitud), que difieren en gran medida en los valores de la actividad biológica (criterio de potencia). Si se definen los valores de corte apropiados, estos criterios permitirían definir cuándo estamos frente a un par de estos. Bajo el principio anterior, se hace necesario:

(i) Definir una métrica que nos permita medir cuantitativamente las diferencias estructurales, así como una forma de representación que capture estas diferencias para cada conjunto de datos en particular.

(ii) Establecer un valor de corte para dicha medida de similitud (o distancia) elegida en el contexto de una forma de representación dada.

(iii) Establecer un valor de corte para las diferencias en actividad biológica en los pares que se cumpla el criterio de similitud.

En cuanto a la métrica de similitud estructural, el coeficiente de Tanimoto es el más ampliamente utilizado en quimioinformática. Por otra parte, como se había mencionado en el Subepígrafe 1.1.5, los *fingerprints* son las representaciones más idóneas para codificar las similitudes globales de las estructuras, por lo tanto, son la forma de representación elegida para esta tarea.

Los valores de corte, tanto de similitud de Tanimoto, como de actividad biológica fueron establecidos. Se tomaron como referencia valores del coeficiente de Tanimoto superiores o iguales a 0.55 para los *fingerprints* del tipo ECFP (por sus siglas en inglés, *Extended Connectivity Fingerprints*) y diferencias de actividad superiores a dos unidades logarítmicas, como se sugiere por Hu y colaboradores en [36].

Es necesario aclarar que se parte de un conjunto de datos de entrenamiento (CDE) que contiene un campo con las representaciones estructurales de las moléculas, un campo con los valores de actividad en datos de potencia y otro campo con las etiquetas de las clases. Además, se generan dos listas, la primera, consiste en una lista de candidatos (LC), que son los compuestos presentes en al menos un par de compuestos que cumplen simultáneamente con los criterios de similitud y diferencia de potencia. Mientras que la segunda contiene los compuestos que se van eliminando (*generadores de cliffs*) en cada ciclo del algoritmo (LE).

Pasos del algoritmo para la detección y eliminación de los *activity cliffs*:

Paso 1: Calcular los valores de cada tipo de *fingerprints* para las estructuras en CDE.

Paso 2: Calcular las matrices de similitud $S = \sum\sum s_{i,j}$ para cada tipo de *fingerprints* cuyos valores fueron calculados en el paso anterior.

Paso 3: Calcular las matrices de distancia entre moléculas $D = \sum\sum d_{i,j}$, a partir de las matrices de similitud antes obtenidas, donde cada valor de distancia vendrá dado por: $d_{i,j} = 1 - s_{i,j}$.

Paso 4: Fusionar las matrices de distancia para obtener una matriz de distancia consenso $MDC = \sum\sum d'_{i,j}$, para todos los tipos de representaciones (*fingerprints*) empleados en el primer paso.

Paso 5: Buscar en el CDE los pares de compuestos (i,j), que tomando como referencia los *fingerprints* del tipo ECFP, cumplan con: $d_{i,j} \leq 0{,}45$ o lo que es lo mismo: $s_{i,j} \geq 0{,}55$.

Paso 6: Teniendo en cuenta el número de pares que cumplen los criterios establecidos en el paso anterior, determinar el valor de corte de similitud para los compuestos en CDE que permita obtener igual distribución a partir de la MDC.

Paso 7: Computar la matriz de diferencias de potencia $P = \sum\sum p_{i,j}$, donde cada $p_{i,j}$ representa el valor absoluto de la diferencia entre las potencias (en unidades logarítmicas) de los compuestos (i,j).

Paso 8: Buscar en P los pares de compuestos (i,j) que cumplan simultáneamente con el valor de corte obtenido en el paso 6 y con: $p_{i,j} \geq 2.0$, y agregarlos a la LC, inicializar LE como una lista vacía.

Paso 9: Para cada candidato en LC calcular: el número de compuestos con los que forma *activity cliffs*.

Paso 10: Agregar a la LE el compuesto que forme la mayor cantidad de *activity cliffs*, actualizar la LC quitando el compuesto movido a la LE y comenzar de nuevo a partir del Paso 9. De existir empate, definir una LC* que contenga solamente estos compuestos, e ir al paso siguiente.

Paso 11: Para cada compuesto en LC*, calcular cuántos *activity cliffs* forman con compuestos de la clase opuesta, agregar a la LE el que forme la mayor cantidad, actualizar la LC y comenzar de nuevo a partir del Paso 9. De existir

empate, actualizar la LC* dejando solamente estos compuestos, e ir al paso siguiente.

Paso 12: Para cada compuesto en LC*, determinar cuáles pertenecen a la clase mayoritaria, si existe uno solo, agregarlo a la LE, actualizar la LC y comenzar de nuevo a partir del Paso 9. De haber empate, actualizar la LC dejando solamente estos compuestos, e ir al paso siguiente.

Paso 13: Para cada compuesto en LC*, buscar el más cercano a la frontera de clasificación, agregarlo a la LE, actualizar la LC* y comenzar de nuevo a partir del Paso 9. Si existe empate, agregar aleatoriamente a la LE, actualizar la LC y comenzar de nuevo a partir del Paso 9.

Paso 14: Eliminar del CDE los compuestos en la LE.

Para ejecutar los pasos 1 y 2 existe gran diversidad de herramientas disponibles. En esta investigación se escogió el paquete *Mayachemtools* [69], el cual se empleó para calcular los valores de 11 tipos de *fingerprints* (entre ellos ECFP) y de las marices de similitud para cada uno de estos.

2.1.3. Cálculo de descriptores moleculares.

Para el cálculo de los descriptores moleculares se escogió el programa ISIDA Fragmentor2011 [70]. Este software, además de calcular una gran variedad de descriptores moleculares, presenta la ventaja de brindarle al usuario un conjunto de opciones adaptables a sus necesidades particulares. El cálculo de los descriptores puede ser limitado a unos pocos, y el tamaño y tipo de fragmentos moleculares a determinar puede ser regulado por el usuario. Una vez aplicado un método de selección de atributos a un conjunto de datos de entrenamiento, es muy sencillo calcular el reducido número de descriptores escogidos para los subconjuntos de prueba y validación externa.

La estructura general de los comandos a emplear para calcular los descriptores moleculares aparece explicada con detalle en [70].

2.2. Pre-procesamiento de los conjuntos de datos seleccionados para la modelación.

Para llevar a cabo el procedimiento propuesto en el epígrafe anterior, se seleccionaron 5 conjuntos de datos reportados en la literatura. En todos los casos se contó con datos de potencia y etiqueta de clase. Por otra parte, no se llevó a cabo ningún procedimiento para balancear los conjuntos de datos, puesto que no es objeto de esta investigación y, además podría alterar/viciar sus resultados. A continuación se describen los conjuntos de datos y el paso de estos por el procedimiento propuesto.

2.2.1. Descripción de los conjuntos de datos

Tetrahymena Pyriformis *(ci700443v).*

Este conjunto de datos fue reportado en [71]. Constituido por 1093 compuestos y valores numéricos de la actividad biológica expresada como el logaritmo de la concentración inhibidora del crecimiento del 50% de las poblaciones de *Tetrahymena Pyriformis* (pIGC50). Después de establecer como valor de corte pIGC50 = 0, la distribución de las clases fue de 448 no-tóxicos y 645 tóxicos.

TRPV1 *(minf00320555).*

Este conjunto de datos fue reportado en [72]. Es una recopilación de 408 antagonistas del *Receptor de Potencial Transitorio V1* (TRPV1, por sus siglas en inglés, *Transient Receptor Potential Vanilloid Type 1*). De ellos 201 activos y 207 inactivos.

BZR, COX-2 y DHFR.

Estos conjuntos de datos fueron reportados en [73]. BZR consiste en 306 ligandos del Receptor Benzodiazepínico, de ellos 157 activos y 149 inactivos. COX-2 es una recopilación de 303 inhibidores de la Ciclooxigenasa-2 (COX-2), de los cuales 148 son etiquetados como activos y 155 como inactivos. DHFR está formado por 393 inhibidores de la Dihidrofolato Reductasa (DHFR), con un total de 126 activos y 267 inactivos.

2.2.2. Proceso de curación de los conjuntos de datos

Los conjuntos de datos antes mencionados fueron sometidos al proceso de curación mencionado en el Subepígrafe 2.1.1. Como resultado de este proceso fueron eliminados compuestos que presentaban elementos traza o poco representados (presentes en menos de tres compuestos), elementos raros o poco comunes en sistemas biológicos, así como los duplicados estructurales presentes. A continuación se muestran los resultados de este proceso para cada conjunto de datos empleado.

Tetrahymena Pyriformis.

Después de realizar el análisis elemental de este conjunto de datos, se determinó que no era necesario eliminar compuestos por este criterio. En la Tabla 2.1 se muestra la distribución elemental de este conjunto de datos, en esta se resumen los valores del número y porciento de los compuestos para cada elemento presente.

Tabla 2.1 Análisis elemental del conjunto de datos *Tetrahymena Pyriformis.*

Elemento Químico	**H**	**C**	**O**	**N**	**S**	**F**	**Cl**	**Br**	**I**
Número de compuestos	1089	1093	877	404	63	36	156	104	10
%	99.6	100	80	37	5.76	3.29	14	10	0.91

Una vez de estandarizadas y protonadas las estructuras, se realizó el análisis de duplicados estructurales, encontrándose seis pares de duplicados de los cuales se eliminaron diez compuestos. En la Tabla 2.2 se muestran los pares identificados, así como los compuestos eliminados por esta vía.

Tabla 2.2 Análisis de duplicados estructurales del conjunto de datos *Tetrahymena Pyriformis*. Se resaltan en negrita los compuestos eliminados.

Par	**Compuesto 1**	**Clase**	**Compuesto 2**	**Clase**
1	**137406**	**0**	**79094**	**0**
2	**156547**	**0**	**107926**	**0**
3	1984061	1	**124072**	**1**
4	**150903**	**0**	**110156**	**0**
5	**141822**	**0**	**26522850**	**0**
6	10051442	0	**142621**	**0**

Nótese que en los pares 1, 2, 4 y 5 se eliminaron ambos compuestos en cada caso. Esto se hizo debido a las apreciables discrepancias entre los valores de potencia (reportados para estos compuestos) en cada uno de estos pares.

TRPV1.

Después de realizar el análisis elemental de este conjunto de datos, se determinó que debía ser eliminado el compuesto con el identificador: " *CHEMBL456982*", el que contiene el elemento fósforo (P), elemento poco representado en el conjunto de datos. En la Tabla 2.3 se muestra la distribución elemental de este conjunto de datos, en esta se resumen los valores del número y porciento de los compuestos para cada elemento presente.

Tabla 2.3 Análisis elemental del conjunto de datos*TRPV1*.

Elemento Químico	**H**	**C**	**O**	**N**	**P**	**S**	**F**	**Cl**	**Br**	**I**
Número de compuestos	408	408	314	405	**1**	59	260	86	8	4
%	100	100	77	99	**0.25**	14	64	21	1.96	0.98

Una vez de estandarizadas y protonadas las estructuras, se realizó el análisis de duplicados estructurales, encontrándose tres pares de duplicados, de los cuales se eliminaron los 6 compuestos, debido a grandes discrepancias en los valores de potencia para cada uno de estos pares. En la Tabla 2.4 se muestran los pares identificados.

Tabla 2.4 Análisis de duplicados estructurales del conjunto de datos *TRPV1*. En este caso todos fueron los compuestos fueron eliminados.

Par	**Compuesto 1**	**Clase**	**Compuesto 2**	**Clase**
1	CHEMBL480447	1	CHEMBL479646	0
2	CHEMBL183321	0	CHEMBL183897	0
3	CHEMBL1630624	0	CHEMBL1630623	0

BZR.

Después de realizar el análisis elemental de este conjunto de datos, se determinó que debían ser eliminados los compuestos con los identificadores: "*Ro07-9238*", *"Bromazepam"* y *"Ro31-0214"*, los cuales contienen los elementos P y Br, elementos poco representados en el conjunto de datos. En la Tabla 2.5 se muestra la distribución

elemental de este conjunto de datos, en esta se resumen los valores del número y porciento de los compuestos para cada elemento presente.

Tabla 2.5 Análisis elemental del conjunto de datos *BZR*.

Elemento Químico	**H**	**C**	**O**	**N**	**P**	**S**	**F**	**Cl**	**Br**
Número de compuestos	306	306	246	306	**1**	21	65	121	**2**
%	100	100	80	100	**0.33**	6.86	21	40	**0.65**

Una vez de estandarizadas y protonadas las estructuras, se realizó el análisis de duplicados estructurales donde no se detectaron pares de compuestos duplicados.

COX-2.

Después de realizar el análisis elemental de este conjunto de datos, se determinó que no era necesario eliminar compuestos por este criterio. En la Tabla 2.6 se muestra la distribución elemental de este conjunto de datos, en esta se resumen los valores del número y porciento de los compuestos para cada elemento presente.

Tabla 2.6 Análisis elemental del conjunto de datos *COX-2*.

Elemento Químico	**H**	**C**	**O**	**N**	**S**	**F**	**Cl**	**Br**
Número de compuestos	303	303	301	259	296	232	73	7
%	100	100	99	85	98	77	24	2.31

Una vez estandarizadas y protonadas las estructuras, se realizó el análisis de duplicados estructurales donde no se detectaron pares de compuestos duplicados.

DHFR.

Después de realizar el análisis elemental de este conjunto de datos, se determinó que no era necesario eliminar compuestos por este criterio. En la Tabla 2.7 se muestra la distribución elemental de este conjunto de datos, en esta se resumen los valores del número y porciento de los compuestos para cada elemento presente.

Tabla 2.7 Análisis elemental del conjunto de datos *DHFR*.

Elemento Químico	**H**	**C**	**O**	**N**	**S**	**F**	**Cl**	**Br**
Número de compuestos	393	393	255	393	56	17	82	10
%	100	100	65	100	14	4.33	21	2.54

Una vez de estandarizadas y protonadas las estructuras, se realizó el análisis de duplicados estructurales donde no se detectaron pares de compuestos duplicados.

2.2.3. Partición de los conjuntos de datos.

Luego de pre-procesados cada uno de los conjuntos de datos, estos fueron divididos en tres subconjuntos: entrenamiento, prueba y validación externa. En una primera etapa, se dividieron en subconjuntos de modelación y validación externa. Para la mayoría de los conjuntos de datos se tomaron las particiones originalmente hechas por los autores de las publicaciones de donde proceden. Solamente fue necesario particionar los conjuntos de datos TRPV1 y ci700443v. Para ello, se seleccionó aleatoriamente el 80% del conjunto original para formar los subconjuntos de modelación y el restante 20% para validación externa. En el caso particular del conjunto de datos ci700443v, debido a su gran tamaño, fue reducido a 636 compuestos antes de este procedimiento. Después, los subconjuntos de modelación fueron particionados en subconjuntos de entrenamiento y prueba, para lo cual se utilizaron en conjunto los tres algoritmos de exclusión de esferas reportados por Golbraikh y colaboradores [54]. Estos algoritmos garantizan que las moléculas presentes en el subconjunto de prueba sean representativas del espacio químico de partida. En la Tabla 2.8 se muestra un resumen de las distribuciones finales para los tres subconjuntos generados por cada conjunto de datos curado.

Tabla 2.8 Distribuciones finales para los tres subconjuntos generados por cada conjunto de datos curado.

	Número de compuestos por subconjuntos de datos (Activos/Inactivos)		
Conjunto de datos	**Entrenamiento**	**Prueba**	**Validación externa**
bzr	139(74/65)	39(19/20)	125(63/62)
ci700443v	399(228/171)	109(64/45)	128(73/55)
cox2	133(62/71)	45(25/20)	125(61/64)
dhfr	175(61/114)	58(23/35)	160(43/117)
minf00320555	244(127/117)	76(38/38)	81(34/47)

2.2.4. Detección y eliminación de los *activity cliffs* más influyentes.

Para llevar a cabo la detección y eliminación de los *activity cliffs* más influyentes se emplearon dos variantes del paso número cuatro del algoritmo mencionado en el sub-epígrafe 2.1.2. En una primera variante, para la detección de los *activity cliffs* solamente se empleó la matriz de distancia obtenida a partir de los valores calculados para los *fingerprints* del tipo ECFP. Esta variante se incluye en la experimentación ya que se basa en criterios reportados en la literatura. La segunda variante consistió en

agregar las matrices de distancia mediante la media geométrica de los valores de distancia obtenidos para cada par único de compuestos. En este experimento se agregaron los valores de las matrices de distancia por media geométrica debido a que esta función de agregación garantiza capturar en mayor medida la similitud que las diferencias existentes entre los pares de moléculas.

2.2.5. Cálculo de los descriptores moleculares

Para calcular los descriptores moleculares hubo que ajustar las configuraciones de los comandos empleados. En determinadas configuraciones (longitudes de fragmentos por encima de 6), el número de descriptores generados era computacionalmente prohibitivo (~10^6). Por esto, la longitud de los fragmentos calculados se estableció entre 2 y 6, valores con los cuales el número de descriptores era procesable. Los tipos de fragmentación a calcular (opción *-t*) empleadas fueron la 3, la 8, la 9 y la 10. En la Tabla 2.9 se muestra el tipo de fragmento que corresponde a cada opción elegida.

Tabla 2.9 Opciones elegidas y tipos de fragmentos correspondientes a estas.

Opción (*-t*)	**Descripción del fragmento**
3	Secuencias de átomos y enlaces.
8	Fragmentos centrados en átomos basándose en secuencias de átomos.
9	Fragmentos centrados en átomos basándose en secuencias de enlaces.
10	Fragmentos centrados en átomos basándose en secuencias de átomos y enlaces.

Una vez calculados los descriptores, fueron removidos los que tenían más del 95% de sus valores igual a cero o constantes. Este proceder constituye el primer filtro aplicado en la reducción de dimensionalidad y busca eliminar los descriptores con poca varianza. A partir de los descriptores obtenidos para los subconjuntos de entrenamiento originales (sin remover *activity cliffs*), se generaron las plantillas con las cuales, se calcularon los descriptores para el resto de los subconjuntos existentes.

2.3. Descripción de la modelación QSAR.

La modelación QSAR se llevó a cabo empleando un *framework* programado en MATLAB [74], el cual fue previamente desarrollado y validado en [75] por el Grupo de Investigación de Simulaciones Moleculares y Diseño de Fármacos del Centro de Bioactivos Químicos. Dicho *framework* se basa en una combinación de tres diferentes

métodos de selección de atributos e igual número de métodos de clasificación, para un total de nueve clasificadores por cada conjunto de datos.

En la selección de atributos fueron empleados tres algoritmos con diferencias conceptuales entre sí, ellos fueron: Algoritmo Genético, *Bagged Trees y Feature Ranking.* Siguiendo el mismo principio de garantizar diversidad, se seleccionaron los tres métodos de clasificación: *Adaboost Ensemble,* Análisis Discriminante Lineal y *Least Squares Support Vector Machines*. La combinación de estos métodos de selección de atributos y clasificación permite la obtención de un conjunto de nueve clasificadores que pueden garantizar la diversidad necesaria para evaluar el desempeño del procedimiento propuesto en el Epígrafe 2.1. Los parámetros de configuración de cada uno de estos métodos de selección de atributos y clasificación se describen en los próximos subepígrafes.

2.3.1. Configuraciones de los métodos de selección de atributos.

Algoritmos genéticos.

Los AG fueron corridos 10 veces, cada vez con una población inicial diferente de 100 individuos y por 250 generaciones. Los cromosomas fueron codificados empleando codificación binaria. Donde el valor 1 en la i-ésima posición significa que el atributo i-ésimo será considerado por la función de ajuste para obtener los modelos, mientras que el valor 0 representa lo contrario. La población inicial es generada aleatoriamente, de manera tal que cada individuo contenga como máximo 10 genes codificados con valor 1.

La función de selección consiste en un torneo de tamaño 2, mientras el operador de cruzamiento combina aleatoriamente cada posición de dos individuos padre para obtener la descendencia. En esencia, la selección por torneo consistió en la selección aleatoria de 2 individuos de la población, de los cuales se seleccionó el de mejor desempeño como el ganador. En el proceso de mutación se seleccionan aleatoriamente dos posiciones de un individuo padre de valores 1 y 0, cuyos valores fueron cambiados a 0 y 1 respectivamente, para generar un nuevo individuo. Las probabilidades de cruce y mutación fueron establecidas en 0.7 y 0.3 respectivamente, y las dos mejores soluciones o individuos automáticamente sobreviven y pasan a la

próxima generación. El problema del sobreajuste fue controlado de dos maneras dentro de la función de evaluación. La primera consistió en emplear un esquema de validación cruzada con 5 particiones del conjunto de entrenamiento, y la segunda, minimizando el Índice de Akaike [76], lo cual tiene como objeto mantener un balance entre el ajuste del modelo y su complejidad.

Bagged Trees.

Durante la selección de atributos por este método, fue entrenado un ensemble de 100 árboles de clasificación, donde cada árbol fue construido usando una muestra de partida independiente del conjunto de entrenamiento. El error de predicción del ensemble fue determinado computando las predicciones de cada árbol sobre los casos del conjunto de entrenamiento con los que no entrenó. Se usó el voto por mayoría para asignar cada predicción de observación y luego comparar estas predicciones con las etiquetas de clase observadas. Para construir los árboles de decisión se establecieron como mínimo una observación por hoja y un subconjunto de atributos igual en número a la raíz cuadrada del número total de atributos fue considerado para cada partición.

El criterio de partición de nodos empleado fue el índice de diversidad de Gini's [77], el cual mide las impurezas de los nodos basándose en la fracción de muestras de cada clase que alcanza ese nodo.

Para realizar la selección de atributos, fueron evaluados varios ensembles, donde el ensemble número i es aquel compuesto por los árboles de clasificación $1, 2, \dots, i$ del total de 100 obtenidos inicialmente. En la evaluación del ensemble óptimo, se emplearon: la exactitud de la clasificación sobre los subconjuntos de entrenamiento, la exactitud de la clasificación sobre los subconjuntos de prueba, y el error de predicción sobre los casos no utilizados en el entrenamiento del ensemble. Una vez determinado el ensemble óptimo, se establece una puntuación para cada atributo. Esta puntuación se basa en el error de predicción al permutar los valores para el atributo en cuestión sobre todas las instancias no empleadas en la construcción del ensemble. Después se seleccionan los 25 mejores atributos.

Feature Ranking.

Para el empleo de la estrategia de selección de atributos por puntuación, fueron seleccionados los primeros 25 atributos basándose en la puntuación derivada de la Prueba de Suma de Rangos de Wilcoxon.

2.3.2. Configuraciones de los métodos de clasificación.

Adaboost Ensembles.

Los ensembles obtenidos por este método fueron entrenados a partir de modelos de LDA entrenados con un solo atributo. Por esta razón, fue entrenado un modelo de LDA por cada atributo utilizando el conjunto de datos de entrenamiento. Luego fueron removidos los modelos de LDA con valores de exactitud menores a 0.5 para el conjunto de datos de entrenamiento.

En cada iteración se minimizó el error ponderado sobre el conjunto de entrenamiento y se retornó un ensemble de modelos de LDA entrenados con un solo atributo. El error ponderado para este ensemble fue calculado y utilizado para actualizar los pesos de cada muestra de entrenamiento.

Análisis Discriminante Lineal.

En la construcción de clasificadores por este método se utilizó la implementación existente en MATLAB [74].

Least Squares Support Vector Machines (LS-SVM).

Este método es una reformulación de las máquinas de soporte vectorial estándar. Para la construcción de los modelos se utilizó el módulo *LS-SVMlab* de MATLAB [78]. Cada atributo fue escalado al intervalo [0,1] y la RBF núcleo empleada fue la definida por: $K(x, x_k) = e^{\frac{-\|x-x_k\|^2}{2\sigma^2}}$. Los parámetros de regularización (γ) y el parámetro σ^2 de la RBF de los modelos fueron optimizados. Para ello fue utilizada una combinación de Recocido Simulado Acoplado (RSA) y la búsqueda por mallas, como está implementado en el módulo de MATLAB antes mencionado. La optimización fue guiada por las tasas de errores de clasificación de los modelos obtenidos por validación cruzada empleando los conjuntos de entrenamiento fraccionados en 5 partes iguales.

El algoritmo de RSA, inicialmente, fue usado para encontrar buenos valores de partida de los parámetros de la RBF núcleo, y luego la estrategia de búsqueda por mallas fue aplicada para el ajuste de estos parámetros.

En los casos de los modelos construidos por la combinación de este método con AG, los parámetros γ y σ^2 fueron optimizados utilizando los conjuntos de entrenamiento completos siguiendo el mismo procedimiento explicado anteriormente.

2.3.3. Establecimiento de los dominios de aplicación de los modelos.

Los dominios de aplicación de los modelos fueron establecidos a partir de los valores máximos y mínimos de los descriptores correspondientes. De esta manera, serán considerados dentro del dominio de aplicación aquellos compuestos cuyos valores de los descriptores estén dentro del rango de los valores máximo y mínimo de dichos descriptores en sus correspondientes subconjuntos de entrenamiento.

2.4. Conclusiones del capítulo

En este capítulo fueron descritos los procedimientos y métodos empleados en la experimentación realizada. Se definió la propuesta del procedimiento previo a la modelación, la cual incluyó el algoritmo de eliminación de *activity cliffs*. Se caracterizaron los conjuntos de datos escogidos para la modelación y se describieron los resultados del pre-procesamiento estructural de estos. Finalmente, se describieron los métodos computacionales empleados en la modelación, así como sus configuraciones.

CAPÍTULO 3. RESULTADOS Y DISCUSIÓN.

En el presente capítulo se exponen y discuten los resultados obtenidos durante la experimentación. En la primera parte se evalúa el efecto de la eliminación de los *activity cliffs* sobre la modelabilidad y la continuidad de las SAR en conjuntos de datos, así como en el proceso de entrenamiento y validación de los modelos. En la segunda parte se evalúa la capacidad predictiva de los modelos obtenidos sobre los conjuntos de validación externa.

Consideraciones previas al análisis.

Para la evaluación de la clasificación de los modelos obtenidos se tuvieron en cuenta cuatro parámetros:

(i) la exactitud: razón entre el número de casos correctamente clasificados y el total;
(ii) la sensibilidad: razón entre el número de casos de la clase de interés (clase = 1 en nuestro caso) correctamente clasificados y el total de compuestos que forman la clase;
(iii) la especificidad: razón entre el número de casos de la otra clase (clase = 0) correctamente clasificados y el total de compuestos que forman la clase;
(iv) el promedio de los valores de estos dos últimos.

Mientras que la exactitud resulta poco adecuada en casos donde las clases están desbalanceadas, el cuarto parámetro mencionado da una evaluación aproximada del balance logrado en la clasificación, motivo por el cual se utilizó en esta investigación.

Para comparar los resultados para las dos variantes de eliminación de *activity cliffs* con los obtenidos para los conjuntos de datos originales, se empleó la Prueba de los Rangos con Signo de Wilcoxon de una sola cola. Esta es una prueba no paramétrica que permite comparar las medianas de dos muestras relacionadas y establecer si una de estas es significativamente superior a la otra. Los valores de las significaciones para esta prueba fueron calculados utilizando el SPSS [79] y en todas las pruebas realizadas se estableció el nivel de significación igual 0,05.

3.1. Influencia de la eliminación de los activity cliffs en el proceso de entrenamiento de los modelos.

En la literatura se reportan diferentes métricas que permiten evaluar la influencia de la eliminación de los *activity cliffs* sobre los conjuntos de entrenamiento previo a la modelación. En el caso de esta investigación se seleccionaron dos de ellas. La primera es el SARI, la cual fue mencionada en el Subepígrafe 1.1.1, y brinda una medida de la naturaleza de las SAR (continua, discontinua o heterogénea). La segunda es el índice de modelabilidad (MODI por sus siglas en inglés, *Modelability Index*), reportado por Golbraikh y colaboradores en [80], la cual da una medida de la posibilidad de obtener modelos predictivos (exactitud > 0.7) a partir de un conjunto de datos dado. El análisis de los resultados basados en estas métricas será tratado en los Subepígrafes 3.1.2 y 3.1.3 respectivamente.

Una vez construidos los modelos, estos se someten a un esquema de validación que permite la selección de los más óptimos. En los Subepígrafes 3.1.4 y 3.1.5, se discuten los resultados obtenidos en este proceso.

3.1.1. Eliminación de los *activity cliffs* de los conjuntos de entrenamiento.

En la Tabla 3.1 se muestra el resumen de la eliminación de los *activity cliffs* utilizando dos vías. En la primera se tuvo en cuenta la similitud de Tanimoto y se emplearon solamente *fingerprints* del tipo ECFP. En la segunda se fusionaron las matrices de similitud obtenidas a partir de diferentes *fingerprints* mediante la media geométrica de los valores.

Tabla 3.1 Resumen de la eliminación de los *activity cliffs* para los conjuntos de entrenamiento empleados.

		Número de *activity cliffs* eliminados			
Conjuntos de datos	**Número inicial de compuestos**	**ECFP**	**%**	**GeoMean**	**%**
bzr	139	6	4,3	4	2,9
ci700443v	399	12	3,0	8	2,0
cox2	133	27	20,3	19	14,3
dhfr	175	4	2,3	3	1,7
minf00320555	244	29	11,9	15	6,1

Nótese que en todos los casos, por la primera vía, la cantidad de compuestos eliminados fue mayor, y que para el subconjunto de entrenamiento del conjunto de

datos con identificador “cox2”, el número de *activity cliffs* removidos fue superior por esta vía, llegando a ser alrededor del 20% del total inicial de compuestos.

3.1.2. Continuidad de los espacios de estructura-actividad (SARI).

Los valores del SARI fueron calculados para los conjuntos de datos de entrenamiento antes y después de aplicar los algoritmos de eliminación de *activity cliffs*, para ello se utilizó el procedimiento propuesto en [20]. En la Tabla 3.2 se resumen los resultados obtenidos para este parámetro, así como los coeficientes de correlación de Pearson entre los cambios en el SARI ($SARI_{\text{sin } activity\ cliffs} - SARI_{\text{con } activity\ cliffs}$) y los porcientos de *activity cliffs* eliminados (ECFP% y GeoMean%) para cada caso.

Tabla 3.2 Valores del SARI para los subconjuntos de entrenamiento antes (Original) y después de aplicar los algoritmos de eliminación de *activity cliffs* (ECFP y GeoMean).

Conjuntos de datos	Original	ECFP	Dif.	ECFP%	GeoMean	Dif.	GeoMean%
bzr	0,2090	0,2083	-0,0007	4,32	0,2097	0,0007	2,88
ci700443v	0,3250	0,3169	-0,0081	3,01	0,3199	-0,0051	2,01
cox2	0,1227	0,1285	**0,0058**	20,30	0,1267	**0,0040**	14,29
dhfr	0,1559	0,1561	0,0002	2,29	0,1564	0,0005	1,71
minf00320555	0,2014	0,2025	**0,0011**	11,89	0,2030	0,0016	6,15
Mean	0,2028	0,2025	-0,0003		0,2031	0,0003	
p-value		0,500			0,313		
Pearson		**0,766**			**0,715**		

En las últimas tres filas de la Tabla 3.3 se muestran las medias (*Mean*), las significaciones ($p - value$) para la Prueba de los Rangos con Signo de Wilcoxon de una sola cola y los coeficientes de correlación de Pearson. Para todos los casos en la tabla anterior, los valores del SARI reflejan que las SAR de los subconjuntos de entrenamiento modelados se caracterizan por ser discontinuas, como se sugiere en [20] para valores de estas magnitudes. Por otra parte, aunque eliminar los *activity cliffs* empleando el algoritmo que realiza agregación de las matrices de similitud por media geométrica mejora los valores de este índice, no lo hace significativamente ($p - value = 0{,}313$). En el otro caso, se puede apreciar que los valores del SARI no mejoraron para los subconjuntos de entrenamiento procesados, aunque no fueron significativamente inferiores ($p - value = 0{,}500$) a los valores de los subconjuntos originales.

En ambos casos, los valores de los coeficientes de correlación de Pearson obtenidos fueron estadísticamente significativos y positivos. Este resultado indica que los cambios en los valores del SARI están significativamente correlacionados con los porcientos de casos eliminados como *activity cliffs*. Adicionalmente, las magnitudes de las mejoras en el SARI, para cada algoritmo en particular, fueron superiores al resto en aquellos casos donde se eliminó más del 10% de instancias como *activity cliffs* (ver valores resaltados en negrita en la Tabla 3.2). Esto está en correspondencia con el trabajo de Smith y Martínez [31], donde se plantea que las mejoras son significativas cuando se remueve más del 10% de los casos de la data original.

Resumiendo, se puede afirmar que en las condiciones de este experimento, no se restauró de manera significativa la continuidad de las SAR, aunque existe elevada correlación entre las mejorías en este parámetro y los porcientos de casos eliminados como *activity cliffs*.

3.1.3. Modelabilidad de los conjuntos de entrenamiento (MODI).

El MODI fue calculado para los conjuntos de datos de entrenamiento antes y después de aplicar los algoritmos de eliminación de *activity cliffs*, para ello se utilizó el procedimiento propuesto en [80]. En la Tabla 3.3 se resumen los resultados obtenidos para este parámetro.

Tabla 3.3 Valores del MODI para los subconjuntos de entrenamiento antes (Original) y después de aplicar los algoritmos de eliminación de *activity cliffs* (ECFP y GeoMean).

Conjuntos de datos	**Original**	**ECFP**	**Dif.**	**ECFP%**	**GeoMean**	**Dif.**	**GeoMean%**
bzr	0,6805	0,7000	0,0195	4,32	0,6830	0,0025	2,88
ci700443v	0,7887	0,7871	-0,0016	3,01	0,7889	0,0002	2,01
cox2	0,5892	0,6311	0,0419	20,30	0,6667	0,0775	14,29
dhfr	0,6969	0,6924	-0,0045	2,29	0,7108	0,0139	1,71
minf00320555	0,6903	0,7537	0,0634	11,89	0,7235	0,0332	6,15
Mean	0,6891	0,7129	0,0237		0,7146	0,0255	
p-value		0,156			**0,031**		
Pearson		**0,973**			**0,7630**		

En las últimas tres filas de la Tabla 3.3 se muestran las medias (*Mean*), las significaciones ($p - value$) para la Prueba de los Rangos con Signo de Wilcoxon de una sola cola y los coeficientes de correlación de Pearson. En [80] Golbraikh y

colaboradores determinaron 0.65 como valor de corte para considerar modelable o no un conjunto de datos. Los valores del MODI para los subconjuntos de entrenamiento originales fueron superiores a este valor de corte, excepto para el de identificador "cox2" (ver valores subrayados en la Tabla 3.3). Este resultado fue mejorado considerablemente siguiendo el procedimiento de eliminación de *activity cliffs* empleando el algoritmo etiquetado como GeoMean, para el cual se sobrepasó el valor de corte mencionado.

En la Tabla 3.3 se puede apreciar que la eliminación de los *activity cliffs*, en las condiciones de este experimento, condujo a mejorar la modelabilidad de los conjuntos de entrenamiento procesados. Los valores de significación ($p - value$) muestran que esta mejoría solamente fue significativa ($p - value = 0{,}031$) para el procedimiento de eliminación de *activity cliffs* empleando el algoritmo etiquetado como GeoMean. Al igual que en el caso de los valores obtenidos para el SARI, los coeficientes de correlación de Pearson de la Tabla 3.3 fueron estadísticamente significativos y muestran que existe elevada correlación entre las mejorías en este parámetro y los porcientos de casos eliminados como *activity cliffs*.

3.1.4. Evaluación de la clasificación sobre los conjuntos de entrenamiento.

Los resultados de la clasificación sobre los conjuntos de entrenamiento fueron resumidos en las Tablas A1.1 y A1.2 del Anexo 1. En la primera se muestran las evaluaciones globales obtenidas de los cuatro parámetros antes mencionados para: los conjuntos de entrenamiento originales (sin previa eliminación de *activity cliffs*) y las dos variantes de los conjuntos de entrenamiento a los cuales se les eliminaron los *activity cliffs* (ECFP y GeoMean). En la segunda se resumen, a modo de comparación, los resultados globales de la clasificación para la clase peor clasificada por los modelos obtenidos para los conjuntos de entrenamiento originales y las otras dos variantes mencionadas. Al final de ambas tablas se muestran los resultados de la Prueba de los Rangos con Signo de Wilcoxon de una sola cola para cada parámetro de evaluación. Los valores de significación mostrados se corresponden con las comparaciones por pares de los desempeños de los modelos obtenidos para las variantes ECFP y GeoMean, con los de los modelos de los conjuntos de entrenamiento originales.

En la Tabla A1.1 se puede apreciar que la eliminación de los *activity cliffs*, por ambas vías, mejora significativamente la exactitud, la sensibilidad, y el promedio entre los valores de la sensibilidad y la especificidad. Por otra parte, no mejora los resultados obtenidos en la especificidad, aunque no resulta significativamente menor para ambos casos.

Al final de la Tabla A1.2 se puede apreciar que la eliminación de los *activity cliffs* por ambas vías, mejora significativamente el desempeño de los modelos en la clasificación de la clase peor clasificada por los modelos obtenidos de los conjuntos de entrenamiento originales. En otras palabras: con la eliminación de los *activity cliffs* se logra balancear la clasificación de los conjuntos de entrenamiento.

3.1.5. Validación estadística de los modelos.

Para la validación estadística de los modelos se emplearon tres variantes: la validación cruzada dejando uno fuera (LOO, por sus siglas en inglés, *Leave One Out*), la validación cruzada usando 5 particiones de los conjuntos de datos y el re-muestreo o *bootstrap*.

En la Tabla A1.3 del Anexo 1 se muestran los resultados obtenidos para estas tres variantes. Al final de esta tabla se resumieron los resultados de la Prueba de los Rangos con Signo de Wilcoxon de una sola cola para cada método de validación. De la comparación establecida por esta prueba estadística se puede inferir que la eliminación de *activity cliffs* mejora la estabilidad de los modelos para todas las variantes, excepto para la validación cruzada usando 5 particiones de los conjuntos obtenidos (ver valor de significación subrayado) para el algoritmo identificado por ECFP, donde fue significativamente menor el desempeño.

3.1.6. Evaluación de la clasificación sobre los conjuntos de prueba.

En el caso de los subconjuntos de prueba, los resultados fueron muy similares a los obtenidos para los de entrenamiento. Para estos, de igual manera (ver la Tabla A2.1 del Anexo 2), fue significativa la mejora en la exactitud, la sensibilidad, y el promedio entre los valores de la sensibilidad y la especificidad. Los resultados para la

especificidad muestran que no existen diferencias significativas entre eliminar o no los *activity cliffs*.

En este caso, la eliminación de los *activity cliffs* por ambas vías, también mejora significativamente el desempeño de los modelos obtenidos de los conjuntos de entrenamiento originales en la clasificación de la clase peor clasificada por estos (ver la Tabla A2.2 del Anexo 2).

La similitud entre los resultados para estos subconjuntos de datos y los obtenidos para los subconjuntos de entrenamiento no es casualidad. Estos son los más parecidos a sus respectivos subconjuntos de entrenamiento, hecho que garantiza la forma en que se particionaron los subconjuntos de modelación (ver Subepígrafe 2.2.3). Entonces, es de esperar que el desempeño de los modelos no difiera mucho para ambos subconjuntos.

3.2. Influencia de la eliminación de los *activity cliffs* sobre la clasificación de los conjuntos de validación externa.

Debido a la selección aleatoria de las moléculas/instancias presentes en los subconjuntos de validación externa y que estos no participan en la modelación, dichos subconjuntos son los que ponen a prueba la capacidad predictiva de los modelos QSAR obtenidos. Generalmente para estos subconjuntos no se espera que los modelos tengan desempeños superiores a los de entrenamiento y prueba.

En las condiciones de esta experimentación, ninguna de las dos variantes de eliminación de *activity cliffs* condujo a mejoras significativas en ninguno de los cuatro parámetros de evaluación, aunque tampoco fueron significativamente inferiores (ver la Tabla A3.1 del Anexo 3). Por otra parte, para ambas variantes, se mejoró la capacidad de los modelos de clasificar correctamente la clase peor predicha por los modelos obtenidos de los conjuntos de datos originales, siendo estadísticamente significativo este resultado para el algoritmo de eliminación de *activity cliffs* etiquetado como ECFP (ver la Tabla A3.2 del Anexo 3).

Este resultado tiene una influencia muy positiva pues alivia uno de los problemas de los algoritmos de aprendizaje automatizado ante el desbalance de las clases, el cual

es precisamente el sesgo de estos al clasificar mejor la clase más representada. El otro efecto positivo de este resultado lo constituye el uso de estos modelos como herramientas de *screening virtual*. En este tipo de tareas se necesita que el clasificador tenga un desempeño adecuado en ambas clases (o sea, que tenga un desempeño balanceado) pues esta herramienta debe ser capaz tanto de ordenar al inicio los compuestos de la clase de interés, como relegar al final de la lista los de la clase contraria.

3.3. Conclusiones del capítulo

De los resultados expuestos en este capítulo, y en las condiciones en las que se realizó la experimentación, se puede concluir que, desde el punto de vista estadístico, las vías de eliminación de los *activity cliffs*:

- Aunque no condujeron a un cambio estadísticamente significativo de la continuidad de las Relaciones Estructura-Actividad, al aplicarlos, se mejoró significativamente la modelabilidad de los conjuntos de entrenamiento procesados; específicamente empleando el algoritmo que realiza agregación de las matrices de similitud por media geométrica (GeoMean).
- Condujeron a mejoras significativas en el proceso de entrenamiento y validación de los modelos.
- Aunque no favorecieron significativamente la clasificación de los subconjuntos de validación externa, de manera general, se mejoró la clasificación de la clase peor clasificada por los subconjuntos de entrenamiento originales, siendo este resultado estadísticamente significativo para el algoritmo de eliminación de *activity cliffs* que no realiza fusión de matrices de similitud, lo que muestra una tendencia a balancear la clasificación.

CONCLUSIONES

Luego de finalizada la investigación y teniendo en cuenta los resultados experimentales obtenidos, se arribó a las siguientes conclusiones:

- Los *activity cliffs* tienen un efecto negativo sobre la capacidad predictiva de los modelos QSAR. Sin embargo, hasta el alcance de la búsqueda realizada, no existen reportes que evidencien la incidencia de remover los *activity cliffs* en la obtención de modelos QSAR.
- Se definió un procedimiento previo a la modelación, el cual incluyó el uso de las herramientas para el pre-procesamiento de las estructuras moleculares, y el algoritmo de eliminación de *activity cliffs* en sus dos variantes.
- Para la evaluación de la hipótesis planteada, se contó con cinco conjuntos de datos y un total de nueve clasificadores basados en la combinación de tres métodos de selección de atributos y tres métodos de clasificación. Luego de llevar a cabo la eliminación de los *activity cliffs*, se pudo comprobar que, desde el punto de vista estadístico:
  (i) Aunque no condujo a un cambio significativo de la continuidad de las SAR, se mejoró significativamente la modelabilidad de los conjuntos de entrenamiento procesados; específicamente empleando el algoritmo que realiza agregación de las matrices de similitud por media geométrica.
  (ii) Condujo a mejoras significativas en el proceso de entrenamiento y validación de los modelos.
  (iii) No favoreció significativamente la clasificación de los subconjuntos de validación externa. Sin embargo, de manera general, mejoró la clasificación de la clase peor clasificada por los modelos obtenidos de los subconjuntos de entrenamiento originales, mostrando una tendencia a balancear la clasificación. Este resultado fue significativo para el algoritmo de eliminación de *activity cliffs* que no realiza fusión de matrices de similitud.

RECOMENDACIONES

La presente investigación apenas es un comienzo de un vasto estudio sobre la factibilidad de eliminar los *activity cliffs*. En la experimentación realizada apenas se exploraron algunas de las variantes posibles. Se recomienda:

(i) continuar dicha exploración con una mayor combinación de variantes;
(ii) probar nuevas formas de agregación de matrices, así como valores de corte de similitud y de potencia;
(iii) estudiar la influencia de eliminar *activity cliffs* sobre multiclasificadores y;
(iv) estudiar la influencia de eliminar *activity cliffs* en tareas de *screening virtual*.

REFERENCIAS BIBLIOGRÁFICAS

[1] D. Fourches, E. Muratov, and A. Tropsha, "Trust, But Verify: On the Importance of Chemical Structure Curation in Cheminformatics and QSAR Modeling Research," *J. Chem. Inf. Model,* vol. 50, pp. 1189–1204, 2010.

[2] G. M. Maggiora, "On Outliers and Activity Cliffss Why QSAR Often Disappoints," *J. Chem. Inf. Model.,* vol. 46, p. 1535, 07/24/2006 2006.

[3] J. Rose, "Methods for Data Analysis," in *Handbook of Chemoinformatics*. vol. 3, J. Gasteier, Ed., ed Weinheim: WILEY-VCH, 2003, pp. 1081-1097.

[4] F. Bajot, "The Use of Qsar and Computational Methods in Drug Design," in *Recent Advances in QSAR Studies*. vol. 8, T. Puzyn, J. Leszczynski, and M. T. Cronin, Eds., ed: Springer Netherlands, 2010, pp. 261-282.

[5] R. Cramer, "The inevitable QSAR renaissance," *Journal of Computer-Aided Molecular Design,* vol. 26, pp. 35-38, 2012/01/01 2012.

[6] V. Kuz'min, A. G. Artemenko, E. Muratov, P. G. Polischuk, L. N. Ognichenko, A. V. Liahovsky*, et al.*, "Virtual Screening and Molecular Design Based on Hierarchical Qsar Technology," in *Recent Advances in QSAR Studies*. vol. 8, T. Puzyn, J. Leszczynski, and M. T. Cronin, Eds., ed: Springer Netherlands, 2010, pp. 127-176.

[7] F. L. Stahura and J. Bajorath, "Virtual Screening Methods that Complement HTS," vol. 7, pp. 259-269, 2004.

[8] A. Tropsha, "Best Practices for QSAR Model Development, Validation, and Exploitation," *Molecular Informatics,* vol. 29, pp. 476 – 488, 2010.

[9] J. L. Medina-Franco, A. B. Yongye, and F. López-Vallejo, "Consensus Models of Activity Landscapes," in *Statistical Modelling of Molecular Descriptors in QSAR/QSPR*, ed: Wiley-VCH Verlag GmbH & Co. KGaA, 2012, pp. 307-326.

[10] Y. Hu and J. Bajorath, "Activity profile relationships between structurally similar promiscuous compounds," *European Journal of Medicinal Chemistry,* vol. 69, pp. 393-398, 11// 2013.

[11] Y. Hu and J. Bajorath, "Extending the Activity Cliff Concept: Structural Categorization of Activity Cliffs and Systematic Identification of Different Types

of Cliffs in the ChEMBL Database," *Journal of Chemical Information and Modeling,* vol. 52, pp. 1806-1811, 2012/07/23 2012.

[12] D. Stumpfe and J. Bajorath, "Exploring Activity Cliffs in Medicinal Chemistry," *Journal of Medicinal Chemistry,* vol. 55, pp. 2932-2942, 2012.

[13] D. Stumpfe, Y. Hu, D. Dimova, and J. Bajorath, "Recent Progress in Understanding Activity Cliffs and Their Utility in Medicinal Chemistry," *Journal of Medicinal Chemistry,* vol. 57, pp. 18-28, 2014/01/09 2013.

[14] J. L. Medina-Franco, "Activity Cliffs: Facts or Artifacts?," *Chem Biol Drug Des,* vol. 81, pp. 553–556, 2013.

[15] D. Horvath, "Quantitative Structure-Activity Relantionships: In Silico Chemistry Or High Tech Alchemy? ," *Rev. Roum. Chim.,* vol. 55, pp. 783-801, 2010.

[16] M. Cruz-Monteagudo, J. L. Medina-Franco, Y. Pérez-Castillo, O. Nicolotti, M. N. D. S. Cordeiro, and F. Borges, "Activity cliffs in drug discovery: Dr Jekyll or Mr Hyde?," *Drug Discovery Today,* vol. 19, pp. 1069-1080, 8// 2014.

[17] L. Peltason, P. Iyer, and J. r. Bajorath, "Rationalizing Three-Dimensional Activity Landscapes and the Influence of Molecular Representations on Landscape Topology and the Formation of Activity Cliffs," *Journal of Chemical Information and Modeling,* vol. 50, pp. 1021-1033, 2010/06/28 2010.

[18] L. P. Jürgen Bajorath, Mathias Wawer, Rajarshi Guha, Michael S. Lajiness, John H. Van Drie, "Navigating structure–activity landscapes," *Drug Discovery Today,* vol. 14, pp. 698-705, 2009.

[19] M. A. M. Johnson, Gerald M., "Concepts and applications of molecular similarity," *Journal of Computational Chemistry,* vol. 13, pp. 539-540, 1990.

[20] L. Peltason and J. Bajorath, "SAR Index: Quantifying the Nature of Structure−Activity Relationships," *Journal of Medicinal Chemistry,* vol. 50, pp. 5571-5578, 2007/11/01 2007.

[21] R. Guha and J. H. Van Drie, "Structure−Activity Landscape Index: Identifying and Quantifying Activity Cliffs," *Journal of Chemical Information and Modeling,* vol. 48, pp. 646-658, 2008/03/01 2008.

[22] J. L. Medina-Franco, K. Martínez-Mayorga, A. Bender, R. M. Marín, M. A. Giulianotti, C. Pinilla*, et al.*, "Characterization of Activity Landscapes Using 2D

and 3D Similarity Methods: Consensus Activity Cliffs," *Journal of Chemical Information and Modeling,* vol. 49, pp. 477-491, 2009/02/23 2009.

[23] A. M. Wassermann, M. Wawer, and J. r. Bajorath, "Activity Landscape Representations for Structure−Activity Relationship Analysis," *Journal of Medicinal Chemistry,* vol. 53, pp. 8209-8223, 2010/12/09 2010.

[24] X. Hu, Y. Hu, M. Vogt, D. Stumpfe, and J. Bajorath, "MMP-Cliffs: Systematic Identification of Activity Cliffs on the Basis of Matched Molecular Pairs," *Journal of Chemical Information and Modeling,* vol. 52, pp. 1138-1145, 2012/05/25 2012.

[25] R. Guha and J. H. Van Drie, "Assessing How Well a Modeling Protocol Captures a Structure−Activity Landscape," *Journal of Chemical Information and Modeling,* vol. 48, pp. 1716-1728, 2008/08/01 2008.

[26] F. Ruggiu, G. Marcou, A. Varnek, and D. Horvath, "ISIDA Property-Labelled Fragment Descriptors," *Molecular Informatics,* vol. 29, pp. 855-868, 2010.

[27] J. L. Durant, B. A. Leland, D. R. Henry, and J. G. Nourse, "Reoptimization of MDL Keys for Use in Drug Discovery," *Journal of Chemical Information and Computer Sciences,* vol. 42, pp. 1273-1280, 2002/11/01 2002.

[28] D. Rogers and M. Hahn, "Extended-Connectivity Fingerprints," *Journal of Chemical Information and Modeling,* vol. 50, pp. 742-754, 2010/05/24 2010.

[29] D. Stumpfe and J. Bajorath, "Similarity searching," *Wiley Interdisciplinary Reviews: Computational Molecular Science,* vol. 1, pp. 260-282, 2011.

[30] D. Stumpfe, D. Dimova, K. Heikamp, and J. Bajorath, "Compound Pathway Model To Capture SAR Progression: Comparison of Activity Cliff-Dependent and -Independent Pathways," *Journal of Chemical Information and Modeling,* vol. 53, pp. 1067-1072, 2013/05/24 2013.

[31] T. M. Michael R. Smith, "Improving classification accuracy by identifying and removing instances that should be misclassified," presented at the The 2011 International Joint Conference on Neural Networks (IJCNN) San Jose, CA, 2011.

[32] B. Byeon, K. Rasheed, and P. Doshi, "Enhancing the Quality of Noisy Training Data Using a Genetic Algorithm and Prototype Selection," in *International Conference on Artificial Intelligence - IC-AI*, 2008, pp. 821-827.

[33] Z. Yang and D. Gao, "Classification for Imbalanced and Overlapping Classes Using Outlier Detection and Sampling Techniques," *Appl. Math. Inf. Sci. ,* vol. 7, pp. 375-381, 2013.

[34] "Waikato Environment for Knowledge Analysis (WEKA)," 3.7.10 ed. New Zealand: University of Waikato, 2013.

[35] M. Smith, T. Martinez, and C. Giraud-Carrier, "An instance level analysis of data complexity," *Machine Learning,* vol. 95, pp. 225-256, 2014/05/01 2014.

[36] Y. Hu, D. Stumpfe, and J. Bajorath, "Advancing the activity cliff concept " *F1000Research,* 2013.

[37] O. Méndez-Lucio, J. Pérez-Villanueva, R. Castillo, and J. L. Medina-Franco, "Identifying Activity Cliff Generators of PPAR Ligands Using SAS Maps," *Molecular Informatics,* vol. 31, pp. 837-846, 2012.

[38] N. Japkowicz, "The Class Imbalance Problem: Significance and Strategies," 2000, pp. 111-117.

[39] N. Japkowicz, "Learning from Imbalanced Data Sets: A Comparison of Various Strategies," in *AAAI Press*, 2000, pp. 10-15.

[40] I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd ed. San Francisco: Morgan Kaufmann, 2005.

[41] A. Cherkasov, E. N. Muratov, D. Fourches, A. Varnek, I. I. Baskin, M. Cronin*, et al.*, "QSAR Modeling: Where Have You Been? Where Are You Going To?," *Journal of Medicinal Chemistry,* vol. 57, pp. 4977-5010, 2014/06/26 2013.

[42] P. Gramatica, S. Cassani, P. P. Roy, S. Kovarich, C. W. Yap, and E. Papa, "QSAR Modeling is not “Push a Button and Find a Correlation”: A Case Study of Toxicity of (Benzo-)triazoles on Algae," *Molecular Informatics* vol. 31, pp. 817 – 835, 2012.

[43] "cxcalc," 6.1.0.110 ed: ChemAxon, 2013.

[44] "JChem for Excel," 6.1.0.110 ed: ChemAxon, 2013.

[45] "Standardizer," 6.1.0.110 ed: ChemAxon, 2012.

[46] "Marvin," 6.1.0.110 ed: ChemAxon, 2013.

[47] N. M. OLBoyle, M. Banck, C. A. James, C. Morley, T. Vandermeersch, and G. R. Hutchison, "Open Babel: An open chemical toolbox," *J Cheminf,* vol. 3, p. 33, 2011.

[48] V. P. Solov'ev and A. Varnek, "EdiSDF," 5.03 ed, 2010, p. (Editor of the Structure Data Files).

[49] N. V. Chawla, Bowyer, K.W., L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research,* vol. 16, pp. 321–357, 2002.

[50] M. Kubat and S. Matwin, "Adressing the curse of imbalanced training sets: onesided selection " in *14th International Conference on Machine Learning*, 1997, pp. 179–186.

[51] S.-J. Yen and Y.-S. Lee, "Under-sampling approaches for improving prediction of the minority class in an imbalanced dataset," in *Intelligent Control and Automation*, ed: Springer, 2006, pp. 731-740.

[52] A. Golbraikh, X. Wang, H. Zhu, and A. Tropsha, "Predictive QSAR Modeling: Methods and Applications in Drug Discovery andChemical Risk Assessment," in *Handbook of Computational Chemistry*, J. Leszczynski, Ed., ed: Springer Netherlands, 2012, pp. 1309-1342.

[53] P. Gramatica, "Chemometric Methods and Theoretical Molecular Descriptors in Predictive QSAR Modeling of the Environmental Behavior of Organic Pollutants," in *Recent Advances in QSAR Studies*. vol. 8, T. Puzyn, J. Leszczynski, and M. T. Cronin, Eds., ed: Springer Netherlands, 2010, pp. 327-366.

[54] A. Golbraikh and A. Tropsha, "Predictive QSAR modeling based on diversity sampling of experimental datasets for the training and test set selection," *Molecular diversity,* vol. 5, pp. 231-243, 2000.

[55] A. Golbraikh, M. Shen, Z. Xiao, Y.-D. Xiao, K.-H. Lee, and A. Tropsha, "Rational selection of training and test sets for the development of validated QSAR models," *Journal of Computer-Aided Molecular Design,* vol. 17, pp. 241-253, 2003/02/01 2003.

[56] D. Ballabio and V. Consonni, "Classification tools in chemistry. Part 1: linear models. PLS-DA," *Analytical Methods,* vol. 5, pp. 3790-3798, 2013.

[57] M. Alvarez-Guerra, D. Ballabio, J. M. Amigo, R. Bro, and J. R. Viguri, "Development of models for predicting toxicity from sediment chemistry by partial least squares-discriminant analysis and counter-propagation artificial neural networks," *Environmental Pollution,* vol. 158, pp. 607-614, 2// 2010.

[58] A. Jurik, R. Reicherstorfer, B. Zdrazil, and G. F. Ecker, "Classification of High-Activity Tiagabine Analogs by Binary QSAR Modeling," *Molecular informatics,* vol. 32, pp. 415-419, 2013.

[59] A. Pérez-Garrido, A. M. Helguera, F. Borges, M. N. D. Cordeiro, V. Rivero, and A. G. Escudero, "Two new parameters based on distances in a receiver operating characteristic chart for the selection of classification models," *Journal of chemical information and modeling,* vol. 51, pp. 2746-2759, 2011.

[60] A. M. Helguera, A. Pérez-Garrido, A. Gaspar, J. Reis, F. Cagide, D. Vina*, et al.*, "Combining QSAR classification models for predictive modeling of human monoamine oxidase inhibitors," *European Journal of Medicinal Chemistry,* vol. 59, pp. 75-90, 1// 2013.

[61] G. Melagraki, A. Afantitis, H. Sarimveis, P. A. Koutentis, J. Markopoulos, and O. Igglessi-Markopoulou, "Optimization of biaryl piperidine and 4-amino-2-biarylurea MCH1 receptor antagonists using QSAR modeling, classification techniques and virtual screening," *Journal of computer-aided molecular design,* vol. 21, pp. 251-267, 2007.

[62] M. Goodarzi, B. Dejaegher, and Y. V. Heyden, "Feature selection methods in QSAR studies," *Journal of AOAC International,* vol. 95, pp. 636-651, 2012.

[63] A. Z. Dudek, T. Arodz, and J. Gálvez, "Computational Methods in Developing Quantitative Structure-Activity Relationships (QSAR): A Review," *Combinatorial Chemistry & High Throughput Screening,* vol. 9, pp. 0-16, 2006.

[64] R. D. Tobias, "An introduction to partial least squares regression," in *Proc. Ann. SAS Users Group Int. Conf., 20th, Orlando, FL*, 1995, pp. 2-5.

[65] V. G. Aguiar-Pulido, Marcos; Cruz-Monteagudo, Maykel ; Rabunal, Juan R.; Dorado, Julian ; Munteanu, Cristian R., "Evolutionary Computation and QSAR Research," *Current Computer Aided-Drug Design,* vol. 9, pp. 206-225, 2013.

[66] O. Ivanciuc, "Drug Design with Machine Learning," in *Encyclopedia of Complexity and Systems Science*, R. A. Meyers, Ed., ed: Springer New York, 2009, pp. 2159-2196.

[67] J. Gasteiger, "Methods for Data Analysis," in *Handbook of Chemoinformatics: From Data to Knowledge in 4 Volumes*, ed Weinheim, Germany: Wiley-VCH Verlag GmbH, 2003.

[68] R. Guha, "The Ups and Downs of Structure–Activity Landscapes," in *Chemoinformatics and Computational Chemical Biology*. vol. 672, J. Bajorath, Ed., ed: Humana Press, 2011, pp. 101-117.

[69] M. Sud, "MayaChemTools: An open source package for computational discovery," in *243rd ACS National Meeting & Exposition, March 25-29 2012, San Diego, CA*, 2012.

[70] G. Marcou, V. Solov'ev, D. Horvath, and A. Varnek, "ISIDA Fragmentor2011-User Manual," 2012.

[71] H. Zhu, A. Tropsha, D. Fourches, A. Varnek, E. Papa, P. Gramatica*, et al.*, "Combinatorial QSAR modeling of chemical toxicants tested against Tetrahymena pyriformis," *Journal of chemical information and modeling,* vol. 48, pp. 766-784, 2008.

[72] D. A. Tsareva and G. F. Ecker, "How far could we go with open data–a case study for TRPV1 antagonists," *Molecular informatics,* vol. 32, pp. 555-562, 2013.

[73] J. J. Sutherland, L. A. O'Brien, and D. F. Weaver, "Spline-Fitting with a Genetic Algorithm: A Method for Developing Classification Structure-Activity Relationships," *J. Chem. Inf. Comput. Sci.,* vol. 43, pp. 1906-1915, 2003.

[74] "MATLAB," R2009a ed: The MathWorks Inc., 2009.

[75] Y. Pérez-Castillo, C. Lazar, J. Taminau, M. Froeyen, M. A. n. Cabrera-Pérez, and A. Nowé, "GA(M)E-QSAR: a novel, fully automatic genetic-algorithm-(meta)-ensembles approach for binary classification in ligand-based drug

design," *Journal of chemical information and modeling,* vol. 52, pp. 2366–2386, 2012.

[76] H. Akaike, "Information theory and an extension of the maximum likelihood principle," in *Second International Symposium on Information Theory*, 1973, pp. 267-281.

[77] C. Gini, "Concentration and Dependency Ratios," *Rivista di politica economica,* vol. 87, pp. 769-790, 1997.

[78] J. A. K. Suykens, T. Van Gestel, J. De Brabanter, B. De Moor, and J. Vandewalle, *Least Squares Support Vector Machines*. Singapore: World Scientific, 2002.

[79] S. 15, "SPSS 15.0 for Windows," ed: SPSS Inc Chicago, IL, 2006.

[80] A. Golbraikh, E. Muratov, D. Fourches, and A. Tropsha, "Data Set Modelability by QSAR," *Journal of Chemical Information and Modeling,* vol. 54, pp. 1-4, 2014/01/27 2014.

ANEXOS

Leyenda:

	Algoritmos de clasificación
GALSSVM	*Genetic Algorithm-Least Square Support Vector Machines.*
GALDA	*Genetic Algorithm-Linear Discriminant Analysis.*
GAAB	*Genetic Algorithm-Adaboost Ensembles.*
BT_LSSVM	*Bagged Trees-Least Square Support Vector Machines.*
BT_LDA	*Bagged Trees-Linear Discriminant Analysis.*
BT_AB	*Bagged Trees-Adaboost Ensembles.*
FR_LSSVM	*Feature Ranking-Least Square Support Vector Machines.*
FR_LDA	*Feature Ranking-Linear Discriminant Analysis.*
FR_AB	*Feature Ranking-Adaboost Ensembles.*
	Parámetros de evaluación
Acc.	Exactitud.
Se.	Sensibilidad.
Sp.	Especificidad.
(Se.+Sp.)/2	Promedio de los valores de Se. y Sp., este da una idea del balance en la clasificación, especialmente cuando los conjuntos de datos están desbalanceados.
Worst Class	Clase peor clasificada, puede referirse a Se. o Sp., en dependencia de que clase haya sido la peor clasificada en el conjunto de entrenamiento.
p-value	Significación Estadística.
Prom.	Promedio.
	Algoritmos de eliminación de *activity cliffs*
ECFP	Algoritmo de eliminación de *activity cliffs* basado en los *fingerprints* del tipo *Extended Connectivity Fingerprints.*
GeoMean	Algoritmo de eliminación de *activity cliffs* basado en la fusión de matrices de similitud por media geométrica de los valores.
	Conjuntos de datos
bzr **ci700443v** **cox2** **dhfr** **minf00320555**	Estos conjuntos de datos aparecen descritos con detalle en el Subepígrafe 2.2.1. Los símbolos empleados aquí también aparecen reflejados.

LOO	Validación cruzada dejando uno fuera o *Leave One Out.*
Boot	Validación por re-muestreo o *bootstrap.*
5-Fold	Validación cruzada usando 5 particiones o *5-folds cross validation.*

Anexo 1. Resultados del proceso de entrenamiento de los modelos.

Tabla A1.1. Desempeño general de los modelos de clasificación obtenidos para los conjuntos de entrenamiento.

	Original				ECFP				GeoMean			
Clasificador	**Acc.**	**Se.**	**Sp.**	**(Se.+Sp.)/2**	**Acc.**	**Se.**	**Sp.**	**(Se.+Sp.)/2**	**Acc.**	**Se.**	**Sp.**	**(Se.+Sp.)/2**
minf00320555												
GALSSVM	0,8320	0,8268	0,8376	0,8322	0,8744	0,9159	0,8333	0,8746	0,8515	0,8718	0,8304	0,8511
GALDA	0,8402	0,8425	0,8376	0,8401	0,8791	0,8879	0,8704	0,8791	0,8646	0,8974	0,8304	0,8639
GAAB	0,8361	0,8205	0,8504	0,8355	0,8558	0,8519	0,8598	0,8558	0,8603	0,8482	0,8718	0,8600
BT_LSSVM	0,8975	0,9449	0,8462	0,8955	0,9302	0,9720	0,8889	0,9304	0,8515	0,8718	0,8304	0,8511
BT_LDA	0,7910	0,7874	0,7949	0,7911	0,8233	0,8224	0,8241	0,8233	0,8428	0,8547	0,8304	0,8425
BT_AB	0,7664	0,7402	0,7949	0,7675	0,8093	0,8224	0,7963	0,8094	0,7642	0,7265	0,8036	0,7650
FR_LSSVM	0,8484	0,8346	0,8632	0,8489	0,9209	0,9346	0,9074	0,9210	0,8472	0,8889	0,8036	0,8462
FR_LDA	0,7459	0,7402	0,7521	0,7461	0,7721	0,7477	0,7963	0,7720	0,7686	0,7607	0,7768	0,7687
FR_AB	0,7869	0,8504	0,7179	0,7842	0,7907	0,8411	0,7407	0,7909	0,7991	0,8547	0,7411	0,7979
dhfr												
GALSSVM	0,9600	0,8852	1,0000	0,9426	0,9883	0,9672	1,0000	0,9836	0,9826	0,9508	1,0000	0,9754
GALDA	0,8686	0,7705	0,9211	0,8458	0,9064	0,8197	0,9545	0,8871	0,8837	0,7541	0,9550	0,8545
GAAB	0,8229	0,8860	0,7049	0,7954	0,8129	0,8545	0,7377	0,7961	0,8605	0,9640	0,6721	0,8180
BT_LSSVM	0,9086	0,7869	0,9737	0,8803	0,9357	0,8525	0,9818	0,9171	0,8779	0,7213	0,9640	0,8426
BT_LDA	0,8629	0,8033	0,8947	0,8490	0,8070	0,7213	0,8545	0,7879	0,8430	0,7705	0,8829	0,8267
BT_AB	0,7714	0,5410	0,8947	0,7179	0,7895	0,6393	0,8727	0,7560	0,7558	0,4918	0,9009	0,6964
FR_LSSVM	0,7257	0,5574	0,8158	0,6866	0,7193	0,8197	0,6636	0,7417	0,7267	0,8197	0,6757	0,7477
FR_LDA	0,6114	0,7377	0,5439	0,6408	0,4678	1,0000	0,1727	0,5864	0,6221	0,7377	0,5586	0,6481
FR_AB	0,5657	0,7213	0,4825	0,6019	0,5673	0,7213	0,4818	0,6016	0,5698	0,7213	0,4865	0,6039
cox2												
GALSSVM	0,8045	0,8065	0,8028	0,8046	0,8868	0,8519	0,9231	0,8875	0,8509	0,9123	0,7895	0,8509
GALDA	0,8346	0,8548	0,8169	0,8359	0,8868	0,9074	0,8654	0,8864	0,8772	0,9123	0,8421	0,8772
GAAB	0,8195	0,7887	0,8548	0,8218	0,8868	0,8654	0,9074	0,8864	0,8421	0,8421	0,8421	0,8421
BT_LSSVM	0,8271	0,7742	0,8732	0,8237	0,9340	0,9074	0,9615	0,9345	0,9035	0,8947	0,9123	0,9035
BT_LDA	0,8421	0,8226	0,8592	0,8409	0,8208	0,8148	0,8269	0,8209	0,8509	0,8070	0,8947	0,8509
BT_AB	0,7820	0,7903	0,7746	0,7825	0,8113	0,7407	0,8846	0,8127	0,8158	0,8246	0,8070	0,8158
FR_LSSVM	0,7293	0,5161	0,9155	0,7158	0,7830	0,8519	0,7115	0,7817	0,7281	0,8421	0,6140	0,7281
FR_LDA	0,7218	0,4839	0,9296	0,7067	0,7264	0,5741	0,8846	0,7293	0,7281	0,8421	0,6140	0,7281
FR_AB	0,7218	0,4839	0,9296	0,7067	0,7736	0,8889	0,6538	0,7714	0,7632	0,6316	0,8947	0,7632

ci700443v												
GALSSVM	0,8972	0,9211	0,8655	0,8933	0,9018	0,9087	0,8929	0,9008	0,9003	0,9279	0,8639	0,8959
GALDA	0,9048	0,9167	0,8889	0,9028	0,8992	0,9087	0,8869	0,8978	0,9054	0,8874	0,9290	0,9082
GAAB	0,8471	0,8304	0,8596	0,8450	0,8734	0,8452	0,8950	0,8701	0,8542	0,8402	0,8649	0,8526
BT_LSSVM	0,9148	0,9342	0,8889	0,9115	0,8656	0,9041	0,8155	0,8598	0,8824	0,9369	0,8107	0,8738
BT_LDA	0,8446	0,8640	0,8187	0,8414	0,8630	0,8904	0,8274	0,8589	0,8517	0,8919	0,7988	0,8454
BT_AB	0,7945	0,7807	0,8129	0,7968	0,8191	0,8219	0,8155	0,8187	0,8031	0,7928	0,8166	0,8047
FR_LSSVM	0,8296	0,8904	0,7485	0,8194	0,7390	0,6849	0,8095	0,7472	0,7366	0,6802	0,8107	0,7454
FR_LDA	0,7268	0,6623	0,8129	0,7376	0,7390	0,6849	0,8095	0,7472	0,7366	0,6802	0,8107	0,7454
FR_AB	0,8271	0,8860	0,7485	0,8173	0,7390	0,6849	0,8095	0,7472	0,7366	0,6802	0,8107	0,7454
bzr												
GALSSVM	0,8489	0,8243	0,8769	0,8506	0,8346	0,8429	0,8254	0,8341	0,8444	0,8571	0,8308	0,8440
GALDA	0,9137	0,9459	0,8769	0,9114	0,9323	0,9571	0,9048	0,9310	0,9111	0,9286	0,8923	0,9104
GAAB	0,8489	0,8615	0,8378	0,8497	0,8647	0,8730	0,8571	0,8651	0,8519	0,8615	0,8429	0,8522
BT_LSSVM	0,9281	0,9459	0,9077	0,9268	0,9398	0,9429	0,9365	0,9397	0,9704	0,9857	0,9538	0,9698
BT_LDA	0,8273	0,8514	0,8000	0,8257	0,7744	0,7714	0,7778	0,7746	0,7630	0,7143	0,8154	0,7648
BT_AB	0,7122	0,6892	0,7385	0,7138	0,7895	0,8286	0,7460	0,7873	0,7185	0,6714	0,7692	0,7203
FR_LSSVM	0,8201	0,7838	0,8615	0,8227	0,8421	0,8143	0,8730	0,8437	0,8296	0,8000	0,8615	0,8308
FR_LDA	0,7266	0,7027	0,7538	0,7283	0,7594	0,7571	0,7619	0,7595	0,7407	0,7429	0,7385	0,7407
FR_AB	0,7266	0,7027	0,7538	0,7283	0,7444	0,7286	0,7619	0,7452	0,7407	0,7286	0,7538	0,7412
	Prom.				**Prom.**	**Dif**	***p-value***		**Prom.**	**Dif**	***p-value***	
Acc.	0,8103				0,8240	0,0137	**0,0030**		0,8157	0,0055	**0,0100**	
Se.	0,7865				0,8321	0,0456	**0,0000**		0,8138	0,0274	**0,0140**	
Sp.	0,8252				0,8192	-0,0060	0,0670		0,8133	-0,0119	0,3200	
(Se.+Sp.)/2	0,8058				0,8256	0,0198	**0,0010**		0,8136	0,0077	**0,0070**	

Tabla A1.2. Influencia de la eliminación de *activity cliffs* sobre la clasificación de la clase peor clasificada para los conjuntos de entrenamiento.

	Original	ECFP		GeoMean	
Clasificador	**Worst Class**	**Worst Class**	**Dif.**	**Worst Class**	**Dif.**
minf00320555					
GALSSVM	0,8268	0,9159	0,0891	0,8718	0,0450
GALDA	0,8376	0,8704	0,0328	0,8304	-0,0072
GAAB	0,8205	0,8519	0,0313	0,8482	0,0277
BT_LSSVM	0,8462	0,8889	0,0427	0,8304	-0,0158
BT_LDA	0,7874	0,8224	0,0350	0,8547	0,0673
BT_AB	0,7402	0,8224	0,0823	0,7265	-0,0137
FR_LSSVM	0,8346	0,9346	0,0999	0,8889	0,0542
FR_LDA	0,7402	0,7477	0,0075	0,7607	0,0205
FR_AB	0,7179	0,7407	0,0228	0,7411	0,0231
dhfr					
GALSSVM	0,8852	0,9672	0,0820	0,9508	0,0656
GALDA	0,7705	0,8197	0,0492	0,7541	-0,0164
GAAB	0,7049	0,7377	0,0328	0,6721	-0,0328
BT_LSSVM	0,7869	0,8525	0,0656	0,7213	-0,0656
BT_LDA	0,8033	0,7213	-0,0820	0,7705	-0,0328
BT_AB	0,5410	0,6393	0,0984	0,4918	-0,0492
FR_LSSVM	0,5574	0,8197	0,2623	0,8197	0,2623
FR_LDA	0,5439	0,1727	-0,3711	0,5586	0,0147
FR_AB	0,4825	0,4818	-0,0006	0,4865	0,0040
cox2					
GALSSVM	0,8028	0,9231	0,1203	0,7895	-0,0133
GALDA	0,8169	0,8654	0,0485	0,8421	0,0252
GAAB	0,7887	0,8654	0,0767	0,8421	0,0534
BT_LSSVM	0,7742	0,9074	0,1332	0,8947	0,1205
BT_LDA	0,8226	0,8148	-0,0078	0,8070	-0,0156
BT_AB	0,7746	0,8846	0,1100	0,8070	0,0324
FR_LSSVM	0,5161	0,8519	0,3357	0,8421	0,3260
FR_LDA	0,4839	0,5741	0,0902	0,8421	0,3582
FR_AB	0,4839	0,8889	0,4050	0,6316	0,1477
ci700443v					
GALSSVM	0,8655	0,8929	0,0274	0,8639	-0,0016
GALDA	0,8889	0,8869	-0,0020	0,9290	0,0401
GAAB	0,8304	0,8452	0,0148	0,8402	0,0098
BT_LSSVM	0,8889	0,8155	-0,0734	0,8107	-0,0782
BT_LDA	0,8187	0,8274	0,0087	0,7988	-0,0199
BT_AB	0,7807	0,8219	0,0412	0,7928	0,0121
FR_LSSVM	0,7485	0,8095	0,0610	0,8107	0,0621
FR_LDA	0,6623	0,6849	0,0227	0,6802	0,0179
FR_AB	0,7485	0,8095	0,0610	0,8107	0,0621
bzr					
GALSSVM	0,8243	0,8429	0,0185	0,8571	0,0328
GALDA	0,8769	0,9048	0,0278	0,8923	0,0154
GAAB	0,8378	0,8571	0,0193	0,8429	0,0050
BT_LSSVM	0,9077	0,9365	0,0288	0,9538	0,0462
BT_LDA	0,8000	0,7778	-0,0222	0,8154	0,0154
BT_AB	0,6892	0,8286	0,1394	0,6714	-0,0178
FR_LSSVM	0,7838	0,8143	0,0305	0,8000	0,0162
FR_LDA	0,7027	0,7571	0,0544	0,7429	0,0402

FR_AB	0,7027	0,7286	0,0259	0,7286	0,0259
Prom.	0,7522	0,8050	0,0528	0,7893	0,0371
p-value		**0,000**		**0,001**	

Tabla A1.3. Resultados de la validación de los modelos obtenidos.

	Original			ECFP			GeoMean		
Clasificador	**LOO**	**Boot**	**5-Fold**	**LOO**	**Boot**	**5-Fold**	**LOO**	**Boot**	**5-Fold**
minf00320555									
GALSSVM	0,8279	0,8041	0,8279	0,8651	0,8301	0,8419	0,8472	0,8171	0,8515
GALDA	0,8320	0,8102	0,8197	0,8651	0,8334	0,8651	0,8515	0,8171	0,8428
GAAB	0,8320	0,7989	0,8320	0,8512	0,8286	0,8465	0,8603	0,8383	0,8603
BT_LSSVM	0,7992	0,7622	0,7910	0,8233	0,8040	0,7953	0,7991	0,7905	0,7948
BT_LDA	0,7500	0,7432	0,7582	0,7860	0,7888	0,7860	0,8122	0,7864	0,7860
BT_AB	0,7213	0,7323	0,7459	0,7628	0,7614	0,7860	0,7249	0,7379	0,7336
FR_LSSVM	0,7951	0,7628	0,7828	0,7814	0,7561	0,7767	0,8253	0,7833	0,8122
FR_LDA	0,7459	0,7388	0,7418	0,7674	0,7718	0,7721	0,7686	0,7616	0,7729
FR_AB	0,7828	0,7204	0,7746	0,7535	0,7564	0,7674	0,7991	0,7477	0,7555
dhfr									
GALSSVM	0,8686	0,8327	0,8686	0,9064	0,8363	0,8538	0,8488	0,8263	0,8663
GALDA	0,8229	0,7974	0,8229	0,8655	0,6996	0,8363	0,8430	0,7965	0,8663
GAAB	0,8057	0,7405	0,7829	0,8012	0,7718	0,8012	0,8372	0,7735	0,8488
BT_LSSVM	0,7771	0,7780	0,8114	0,8304	0,7917	0,8304	0,7965	0,7639	0,7674
BT_LDA	0,7429	0,7340	0,7600	0,7485	0,7365	0,7135	0,7907	0,7601	0,7674
BT_AB	0,7371	0,7132	0,7314	0,7602	0,7320	0,7661	0,7151	0,6903	0,7267
FR_LSSVM	0,7257	0,6768	0,7029	0,7076	0,6791	0,7076	0,7093	0,6841	0,7151
FR_LDA	0,6114	0,5652	0,6114	0,4678	0,4681	0,4678	0,6221	0,5761	0,5988
FR_AB	0,5657	0,5551	0,5657	0,5673	0,5537	0,5673	0,5698	0,5607	0,5698
cox2									
GALSSVM	0,7895	0,7508	0,8045	0,8491	0,8123	0,8679	0,8509	0,7688	0,8246
GALDA	0,8120	0,7597	0,8120	0,8774	0,8253	0,8491	0,8684	0,7885	0,8421
GAAB	0,8120	0,7369	0,7820	0,8679	0,8373	0,8868	0,8421	0,8016	0,8333
BT_LSSVM	0,7368	0,7117	0,7444	0,8113	0,7546	0,8208	0,7895	0,7656	0,7807
BT_LDA	0,7444	0,7235	0,7519	0,7238	0,7169	0,0000	0,7719	0,7361	0,7281
BT_AB	0,7293	0,6991	0,7519	0,7736	0,7315	0,7642	0,7895	0,7520	0,7544
FR_LSSVM	0,7143	0,6898	0,7293	0,7547	0,7111	0,7547	0,7281	0,7162	0,7281
FR_LDA	0,7218	0,6908	0,7218	0,7264	0,7142	0,7264	0,7281	0,7248	0,7281
FR_AB	0,7218	0,6994	0,7218	0,7075	0,7019	0,7264	0,5965	0,6986	0,6930
ci700443v									
GALSSVM	0,8772	0,8591	0,8647	0,8863	0,8586	0,8786	0,8875	0,8681	0,8875
GALDA	0,8822	0,8595	0,8747	0,8760	0,8582	0,8682	0,8772	0,8602	0,8875
GAAB	0,8446	0,8340	0,8471	0,8734	0,8445	0,8605	0,8517	0,8327	0,8542
BT_LSSVM	0,8546	0,8353	0,8647	0,8501	0,8409	0,8475	0,8568	0,8311	0,8593
BT_LDA	0,8221	0,8219	0,8221	0,8398	0,8363	0,8320	0,8465	0,8340	0,8440
BT_AB	0,7794	0,7829	0,7970	0,8165	0,8115	0,8165	0,8056	0,7808	0,8031
FR_LSSVM	0,8296	0,8286	0,8296	0,7390	0,7373	0,7390	0,7366	0,7354	0,7366
FR_LDA	0,7268	0,7268	0,7268	0,7390	0,7373	0,7390	0,7366	0,7354	0,7366
FR_AB	0,8271	0,7881	0,7794	0,7390	0,7373	0,7390	0,7366	0,7354	0,7366
bzr									
GALSSVM	0,8489	0,7513	0,8345	0,8346	0,7160	0,8346	0,8370	0,6583	0,8444
GALDA	0,8921	0,8592	0,9065	0,8872	0,8526	0,8947	0,8815	0,7491	0,8815
GAAB	0,7986	0,8034	0,8273	0,8271	0,8201	0,8421	0,8370	0,7769	0,8370
BT_LSSVM	0,8058	0,7581	0,7986	0,7519	0,6918	0,7444	0,7778	0,7455	0,7556
BT_LDA	0,7626	0,7318	0,7482	0,7744	0,7560	0,7669	0,7481	0,6850	0,7037
BT_AB	0,7122	0,6813	0,6906	0,7820	0,7374	0,7744	0,6519	0,6650	0,7111

FR_LSSVM	0,7698	0,7363	0,7698	0,7970	0,7618	0,7820	0,7852	0,7409	0,7704
FR_LDA	0,7266	0,7260	0,7266	0,7218	0,7249	0,7293	0,7333	0,7350	0,7333
FR_AB	0,7266	0,7260	0,7266	0,7444	0,7459	0,7444	0,7407	0,7393	0,7407
	Prom.			**Prom.**	**Dif**	***p-value***	**Prom.**	**Dif**	***p-value***
LOO	0,7780			0,7885	0,0104	**0,0050**	0,7847	0,0067	**0,0110**
Boot	0,7519			0,7616	0,0097	**0,0040**	0,7549	0,0030	**0,0370**
5-Fold	0,7775			0,7691	-0,0083	**0,0140**	0,7816	0,0041	0,0560

Anexo 2. Resultados del desempeño de los modelos sobre los conjuntos de prueba.

Tabla A2.1. Desempeño de los modelos de clasificación obtenidos para los conjuntos de prueba.

	Original				ECFP				GeoMean			
Clasificador	**Acc.**	**Se.**	**Sp.**	**(Se.+Sp.)/2**	**Acc.**	**Se.**	**Sp.**	**(Se.+Sp.)/2**	**Acc.**	**Se.**	**Sp.**	**(Se.+Sp.)/2**
minf00320555												
GALSSVM	0,8026	0,7895	0,8158	0,8026	0,8026	0,8684	0,7368	0,8026	0,8158	0,8421	0,7895	0,8158
GALDA	0,8026	0,8158	0,7895	0,8026	0,7895	0,7895	0,7895	0,7895	0,8026	0,8421	0,7632	0,8026
GAAB	0,8158	0,7895	0,8421	0,8158	0,8026	0,7895	0,8158	0,8026	0,8158	0,7895	0,8421	0,8158
BT_LSSVM	0,7632	0,8158	0,7105	0,7632	0,7632	0,8158	0,7105	0,7632	0,8026	0,8158	0,7895	0,8026
BT_LDA	0,8158	0,7895	0,8421	0,8158	0,8026	0,7895	0,8158	0,8026	0,8026	0,8158	0,7895	0,8026
BT_AB	0,8026	0,7895	0,8158	0,8026	0,7895	0,8158	0,7632	0,7895	0,8289	0,7632	0,8947	0,8289
FR_LSSVM	0,7763	0,7632	0,7895	0,7763	0,7763	0,8158	0,7368	0,7763	0,7632	0,8421	0,6842	0,7632
FR_LDA	0,8158	0,8158	0,8158	0,8158	0,8026	0,8158	0,7895	0,8026	0,8026	0,8158	0,7895	0,8026
FR_AB	0,7895	0,8947	0,6842	0,7895	0,8158	0,9211	0,7105	0,8158	0,7895	0,8947	0,6842	0,7895
dhfr												
GALSSVM	0,7931	0,6957	0,8571	0,7764	0,8276	0,7826	0,8571	0,8199	0,8621	0,6957	0,9714	0,8335
GALDA	0,8276	0,6522	0,9429	0,7975	0,8103	0,6522	0,9143	0,7832	0,7759	0,5652	0,9143	0,7398
GAAB	0,7586	0,8857	0,5652	0,7255	0,7759	0,8286	0,6957	0,7621	0,7759	0,9429	0,5217	0,7323
BT_LSSVM	0,7759	0,5652	0,9143	0,7398	0,8621	0,8696	0,8571	0,8634	0,7414	0,4783	0,9143	0,6963
BT_LDA	0,7931	0,6957	0,8571	0,7764	0,7759	0,7826	0,7714	0,7770	0,7586	0,5652	0,8857	0,7255
BT_AB	0,7069	0,5217	0,8286	0,6752	0,7241	0,5652	0,8286	0,6969	0,7241	0,5217	0,8571	0,6894
FR_LSSVM	0,6379	0,3913	0,8000	0,5957	0,6724	0,7391	0,6286	0,6839	0,6897	0,7391	0,6571	0,6981
FR_LDA	0,5862	0,7391	0,4857	0,6124	0,5345	1,0000	0,2286	0,6143	0,5862	0,7391	0,4857	0,6124
FR_AB	0,5345	0,7391	0,4000	0,5696	0,5345	0,7391	0,4000	0,5696	0,5345	0,7391	0,4000	0,5696
cox2												
GALSSVM	0,8222	0,7600	0,9000	0,8300	0,8889	0,8800	0,9000	0,8900	0,8667	0,8400	0,9000	0,8700
GALDA	0,8222	0,7600	0,9000	0,8300	0,9111	0,8800	0,9500	0,9150	0,8667	0,8400	0,9000	0,8700
GAAB	0,8444	0,8500	0,8400	0,8450	0,8444	0,9000	0,8000	0,8500	0,8222	0,9000	0,7600	0,8300
BT_LSSVM	0,8000	0,7200	0,9000	0,8100	0,8222	0,7600	0,9000	0,8300	0,8667	0,8800	0,8500	0,8650
BT_LDA	0,7778	0,7200	0,8500	0,7850	0,8000	0,7200	0,9000	0,8100	0,8667	0,8000	0,9500	0,8750
BT_AB	0,7333	0,6400	0,8500	0,7450	0,7333	0,6000	0,9000	0,7500	0,7556	0,7200	0,8000	0,7600
FR_LSSVM	0,7111	0,5200	0,9500	0,7350	0,7556	0,7200	0,8000	0,7600	0,7556	0,7200	0,8000	0,7600
FR_LDA	0,6889	0,4800	0,9500	0,7150	0,7111	0,5200	0,9500	0,7350	0,7333	0,7200	0,7500	0,7350
FR_AB	0,6889	0,4800	0,9500	0,7150	0,8000	0,8400	0,7500	0,7950	0,7333	0,6000	0,9000	0,7500

ci700443v												
GALSSVM	0,8716	0,8906	0,8444	0,8675	0,8807	0,8750	0,8889	0,8819	0,8624	0,8438	0,8889	0,8663
GALDA	0,8899	0,9063	0,8667	0,8865	0,8716	0,8594	0,8889	0,8741	0,8716	0,8125	0,9556	0,8840
GAAB	0,8532	0,8000	0,8906	0,8453	0,8440	0,8222	0,8594	0,8408	0,8349	0,7778	0,8750	0,8264
BT_LSSVM	0,8440	0,8438	0,8444	0,8441	0,8716	0,9063	0,8222	0,8642	0,8716	0,9375	0,7778	0,8576
BT_LDA	0,8716	0,8906	0,8444	0,8675	0,8532	0,8906	0,8000	0,8453	0,8532	0,9063	0,7778	0,8420
BT_AB	0,7798	0,7969	0,7556	0,7762	0,7798	0,7813	0,7778	0,7795	0,7615	0,7500	0,7778	0,7639
FR_LSSVM	0,7890	0,8906	0,6444	0,7675	0,7798	0,7813	0,7778	0,7795	0,7798	0,7813	0,7778	0,7795
FR_LDA	0,7798	0,7813	0,7778	0,7795	0,7798	0,7813	0,7778	0,7795	0,7798	0,7813	0,7778	0,7795
FR_AB	0,7890	0,8906	0,6444	0,7675	0,7798	0,7813	0,7778	0,7795	0,7798	0,7813	0,7778	0,7795
bzr												
GALSSVM	0,7436	0,7895	0,7000	0,7447	0,7179	0,7895	0,6500	0,7197	0,7692	0,8421	0,7000	0,7711
GALDA	0,7692	0,9474	0,6000	0,7737	0,8205	0,9474	0,7000	0,8237	0,7692	0,9474	0,6000	0,7737
GAAB	0,6923	0,6000	0,7895	0,6947	0,7692	0,7500	0,7895	0,7697	0,7692	0,7500	0,7895	0,7697
BT_LSSVM	0,7436	0,8947	0,6000	0,7474	0,7436	0,8421	0,6500	0,7461	0,7692	0,8421	0,7000	0,7711
BT_LDA	0,7692	0,8421	0,7000	0,7711	0,6667	0,6842	0,6500	0,6671	0,6923	0,6842	0,7000	0,6921
BT_AB	0,6667	0,6316	0,7000	0,6658	0,7179	0,7895	0,6500	0,7197	0,7179	0,7368	0,7000	0,7184
FR_LSSVM	0,7436	0,7895	0,7000	0,7447	0,7436	0,7895	0,7000	0,7447	0,7436	0,7895	0,7000	0,7447
FR_LDA	0,6667	0,6842	0,6500	0,6671	0,6923	0,7368	0,6500	0,6934	0,6923	0,7368	0,6500	0,6934
FR_AB	0,6667	0,6842	0,6500	0,6671	0,6667	0,6842	0,6500	0,6671	0,6667	0,6842	0,6500	0,6671
	Prom.				**Prom.**	**Dif**	***p-value***		**Prom.**	**Dif**	***p-value***	
Acc.	0,7648				0,7757	0,0109	**0,019**		0,7761	0,0113	**0,012**	
Se.	0,7474				0,7931	0,0458	**0,004**		0,7736	0,0262	0,066	
Sp.	0,7789				0,7635	-0,0153	0,089		0,7738	-0,0051	0,345	
(Se.+Sp.)/2	0,7631				0,7783	0,0152	**0,003**		0,7737	0,0106	**0,020**	

Tabla A2.2. Influencia de la eliminación de activity cliffs sobre la clasificación de la clase peor clasificada para los conjuntos de prueba.

	Original	ECFP		GeoMean	
Clasificador	**Worst Class**	**Worst Class**	**Dif.**	**Worst Class**	**Dif.**
minf00320555					
GALSSVM	0,7895	0,8684	0,0789	0,8421	0,0526
GALDA	0,7895	0,7895	0,0000	0,7632	-0,0263
GAAB	0,7895	0,7895	0,0000	0,7895	0,0000
BT_LSSVM	0,7105	0,7105	0,0000	0,7895	0,0789
BT_LDA	0,7895	0,7895	0,0000	0,8158	0,0263
BT_AB	0,7895	0,8158	0,0263	0,7632	-0,0263
FR_LSSVM	0,7632	0,8158	0,0526	0,8421	0,0789
FR_LDA	0,8158	0,7895	-0,0263	0,7895	-0,0263
FR_AB	0,6842	0,7105	0,0263	0,6842	0,0000
dhfr					
GALSSVM	0,6957	0,7826	0,0870	0,6957	0,0000
GALDA	0,6522	0,6522	0,0000	0,5652	-0,0870
GAAB	0,5652	0,6957	0,1304	0,5217	-0,0435
BT_LSSVM	0,5652	0,8696	0,3043	0,4783	-0,0870
BT_LDA	0,6957	0,7826	0,0870	0,5652	-0,1304
BT_AB	0,5217	0,5652	0,0435	0,5217	0,0000
FR_LSSVM	0,3913	0,7391	0,3478	0,7391	0,3478
FR_LDA	0,4857	0,2286	-0,2571	0,4857	0,0000
FR_AB	0,4000	0,4000	0,0000	0,4000	0,0000
cox2					
GALSSVM	0,7600	0,8800	0,1200	0,8400	0,0800
GALDA	0,7600	0,8800	0,1200	0,8400	0,0800
GAAB	0,8400	0,8000	-0,0400	0,7600	-0,0800
BT_LSSVM	0,7200	0,7600	0,0400	0,8800	0,1600
BT_LDA	0,7200	0,7200	0,0000	0,8000	0,0800
BT_AB	0,6400	0,6000	-0,0400	0,7200	0,0800
FR_LSSVM	0,5200	0,7200	0,2000	0,7200	0,2000
FR_LDA	0,4800	0,5200	0,0400	0,7200	0,2400
FR_AB	0,4800	0,8400	0,3600	0,6000	0,1200
ci700443v					
GALSSVM	0,8444	0,8889	0,0444	0,8889	0,0444
GALDA	0,8667	0,8889	0,0222	0,9556	0,0889
GAAB	0,8000	0,8222	0,0222	0,7778	-0,0222
BT_LSSVM	0,8438	0,9063	0,0625	0,9375	0,0938
BT_LDA	0,8444	0,8000	-0,0444	0,7778	-0,0667
BT_AB	0,7556	0,7778	0,0222	0,7778	0,0222
FR_LSSVM	0,6444	0,7778	0,1333	0,7778	0,1333
FR_LDA	0,7778	0,7778	0,0000	0,7778	0,0000
FR_AB	0,6444	0,7778	0,1333	0,7778	0,1333
bzr					
GALSSVM	0,7000	0,6500	-0,0500	0,7000	0,0000
GALDA	0,6000	0,7000	0,1000	0,6000	0,0000
GAAB	0,6000	0,7500	0,1500	0,7500	0,1500
BT_LSSVM	0,6000	0,6500	0,0500	0,7000	0,1000
BT_LDA	0,7000	0,6500	-0,0500	0,7000	0,0000
BT_AB	0,6316	0,7895	0,1579	0,7368	0,1053

FR_LSSVM	0,7000	0,7000	0,0000	0,7000	0,0000
FR_LDA	0,6500	0,6500	0,0000	0,6500	0,0000
FR_AB	0,6500	0,6500	0,0000	0,6500	0,0000
Prom.	0,6815	0,7360	0,0545	0,7237	0,0422
p		**0,000**			**0,001**

Anexo 3. Resultados del desempeño de los modelos sobre los conjuntos de validación externa.

Tabla A3.1. Desempeño de los modelos de clasificación obtenidos para los conjuntos de validación externa.

	Original				ECFP				GeoMean			
Clasificador	**Acc.**	**Se.**	**Sp.**	**(Se.+Sp.)/2**	**Acc.**	**Se.**	**Sp.**	**(Se.+Sp.)/2**	**Acc.**	**Se.**	**Sp.**	**(Se.+Sp.)/2**
minf00320555												
GALSSVM	0,7284	0,7353	0,7234	0,7293	0,6914	0,7059	0,6809	0,6934	0,7407	0,7353	0,7447	0,7400
GALDA	0,6667	0,6176	0,7021	0,6599	0,7160	0,7059	0,7234	0,7146	0,7407	0,7647	0,7234	0,7441
GAAB	0,7037	0,6809	0,7353	0,7081	0,7160	0,7447	0,6765	0,7106	0,6914	0,6383	0,7647	0,7015
BT_LSSVM	0,7037	0,7941	0,6383	0,7162	0,7160	0,7647	0,6809	0,7228	0,7160	0,7353	0,7021	0,7187
BT_LDA	0,7407	0,7353	0,7447	0,7400	0,7531	0,6765	0,8085	0,7425	0,7037	0,7647	0,6596	0,7121
BT_AB	0,7531	0,7059	0,7872	0,7466	0,7160	0,6176	0,7872	0,7024	0,7284	0,6765	0,7660	0,7212
FR_LSSVM	0,6914	0,7059	0,6809	0,6934	0,6914	0,7647	0,6383	0,7015	0,7037	0,7647	0,6596	0,7121
FR_LDA	0,7160	0,6471	0,7660	0,7065	0,7160	0,6471	0,7660	0,7065	0,7037	0,6471	0,7447	0,6959
FR_AB	0,7160	0,7941	0,6596	0,7268	0,7284	0,7941	0,6809	0,7375	0,7160	0,7941	0,6596	0,7268
dhfr												
GALSSVM	0,8063	0,6744	0,8547	0,7646	0,7563	0,6977	0,7778	0,7377	0,7188	0,7442	0,7094	0,7268
GALDA	0,8438	0,7442	0,8803	0,8123	0,7500	0,7442	0,7521	0,7482	0,7875	0,6744	0,8291	0,7517
GAAB	0,7125	0,7692	0,5581	0,6637	0,7500	0,8120	0,5814	0,6967	0,7563	0,8376	0,5349	0,6862
BT_LSSVM	0,7750	0,5814	0,8462	0,7138	0,7438	0,6279	0,7863	0,7071	0,7563	0,5349	0,8376	0,6862
BT_LDA	0,7063	0,6512	0,7265	0,6888	0,7750	0,7442	0,7863	0,7653	0,7625	0,6977	0,7863	0,7420
BT_AB	0,7438	0,5349	0,8205	0,6777	0,7375	0,5581	0,8034	0,6808	0,7625	0,5116	0,8547	0,6832
FR_LSSVM	0,6750	0,2326	0,8376	0,5351	0,6875	0,5116	0,7521	0,6319	0,6813	0,4884	0,7521	0,6203
FR_LDA	0,5500	0,6279	0,5214	0,5746	0,3750	0,9767	0,1538	0,5653	0,5500	0,6279	0,5214	0,5746
FR_AB	0,5000	0,6512	0,4444	0,5478	0,5000	0,6512	0,4444	0,5478	0,5000	0,6512	0,4444	0,5478
cox2												
GALSSVM	0,7280	0,7869	0,6719	0,7294	0,6640	0,6557	0,6719	0,6638	0,6480	0,7213	0,5781	0,6497
GALDA	0,7200	0,7869	0,6563	0,7216	0,6800	0,7213	0,6406	0,6810	0,6800	0,7869	0,5781	0,6825
GAAB	0,6640	0,5781	0,7541	0,6661	0,6240	0,4844	0,7705	0,6274	0,6960	0,6250	0,7705	0,6977
BT_LSSVM	0,7040	0,7213	0,6875	0,7044	0,6080	0,5738	0,6406	0,6072	0,6480	0,7213	0,5781	0,6497
BT_LDA	0,6960	0,7213	0,6719	0,6966	0,6400	0,7377	0,5469	0,6423	0,6560	0,7049	0,6094	0,6571
BT_AB	0,6720	0,7541	0,5938	0,6739	0,6560	0,6557	0,6563	0,6560	0,6000	0,7213	0,4844	0,6028
FR_LSSVM	0,6240	0,5246	0,7188	0,6217	0,6560	0,8033	0,5156	0,6595	0,6160	0,8525	0,3906	0,6215
FR_LDA	0,6880	0,5574	0,8125	0,6849	0,6640	0,6066	0,7188	0,6627	0,6080	0,8525	0,3750	0,6137

FR_AB	0,6880	0,5574	0,8125	0,6849	0,6480	0,8689	0,4375	0,6532	0,5760	0,5738	0,5781	0,5759
ci700443v												
GALSSVM	0,7969	0,9041	0,6545	0,7793	0,7969	0,8630	0,7091	0,7861	0,8125	0,8767	0,7273	0,8020
GALDA	0,7891	0,8767	0,6727	0,7747	0,7969	0,8493	0,7273	0,7883	0,8047	0,8493	0,7455	0,7974
GAAB	0,7578	0,6727	0,8219	0,7473	0,8047	0,7455	0,8493	0,7974	0,7656	0,6727	0,8356	0,7542
BT_LSSVM	0,7969	0,8767	0,6909	0,7838	0,8203	0,9178	0,6909	0,8044	0,7891	0,8904	0,6545	0,7725
BT_LDA	0,7188	0,8219	0,5818	0,7019	0,7500	0,8493	0,6182	0,7337	0,7422	0,8630	0,5818	0,7224
BT_AB	0,7344	0,8219	0,6182	0,7200	0,7891	0,8630	0,6909	0,7770	0,7656	0,8356	0,6727	0,7542
FR_LSSVM	0,7422	0,9041	0,5273	0,7157	0,6406	0,6575	0,6182	0,6379	0,6406	0,6575	0,6182	0,6379
FR_LDA	0,6406	0,6575	0,6182	0,6379	0,6406	0,6575	0,6182	0,6379	0,6406	0,6575	0,6182	0,6379
FR_AB	0,7422	0,9041	0,5273	0,7157	0,6406	0,6575	0,6182	0,6379	0,6406	0,6575	0,6182	0,6379
bzr												
GALSSVM	0,6800	0,6032	0,7581	0,6806	0,7200	0,6667	0,7742	0,7204	0,7280	0,6984	0,7581	0,7282
GALDA	0,7520	0,7619	0,7419	0,7519	0,7440	0,7460	0,7419	0,7440	0,6960	0,6508	0,7419	0,6964
GAAB	0,7040	0,7258	0,6825	0,7042	0,6960	0,7581	0,6349	0,6965	0,7120	0,7742	0,6508	0,7125
BT_LSSVM	0,7200	0,6825	0,7581	0,7203	0,7200	0,6984	0,7419	0,7202	0,7280	0,6667	0,7903	0,7285
BT_LDA	0,6880	0,6349	0,7419	0,6884	0,6400	0,5079	0,7742	0,6411	0,6720	0,6032	0,7419	0,6726
BT_AB	0,6160	0,5397	0,6935	0,6166	0,6640	0,6349	0,6935	0,6642	0,6160	0,5079	0,7258	0,6169
FR_LSSVM	0,6800	0,6508	0,7097	0,6802	0,6800	0,6508	0,7097	0,6802	0,6800	0,6349	0,7258	0,6804
FR_LDA	0,6160	0,5714	0,6613	0,6164	0,6400	0,6190	0,6613	0,6402	0,6560	0,6508	0,6613	0,6560
FR_AB	0,6160	0,5714	0,6613	0,6164	0,6160	0,5714	0,6613	0,6164	0,6160	0,5714	0,6613	0,6164
	Prom.				**Prom.**	**Dif**	***p-value***		**Prom.**	**Dif**	***p-value***	
Acc.	0,7046				0,6924	-0,0122	0,109		0,6923	-0,0123	0,078	
Se.	0,6901				0,7059	0,0159	0,226		0,7015	0,0114	0,295	
Sp.	0,6984				0,6758	-0,0226	0,210		0,6704	-0,0280	0,097	
(Se.+Sp.)/2	0,6942				0,6909	-0,0034	0,327		0,6859	-0,0083	0,158	

Tabla A3.2. Influencia de la eliminación de activity cliffs sobre la clasificación de la clase peor clasificada para los conjuntos de validación externa.

	Original	ECFP		GeoMean	
Clasificador	**Worst Class**	**Worst Class**	**Dif.**	**Worst Class**	**Dif.**
minf00320555					
GALSSVM	0,7234	0,6809	-0,0426	0,7447	0,0213
GALDA	0,6176	0,7059	0,0882	0,7647	0,1471
GAAB	0,6809	0,7447	0,0638	0,6383	-0,0426
BT_LSSVM	0,6383	0,6809	0,0426	0,7021	0,0638
BT_LDA	0,7353	0,6765	-0,0588	0,7647	0,0294
BT_AB	0,7059	0,6176	-0,0882	0,6765	-0,0294
FR_LSSVM	0,6809	0,6383	-0,0426	0,6596	-0,0213
FR_LDA	0,6471	0,6471	0,0000	0,6471	0,0000
FR_AB	0,6596	0,6809	0,0213	0,6596	0,0000
dhfr					
GALSSVM	0,6744	0,6977	0,0233	0,7442	0,0698
GALDA	0,7442	0,7442	0,0000	0,6744	-0,0698
GAAB	0,5581	0,5814	0,0233	0,5349	-0,0233
BT_LSSVM	0,5814	0,6279	0,0465	0,5349	-0,0465
BT_LDA	0,6512	0,7442	0,0930	0,6977	0,0465
BT_AB	0,5349	0,5581	0,0233	0,5116	-0,0233
FR_LSSVM	0,2326	0,5116	0,2791	0,4884	0,2558
FR_LDA	0,5214	0,1538	-0,3675	0,5214	0,0000
FR_AB	0,4444	0,4444	0,0000	0,4444	0,0000
cox2					
GALSSVM	0,6719	0,6719	0,0000	0,5781	-0,0938
GALDA	0,6563	0,6406	-0,0156	0,5781	-0,0781
GAAB	0,5781	0,4844	-0,0938	0,6250	0,0469
BT_LSSVM	0,6875	0,6406	-0,0469	0,5781	-0,1094
BT_LDA	0,6719	0,5469	-0,1250	0,6094	-0,0625
BT_AB	0,5938	0,6563	0,0625	0,4844	-0,1094
FR_LSSVM	0,5246	0,8033	0,2787	0,8525	0,3279
FR_LDA	0,5574	0,6066	0,0492	0,8525	0,2951

FR_AB	0,5574	0,8689	0,3115	0,5738	0,0164
ci700443v					
GALSSVM	0,6545	0,7091	0,0545	0,7273	0,0727
GALDA	0,6727	0,7273	0,0545	0,7455	0,0727
GAAB	0,6727	0,7455	0,0727	0,6727	0,0000
BT_LSSVM	0,6909	0,6909	0,0000	0,6545	-0,0364
BT_LDA	0,5818	0,6182	0,0364	0,5818	0,0000
BT_AB	0,6182	0,6909	0,0727	0,6727	0,0545
FR_LSSVM	0,5273	0,6182	0,0909	0,6182	0,0909
FR_LDA	0,6182	0,6182	0,0000	0,6182	0,0000
FR_AB	0,5273	0,6182	0,0909	0,6182	0,0909
bzr					
GALSSVM	0,6032	0,6667	0,0635	0,6984	0,0952
GALDA	0,7419	0,7419	0,0000	0,7419	0,0000
GAAB	0,6825	0,6349	-0,0476	0,6508	-0,0317
BT_LSSVM	0,6825	0,6984	0,0159	0,6667	-0,0159
BT_LDA	0,6349	0,5079	-0,1270	0,6032	-0,0317
BT_AB	0,5397	0,6349	0,0952	0,5079	-0,0317
FR_LSSVM	0,6508	0,6508	0,0000	0,6349	-0,0159
FR_LDA	0,5714	0,6190	0,0476	0,6508	0,0794
FR_AB	0,5714	0,5714	0,0000	0,5714	0,0000
Prom.	**0,6172**	**0,6404**	**0,0232**	**0,6395**	**0,0223**
p-value		**0,031**		0,135	