UNIVERSIDAD NACIONAL DEL SUR

TESIS DE DOCTOR EN CIENCIAS DE LA COMPUTACIÓN

Desarrollo de Métodos Analíticos y de Predicción para Informática Molecular Basados en Técnicas de Aprendizaje Automático y Visualización

María Jimena Martínez

Bahía Blanca Argentina

2017

Prefacio

Esta Tesis es presentada como parte de los requisitos para optar al grado académico de Doctor en Ciencias de la Computación, de la Universidad Nacional del Sur, y no ha sido presentada previamente para la obtención de otro título en esta Universidad u otras. La misma contiene los resultados obtenidos en investigaciones llevadas a cabo en el Departamento de Ciencias de la Computación y en el Instituto de Ciencias e Ingeniería de la Computación, durante el período comprendido entre el 18 de septiembre de 2012 y el 3 de mayo de 2017, bajo la dirección del Dr. Gustavo E. Vazquez, profesor y director del Departamento de Informática y Ciencias de la Computación de la Universidad Católica del Uruguay , y del Dr. Axel J. Soto, investigador asociado del Centro Nacional de Minería de Texto (NaCTeM) en la Escuela de Informática de la Universidad de Manchester, Inglaterra.

María Jimena Martínez

Bahía Blanca, 3 de mayo de 2017

Universidad Nacional del Sur
Secretaría General de Posgrado y Educación Continua

La presente tesis ha sido aprobada el …/…/……., mereciendo la calificación de …… (……..)

Agradecimientos

Esta tesis pudo terminarse gracias a la ayuda de muchas personas, es por eso que todas ellas son también autores de este trabajo.

En primer lugar, agradezco profundamente a mi marido Lionel y a mi hijo Surya por la paciencia, la comprensión y el apoyo durante toda esta etapa de manera incondicional. Inmensamente agradecida por el amor que me brindan cada día.

A mi madre Mónica, a mis abuelos María Luisa y Adolfo, a mis tíos Susana y José Luis, y a Javier, quienes de igual manera me han alentado, brindándome su ayuda y amor durante toda mi vida. Mi mayor gratitud hacia ellos.

También agradezco a mi familia política por brindarme siempre su cariño.

Quisiera agradecer también a mis directores, Gustavo y Axel por guiarme en el camino de la investigación. A Ignacio y Mónica, quienes también han sido directores para mí. Todos ellos me han ayudado con paciencia y me han acompañado en este camino, con una inmensa calidad humana. Gracias por darme esta oportunidad y transmitirme sus conocimientos.

A todo el grupo de trabajo y a mis compañeros de oficina por los días compartidos, y a mis amigas de siempre, especialmente a Julieta, con quien hemos compartido muchos años de amistad y de estudio.

Por último, quiero agradecer al Consejo Nacional de Investigaciones Científicas y Técnicas, al Departamento de Ciencias e Ingeniería de la Computación y al Instituto de Ciencias e Ingeniería de la Computación por brindarme los medios necesarios para el desarrollo de esta tesis.

Resumen

Los distintos procesos involucrados en la industria química deben ser estudiados cuidadosamente con el fin de obtener productos de calidad al menor costo y causando el mínimo daño al medio ambiente (ej. industria de polímeros sintéticos y diseño racional de fármacos). Hace ya varios años que distintos métodos computacionales son utilizados en la industria química con el fin de lograr esos objetivos. En particular, el modelado QSAR/QSPR es una técnica de gran interés dentro del área de la informática molecular, ya que permite correlacionar de manera cuantitativa características estructurales de una entidad química con una determinada propiedad físico-química o actividad biológica.

El objetivo de esa tesis fue desarrollar distintas metodologías para asistir a expertos en informática molecular en el proceso de predicción de propiedades fisicoquímicas o de actividad biológica. Más específicamente, las técnicas desarrolladas se enfocan en incorporar al proceso de modelado predictivo QSAR/QSPR, el conocimiento del experto en el dominio. De esta manera se logran mejorar ciertas características de los modelos, tales como su interpretación en términos físicos-químicos, las cuales permite aumentar la generalidad del modelo. Al respecto, se ha implementado una herramienta de analítica visual, denominada VIDEAN, que combina métodos estadísticos con visualizaciones interactivas para elegir un conjunto de descriptores que predigan una determinada propiedad objetivo. Otro de los aportes de esta tesis está relacionado con el dominio de aplicación de un modelo QSAR/QSPR. En este sentido, se ha implementado una técnica para determinar el dominio de aplicación de modelos de clasificación. Esto representa una novedad dado que la mayoría de las técnicas desarrolladas para este fin apuntan exclusivamente a los modelos de regresión.

Los métodos implementados han sido evaluados mediante el estudio de propiedades de relevancia para tres campos de aplicación: el diseño racional de fármacos, el diseño de materiales poliméricos (plásticos) y las ciencias ambientales. Con este fin, se han desarrollado numerosos modelos predictivos de regresión y clasificación. En el área de diseño racional de fármacos, las propiedades que se estudiaron están relacionadas con el comportamiento ADMET (absorción, distribución, metabolismo, excreción y toxicidad) de los mismos: *absorción intestinal humana (Human Intestinal Absorption, HIA)* y el *pasaje de la barrera hemato-encefálica (Blood-Brain Barrier, BBB)*, ambas esenciales para el desarrollo de nuevos fármacos. En el campo de los materiales poliméricos, se exploraron varias propiedades mecánicas, que proporcionan información relacionada con la ductilidad, resistencia y rigidez del material polimérico; y que, junto con otras propiedades, definen su perfil de aplicación estructural. Estas propiedades son: *elongación a la rotura (elongation at break)*, *resistencia a tensión en la rotura (tensile strength at break)* y *módulo elástico (tensile modulus)*. En el área de medioambiente, la propiedad que se estudió fue el *coeficiente de distribución sangre-hígado* (*log* P_{liver}) en compuestos orgánicos volátiles (VOCs), que son gases que se emiten de ciertos sólidos o líquidos y que son ampliamente utilizados como ingredientes en productos para el hogar (pinturas, los barnices, productos de limpieza, desinfección, cosmética, entre otros). Los resultados de estudios de este tipo de propiedades brindan un panorama de cómo se distribuyen estos tipos de compuestos en el organismo y pueden emplearse para la evaluación de riesgos y toma de decisiones en materia de salud pública.

Abstract

The various processes involved in the chemical industry must be carefully studied in order to obtain quality products at the lowest cost and causing the least damage to the environment (e.g. synthetic polymer industry and rational drug design). During the last two decades, different computational methods have been used in the chemical industry in order to achieve these objectives. In particular, QSAR/QSPR modeling is a technique of great interest in the area of molecular informatics, since it allows to quantitatively correlate structural characteristics of a chemical entity with a given physical-chemical or biological activity.

The objective of this thesis was to develop different methodologies to assist molecular computing experts in the process of predicting physicochemical or biological activity properties. More specifically, the techniques developed focus on incorporating domain expert's knowledge into the traditional automated predictive modeling process. In this way, certain characteristics of the models can be improved, such as their interpretation in physical-chemical terms, which allow to increase the generality on the model. In this sense, a visual analytics tool, called VIDEAN, has been implemented to combine statistical methods with interactive visualizations to choose a set of molecular descriptors that predict a specific target property. Another contribution of this thesis focuses on the implementation of a technique to determine the applicability domain of QSAR/QSPR classification models. In this regard, a technique has been implemented to determine the applicability domain of classification models. This represents a novelty given that most of the techniques developed for this purpose are exclusively intended for regression models.

Implemented methods have been evaluated using target properties of relevance in three application areas: rational drug design, design of polymeric materials (plastics) and environmental sciences. To this end, different predictive regression and classification models were proposed that overcome in performance and interpretability to other traditional models have been developed. To this end, numerous regression and classification models have been developed. In rational drug design, the properties that were studied are related to the ADMET behavior (absorption, distribution, metabolism, excretion and toxicity): Human Intestinal Absorption (HIA) and Blood-brain barrier (BBB), both essential for the development of new drugs. In the field of polymeric materials, various mechanical properties, which provide information related to the ductility, strength and rigidity of the polymeric material were explored, and which, along with other properties define its structural application profile. These properties are: *elongation at break*, *tensile strength at break* and *tensile modulus*. In environment area, the property studied was the blood - liver distribution coefficient (log P_{liver}) in volatile organic compounds (VOCs), which are gases that are emitted from certain solids or liquids and are widely used as ingredients in products for the home (paints, varnishes, cleaning products, disinfection, cosmetics, among others). The results obtained from this studies provide an overview of how these types of compounds are distributed in the body and can be used for risk assessment and public health decision making.

Lista de Publicaciones

Publicaciones indexadas en SCI Thompson Reuters y Scopus

- Ignacio Ponzoni, Victor Sebastian, Carlos Requena, Carlos Roca, **María J. Martínez**, Fiorella Cravero, Mónica F. Díaz, Juan A. Páez, Ramón Gómez Arrayas, Javier Adrio, Nuria E. Campillo· "Hybridizing Feature Selection and Feature Learning Approaches in QSAR Modeling for Drug Discovery**". Scientific Reports**; 7,2404 (2017).
- Fiorella Cravero, **María J. Martínez**, Mónica F. Díaz, Ignacio Ponzoni. "QSAR Classification Models for Predicting Affinity to Blood or Liver of Volatile Organic Compounds in e-Health". **Lecture Notes in Computer Science. International Work-Conference on Bioinformatics and Biomedical Engineering (IWBBIO)**; Springer International Publishing AG 2017; Part II, LNBI 10209, pp. 1–10 (2017).
- Fiorella Cravero, **María J. Martínez**, Gustavo E. Vazquez, Mónica F. Díaz, Ignacio Ponzoni. "Feature Learning applied to the Estimation of Tensile Strength at Break in Polymeric Material Design". **Journal of Integrative Bioinformatics(JIB)**; vol. 13 p. 286 – 301(2016).
- Fiorella Cravero, **María J. Martínez**, Gustavo E. Vazquez, Mónica F. Díaz, Ignacio Ponzoni. "Intelligent Systems for Predictive Modelling in Cheminformatics: QSPR Models for Material Design using Machine Learning and Visual Analytics Tools". **Advances in Intelligent Systems and Computing. 10th International Conference on Practical Applications of Computational Biology & Bioinformatics**. Vol. 477, (2016).
- **María J. Martínez**, Ignacio Ponzoni, Mónica F. Díaz, Gustavo E. Vazquez, Axel J. Soto. "Visual analytics in cheminformatics: user-supervised descriptor selection for QSAR methods". **Journal of Cheminformatics**; vol. 7 p. 1 – 17 (2015).
- Damián Palomba, **María J. Martínez**, Ignacio Ponzoni, Mónica F. Díaz, Gustavo E. Vazquez, Axel J. Soto. "QSPR Models for Predicting Log Pliver Values for Volatile Organic Compounds Combining Statistical Methods and

Domain Knowledge". **MOLECULES**, **Special Issue "QSAR and Its Applications"**; p. 14937 – 14953 (2012).

Otras publicaciones y comunicaciones en conferencias no indexadas en Scopus

- **María J. Martínez**, Fiorella Cravero, Mónica F. Díaz, Ignacio Ponzoni. "QSPR Modeling Applied to High Molecular Weight Polymers: Ductility Characterization from Elongation at Break **VIII International Symposium on Materials**. **MATERIAIS 2017.** Aveiro, Portugal (2017).
- Fiorella Cravero, Santiago Schustik, **María J. Martínez**, Ignacio Ponzoni, Mónica F. Díaz. "Macro Approach to Molecular Modelling of Linear Polymers Applied to Estimation of Tensile Modulus for New Materials Development". **VIII International Symposium on Materials**. **MATERIAIS 2017.** Aveiro, Portugal (2017).
- **María J. Martínez**, Fiorella Cravero, Mónica F. Díaz, Ignacio Ponzoni. "Unsupervised Learning Based on Deep Learning Applied to the Identification of Applicability Domain of QSAR Models". **Fourth International Society for Computational Biology Latin America Bioinformatics Conference (ISCB-LA)**. CABA, Buenos Aires (2016).
- Fiorella Cravero**, María J. Martínez**, Mónica F. Díaz,Ignacio Ponzoni."Fuzzy Clustering: Identification of Similar Compounds for Virtual Screening in Rational Drug Design". **Fourth International Society for Computational Biology Latin America Bioinformatics Conference (ISCB-LA)**. CABA, Buenos Aires (2016).
- Fiorella Cravero, **María J. Martínez**, Mónica Fátima Díaz, Gustavo Esteban Vázquez, Ignacio Ponzoni. "An integral framework for QSAR modelling using computational intelligence and visual analytics". **VI Argentinian Conference on Bioinformatics and Computational Biology (6CAB2C)**. Lugar: Bahía Blanca, Buenos Aires (2015).
- Fiorella Cravero, **María J. Martínez**, Ignacio Ponzoni, Gustavo E. Vazquez, Mónica F. Díaz. "Desarrollo de Modelos QSPR asistido por técnicas de Analítica Visual para la Predicción de Propiedades Mecánicas de Polímeros Lineales". **XI Simposio Argentino de Polímeros**. Santa Fe (2015).

- Fiorella Cravero, **María J. Martínez**, Ignacio Ponzoni, Gustavo E. Vazquez, Mónica F. Díaz. "Predicción del Módulo Elástico para Polímeros Lineales Aplicando Analítica Visual y Aprendizaje Automático". **XI Simposio Argentino de Polímeros**. Santa Fe (2015).

- **María J. Martinez**, Fiorella Cravero, Gustavo E. Vazquez, Mónica F. Diaz, Axel J. Soto, Ignacio Ponzoni. "Interactive Visual Analysis Methodology for Improving Descriptor Selection in QSPR: First Steps". **V Congreso Argentino de Bioinformática y Biología Computacional (5CAB2C)**. Bariloche, Río Negro (2014).

- Damián Palomba, **María J. Martínez**, Fiorella Cravero, Axel J. Soto, Gustavo E. Vazquez, Ignacio Ponzoni, Mónica F. Díaz. "Predicción de Propiedades Mecánicas del Ensayo de Tensión para Polímeros Lineales. Modelado QSPR con Inteligencia Computacional y Análisis Visual Interactivo". **30º Congreso Argentino de Química**. CABA, Buenos Aires (2014).

- Damián Palomba, **María J. Martínez**, Ignacio Ponzoni, Mónica F. Díaz, Gustavo E. Vazquez, Axel J. Soto. "Prediction of blood to liver coefficients for volatile organic compounds: a cheminformatics approach". **III Congreso Argentino de Bioinformática y Biología Computacional (3CAB2C**) 2010. Oro Verde, Entre Ríos (2012).

- Axel J. Soto, **María J. Martínez**, Rocío L. Cecchini, Gustavo E. Vazquez, Ignacio Ponzoni. "A Prototype Software Tool for Selection of Relevant Descriptors in QSAR Models". **I Congreso Argentino de Bioinformática y Biología Computacional (1CAB2C)**. Quilmes, Buenos Aires (2010).

Índice General

Capítulo 1: Introducción

1.1 Aspectos Generales

La informática molecular es un área reciente de investigación destinada al desarrollo de métodos computacionales que faciliten el estudio de los fenómenos y principios que rigen a las moléculas y sus interacciones (Baumann et al., 2014). Este campo de investigación, que emerge de la Bioinformática y la Quimioinformática tiene múltiples áreas de aplicación, que van desde el diseño racional de fármacos (Gasteiger, 2016) hasta el diseño de materiales poliméricos (Afantitis et al., 2005; Palomba et al., 2014)

Dentro de la Informática Molecular, podemos encontrar como un área precursora el desarrollo de métodos para la inferencia de modelos QSAR/QSPR (Quatitative Structure-Activity/Property Relationship). Estos modelos permiten establecer relaciones entre características de un compuesto químico y una determinada propiedad fisicoquímica o biológica (Danishuddin y Khan, 2016). Es importante aclarar que los modelos QSAR/QSPR no sólo relacionan una propiedad a partir de características estructurales, sino que esta relación puede obtenerse a partir de cualquier descriptor, es decir, cualquier información determinable en forma objetiva sobre un aspecto de una molécula.

Hoy en día es necesario que las predicciones de métodos computacionales basados en QSAR/QSPR sean lo suficientemente precisas, para que pueda hacerse efectiva la aplicación de estos modelos en casos industriales reales de gran escala, de una manera confiable. La implementación de estos métodos para realizar experimentaciones *in silico*, representa un ahorro muy importante tanto desde el punto de vista económico, ya que se pueden hacer pruebas

virtuales a un muy bajo costo, como así también de tiempo y desarrollo (Gajewicz et al., 2012).

Para lograr obtener buenos modelos predictivos, se utilizan distintas metodologías de Aprendizaje Automático (Machine Learning), disciplina que consiste básicamente en detectar patrones y encontrar relaciones que permitan generalizar el conjunto de datos que es analizado (Bishop, 2006).

En esta tesis, la propuesta es integrar la Analítica Visual al proceso de desarrollo de modelos QSAR/QSPR y así aprovechar al máximo el análisis y la exploración de los datos combinando el razonamiento humano con el poder de procesamiento de las computadoras (Keim et al., 2010). Además, se busca analizar distintas técnicas de Aprendizaje Automático para diseñar e implementar metodologías que permitan determinar el dominio de aplicación de modelos de clasificación QSAR/QSPR.

1.2 Objetivos Generales y Específicos

El foco principal de esta tesis es la investigación de técnicas de analítica visual y aprendizaje automático aplicados en el modelado QSAR/QSPR. Lo que se busca es desarrollar metodologías que asistan en la inferencia de la estructura relacional presente entre moléculas, descriptores y propiedades fisicoquímicas o de actividad biológica. En tal sentido, las tareas de investigación son abordadas en un marco multidisciplinario que abarca la interacción con investigadores del área de Ciencias de la Computación, Farmacia, Química y también Bioinfomática. La principal contribución de este trabajo está vinculada con la Informática Molecular.

Más específicamente, podemos identificar dos etapas importantes en el modelado: el proceso de análisis y selección de descriptores moleculares, por un

lado, y la determinación del dominio de aplicación de los modelos desarrollados, por el otro (Cherkasov et al., 2014). En este contexto los objetivos específicos son:

- Proponer e implementar técnicas de analítica visual para contribuir al proceso de selección de descriptores, involucrando al experto en el dominio para poder lograr una mejor interpretación de los modelos en términos físicos químicos (no sólo en términos estadísticos).
- Desarrollar metodologías para la identificación del dominio de aplicación de un modelo de clasificación QSAR/QSPR.
- Generar nuevos modelos aplicados a la predicción de propiedades de interés en el diseño racional de fármacos, el diseño de materiales poliméricos y las ciencias ambientales.
- Comparar el desempeño de modelos generados con metodologías clásicas (modelos obtenidos mediante un proceso automatizado) versus modelos generados utilizando nuevos enfoques (agregando al proceso automatizado el conocimiento del experto).
- Testear y validar las técnicas implementadas con distintos casos de estudios.

1.3 Estructura y Contenido

Esta tesis se encuentra organizada en 7 capítulos. El capítulo 2 está orientado a introducir conceptos básicos de la analítica visual y de algoritmos de aprendizaje automático, para aquellos lectores que no estén familiarizados con estos conceptos. En el capítulo 3 se abordan los conceptos del modelado QSAR/QSPR, introduciendo nociones claves para entender el desarrollo de los capítulos siguientes. El capítulo 4 hace una presentación de diferentes metodologías para

la selección y aprendizaje de descriptores, así como también un análisis y aplicación de estas. También se presentan una serie de modelos de regresión generados para predecir una propiedad dentro del campo de las ciencias ambientales, utilizando un enfoque semi-automático combinado. En el capítulo 5 se presenta una de las principales contribuciones de esta tesis, una herramienta de analítica visual para analizar descriptores moleculares. Se analizan varios enfoques, tales como, la comparación de modelos generados utilizando selección de descriptores versus aprendizaje de descriptores y también utilizando una hibridización de estas dos técnicas, en distintos campos de aplicación. En el capítulo 6 se presenta otra de las contribuciones de esta tesis, referida a la identificación del dominio de aplicación para modelos QSAR/QSPR de clasificación y su validación con diferentes casos de estudio. Finalmente, en el capítulo 7 se presentan las conclusiones, principales aportes de esta tesis y trabajos futuros que podrían realizarse.

Capítulo 2: Conceptos de Analítica Visual y Aprendizaje Automático

En el presente capítulo se desarrollarán conceptos básicos relacionados a la Analítica Visual y al Aprendizaje Automático, por lo que está destinado a aquellos lectores que no poseen conocimientos en estas áreas. Se presentarán distintas definiciones y clasificaciones para comprender conceptos que serán utilizados en los siguientes capítulos. Para una lectura más profunda sobre Analítica Visual y Aprendizaje Automático se recomienda consultar los siguientes libros respectivamente: Thomas and Cook (2005), Keim et al. (2010) y Mitchell and McGraw (1997), Bishop (2006).

2.1 Conceptos de Analítica Visual

La Analítica Visual (AV) es un campo interdisciplinario emergente en los últimos años que tiene como objetivo combinar las habilidades racionales del ser humano con las visualizaciones interactivas. La implementación de herramientas y metodologías de AV aspiran a:

- Permitir la exploración de datos para inferir nuevo conocimiento.
- Facilitar la detección de patrones esperados y ofrecer posibilidades para descubrir aquellos inesperados.
- Permitir observar grandes cantidades de datos desde diferentes puntos de vista y de manera simultánea.
- Presentar resultados que contribuyan de manera clara en la toma de decisiones.

2.1.1 Introducción a la Analítica Visual

De manera más específica a lo mencionado anteriormente, se puede decir que la AV es un campo multidisciplinario que combina el razonamiento humano con las capacidades analíticas de los ordenadores (Keim et al., 2010). Su objetivo es proporcionar soluciones al problema de la sobrecarga de información (Levitin, 2014), así como también contribuir en la detección de patrones en los datos de manera exploratoria (Keim et al., 2006). Uno de los primeros enfoques en la analítica visual fue propuesto por Turkey (1977) mediante el análisis de datos interactivos. Posteriormente, las mejoras en las interfaces gráficas de usuario y dispositivos de interacción han contribuido a la investigación en la visualización de la información. Más tarde, se reconoció el potencial de integrar el conocimiento del usuario, lo cual llevó al desarrollo de los campos de la exploración visual de datos y la de minería de datos (Keim, 2001).

El término Analítica Visual fue utilizado por primera vez por Wong y Thomas en 2004 (Wong y Thomas, 2004). La AV fue definida como la ciencia del razonamiento analítico facilitada por interfaces interactivas de hombre-máquina (Thomas y Cook, 2005). Otra definición más actual propuesta en (Keim et al., 2010), dice que la AV combina técnicas de análisis automatizados con visualizaciones interactivas para una comprensión, razonamiento y toma de decisiones efectivos sobre la base de conjuntos de datos muy grandes y complejos.

2.1.2 Aplicación en Informática Molecular y Bioinformática

En diversas áreas, tales como biología, medicina, física, astronomía e inteligencia de negocios, entre muchas otras, podemos observar la aplicación de la AV. A

continuación, se hará especial foco en la importancia de aplicarla en los campos de Bioinformática e Informática Molecular.

Tanto la Bioinformática como la Informática Molecular son disciplinas en las que se utilizan diferentes técnicas computacionales para resolver problemas específicos de cada área. Entre estos problemas se encuentran: análisis de secuencias, expresión génica, redes biológicas y biología de sistemas, en Bioinformática; y diseño racional de fármacos, modelado QSAR/QSPR y cribado virtual, en Informática Molecular (Attwood et al., 2016; Gasteiger, 2016). Todos ellos implican el análisis y manejo de grandes cantidades de datos biológicos y/o químicos. En este sentido, el análisis de datos es uno de los mayores desafíos en estas interdisciplinas debido a la gran cantidad de información disponible y la complejidad de dicha información. Cuestiones como la procedencia de los datos (estos pueden obtenerse de diferentes dominios y almacenarse en distintos formatos) y la complejidad y heterogeneidad de los mismos, están a la orden del día. La integración de esta información tan diversa se convierte en una cuestión clave para el análisis (Alyass et al., 2015; Basak et al., 2015).

Planteado este escenario, la aplicación de VA se vuelve esencial debido a la gran cantidad de datos a analizar, aportando entre otras ventajas, el análisis interactivo y exploratorio de los datos. En este sentido, desde hace algunos años que la AV viene siendo aplicada a diferentes problemas en el campo de la Bioinformática. Algunos ejemplos son: (i) SpRay (Dietzsch, Heinrich, Nieselt, y Bartz, 2009) para la exploración visual de datos de expresiones de genes; (ii) GenAMap (Curtis et al., 2011) un sistema de AV que permite explorar la estructura del genoma humano; (iii) BiNA (Gerasch et al., 2014) para el análisis visual de datos de redes biológicas; y (iv) una plataforma de AV presentada recientemente para genómica (Freese et al., 2016). En el área de Informática Molecular no existen tantos problemas donde se ha aplicado AV en toda su dimensión. Algunos ejemplos que entran dentro de esta categoría son: (i) Scaffold Hunter (Wetzel et al., 2009) y ConTour (Partl et al., 2014), las cuales presentan herramientas para el análisis de datos moleculares y sus interacciones; (ii) QSARINS (Gramatica et al.,

2013) y ChesMapper 2.0 (Gütlein et al., 2014), las cuales se utilizan para el modelado QSAR/QSPR. En este área se encuadra uno de los aportes realizados en el marco de esta tesis; una herramienta de AV para el análisis visual e interactivo de descriptores moleculares, llamada VIDEAN (Martínez et al., 2015).

2.1.3 Proceso de la Analítica Visual

En el proceso de la AV se combina el análisis de datos automatizado con el conocimiento del experto para lograr una efectiva exploración de los datos. En la Figura 2.1, se muestra una representación esquemática de este proceso, planteada en el libro de Keim et al. (2010). Por un lado, tenemos los datos, los cuáles habitualmente suelen ser pre-procesados o transformados para derivar diferentes representaciones de los mismos. Luego, el usuario puede decidir entre realizar una exploración visual de los datos o aplicar algún método automatizado. En el caso de llevar a cabo un análisis automatizado, se utilizará algún método de minería de datos con el fin de generar un modelo. Este modelo podrá ser evaluado y mejorado, utilizando la exploración visual del mismo. La alternancia entre los métodos visuales y automáticos es habitual en el proceso de AV y lleva a un perfeccionamiento y verificación continuos de los resultados preliminares que se van obteniendo. De esta manera se incorpora el conocimiento del usuario al proceso.

En el caso en que la primera etapa elegida sea la exploración visual, el usuario debe confirmar su hipótesis utilizando algún método automatizado. Los resultados que se van obteniendo en las visualizaciones se pueden usar para dirigir ese análisis.

Por último, el conocimiento en el proceso de la AV puede obtenerse tanto de la visualización y el análisis automatizado, como de las interacciones entre las visualizaciones, los modelos y los usuarios.

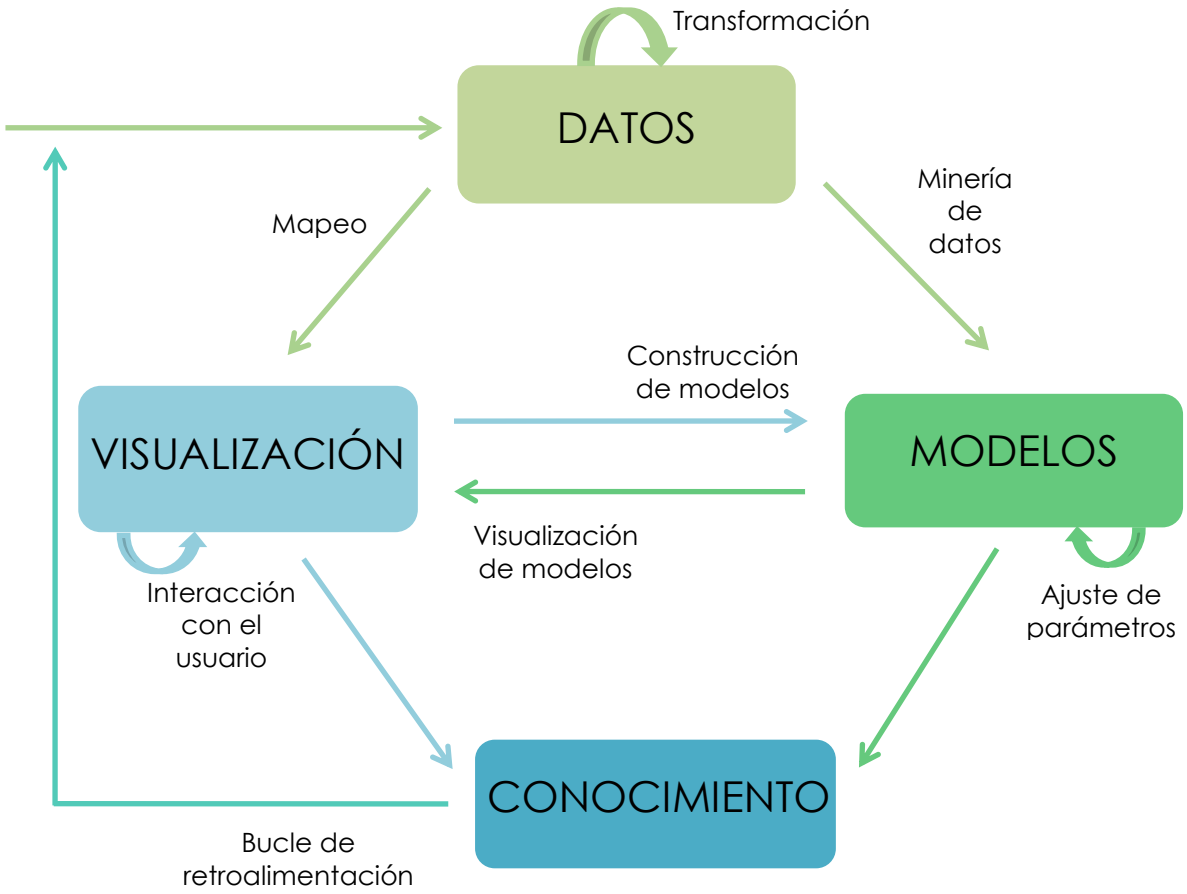


Figura 2.1 Proceso de la Analítica Visual. Este esquema muestra el flujo entre los datos, los modelos creados a partir de los datos, la visualización y el conocimiento del experto. La línea del proceso que involucra a los modelos (en la parte derecha de la figura) representa el análisis automatizado de los datos, mientras que la línea que involucra a la visualización (parte izquierda de la figura) representa la exploración visual de los datos.

Como se mencionó al comienzo de la sección, la AV integra métodos y características de distintas disciplinas. En la Figura 2.2, se puede ver un esquema con las disciplinas relacionadas y el soporte necesario para que la vinculación entre éstas sea posible, tal como se presenta en libro de Keim et al. (2010). La *visualización* es el núcleo de esta integración. A continuación, se enumeran y explican de manera breve, cada una de las disciplinas involucradas.

- Visualización: generalmente suele clasificarse en *visualización científica* y *visualización de información*. La primera, se utiliza para la visualización de

fenómenos reales (por ejemplo: médicos, biológicos o meteorológicos), y se hace hincapié en la representación exacta de tales fenómenos. El segundo tipo de visualización, a diferencia de la primera, se centra en la representación de información abstracta de una manera intuitiva. Ejemplos de ambas técnicas de visualización pueden ser: imágenes y animaciones tridimensionales para el caso de visualización *científica* y grafos, mapas de árboles, histogramas, coordenadas paralelas, mapas, entre muchos otros para visualización *de información*. Esta clasificación de la visualización ha generado varios cuestionamientos en los últimos años, debido a si realmente sigue siendo necesario realizar esta diferenciación entre las visualizaciones *científicas* y *de información*. En Rhyne (2003) y Rhyne et al. (2003) se pueden encontrar discusiones al respecto. En el primer reporte, se hace un repaso de la perspectiva histórica de esta clasificación y se analiza si es realmente necesaria; mientras que, en el segundo, investigadores del área brindan su opinión respecto a este tema.

- Percepción y Cognición humana: en el contexto de la AV, estos conceptos están relacionados con la interacción entre el ser humano y la computadora. La percepción visual es la forma en que el cerebro interpreta los estímulos para hacer una interpretación de lo que ve. Por otra parte, la cognición es la capacidad humana de conocer a través de la percepción, es decir, la capacidad de razonar, entender y hacer inferencias a partir de la información que se analiza. Estos conceptos son esenciales en VA, ya que proporcionan la parte del conocimiento humano al proceso. En este sentido, existen trabajos como Green et al. (2009) y Hegarty (2011) que brindan recomendaciones para que las herramientas de VA integren de una manera adecuada las capacidades de los seres humanos y de las computadoras.
- Manejo de Datos: es otro de los conceptos fundamentales ya que implica el manejo de las bases de datos que utilizaremos para analizar. Hoy en día estas bases de datos pueden contener datos heterogéneos (números,

texto, audio, video, datos no estructurados) y se debe encontrar una buena representación de toda la información, así como también poder llevar a cabo un proceso de descarte de datos perdidos o inexactos. En la actualidad está aumentando el uso de diferentes técnicas de análisis para hacer frente a estos problemas de datos, y así poder procesar los datos automáticamente (Dharma Rahadi et al., 2016; Xiaofeng y Xiang, 2013).

- Minería de Datos: se refieren al proceso computacional mediante el cual se intenta descubrir determinados patrones en grandes cantidades de datos. Para lograr este objetivo se utilizan métodos de aprendizaje automático, estadística y sistemas de bases de datos. Un análisis profundo de esta disciplina puede encontrarse en el libro de Witten et al. (2016).

- Análisis de datos espacio-temporales: los datos espacio-temporales se han vuelto muy importantes en los últimos años. Esto se debe a que muchas aplicaciones del mundo real, tales como los sistemas de información geográfica, los servicios basados en localización, entre otros, necesitan almacenar características espaciales y/o temporales de los datos. En este sentido, existen varios trabajos en los que se presentan herramientas de analítica visual para modelar y analizar este tipo de datos, entre ellos podemos citar a Guo et al. (2009), Hao et al. (2011) y Andrienko y Andrienko (2013).

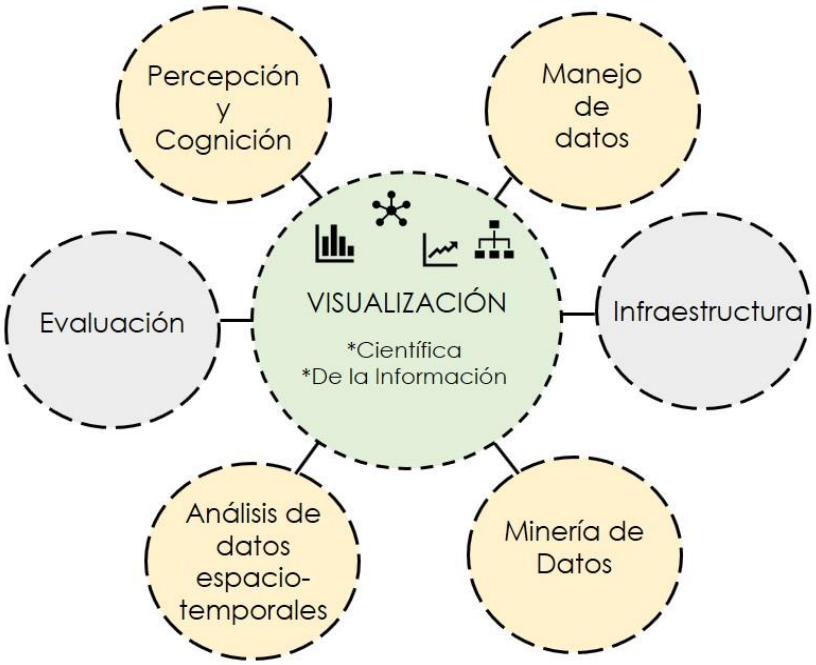


Figura 2.2 Disciplinas integradas en la Analítica Visual. En este diagrama se puede observar que la AV integra la visualización (tanto científica como de la información) con otras disciplinas (percepción y cognición, manejo de datos, minería de datos y análisis de datos espacio-temporales). Mientras que, a otro nivel, esta integración se hace posible dependiendo de la disponibilidad de infraestructura y las posibilidades de evaluación.

- Evaluación: evaluar la eficacia de una herramienta de VA es un tema importante y no trivial. La evaluación es difícil debido a la naturaleza exploratoria de la analítica visual, la potencial variedad de usuarios y la diversidad de datos. Hay muchos enfoques diferentes para realizar la evaluación. En Carpendale (2008) se presenta una visión general de las diferentes metodologías empíricas para la evaluación. Hace especial hincapié en lograr un equilibrio entre generalidad, precisión y realismo; además presenta un análisis exhaustivo de las diferentes técnicas de evaluación cualitativa y cuantitativa. Un enfoque que pretende analizar la interacción entre el diseño y la evaluación de una herramienta VA se presenta en Munzner (2009). Aquí, el autor desarrolla un modelo con cuatro

capas y proporciona orientación para determinar los enfoques de validación mediante la identificación de potenciales problemas existentes en cada capa. Otro trabajo más reciente de revisión exhaustiva relacionado con métodos empíricos es Lam et al. (2012), en el cual se presentan siete escenarios relacionados con diferentes aspectos de la evaluación. En el mismo se brindan ejemplos de casos de estudio y se plantean diferentes preguntas con el objetivo de guiar al lector en cómo realizar una experiencia propia en la evaluación de una herramienta.

- Infraestructura: es de suma importancia para asegurar que las herramientas de AV funcionen de manera eficiente y efectiva. La infraestructura vincula todos los procesos, funciones y servicios que requieran estas herramientas para brindarle una experiencia armoniosa al usuario que está explorando e interactuando con los datos. En general, no resulta una tarea sencilla debido a que cada aplicación de AV requiere que se utilicen las tecnologías apropiadas, así como también que se puedan reutilizar distintos componentes comunes entre aplicaciones, para obtener herramientas más adaptables y rápidas de desarrollar.

2.2 Conceptos de Aprendizaje Automático

El Aprendizaje Automático (AA) es una rama de la inteligencia artificial, cuyo objetivo es desarrollar técnicas que le permitan a las computadoras *aprender* patrones en los datos. De manera más específica, esta disciplina busca diseñar algoritmos capaces de generalizar comportamientos a partir de cierta información suministrada como ejemplos, induciendo así al conocimiento.

2.2.1 Introducción al Aprendizaje Automático

El AA se focaliza en la creación de modelos, a partir de un conjunto de datos proporcionado como entrada. Los datos son la materia prima para la construcción de los modelos y son llamados *datos de entrenamiento* (Figura 2.3), ya que son ejemplos que le sirven al modelo para poder entrenar y aprender de ellos. Generalmente, se presentan en un formato de tabla, donde cada fila representa una *instancia* (también puede llamarse *ejemplo* o *muestra*) y cada columna un *atributo* (también llamada *característica o descriptor*). Una de las columnas de la tabla (normalmente la última) representa un atributo especial, llamado atributo objetivo o destino, que es el que asigna a cada instancia de la tabla una etiqueta determinada (valor real o etiqueta de clase) y es el atributo que se quiere predecir. Siendo este el caso, estaremos hablando de un enfoque de aprendizaje *supervisado*, ya que tenemos un conjunto de datos de entrenamiento donde cada instancia está asociada a un valor (atributo objetivo o destino). Si el conjunto de datos no tiene este valor asociado, estaremos frente a un enfoque de aprendizaje *no supervisado*. Hablaremos con más detalle de los diferentes enfoques en la próxima sección.

Una vez definidos los datos de entrada, el paso siguiente es aplicar algún método de AA para generar uno o más modelos. Un modelo es una expresión matemática que describe como se relacionan un conjunto de atributos. Algunos ejemplos de métodos para generar modelos son: regresión lineal, redes neuronales y árboles de decisión, entre muchos otros que explicaremos en detalle más adelante. Estos métodos ofrecen como resultado el modelo propiamente dicho (en enfoque supervisado), o el descubrimiento de datos similares o la detección de patrones en los datos (en el caso de un enfoque no supervisado).

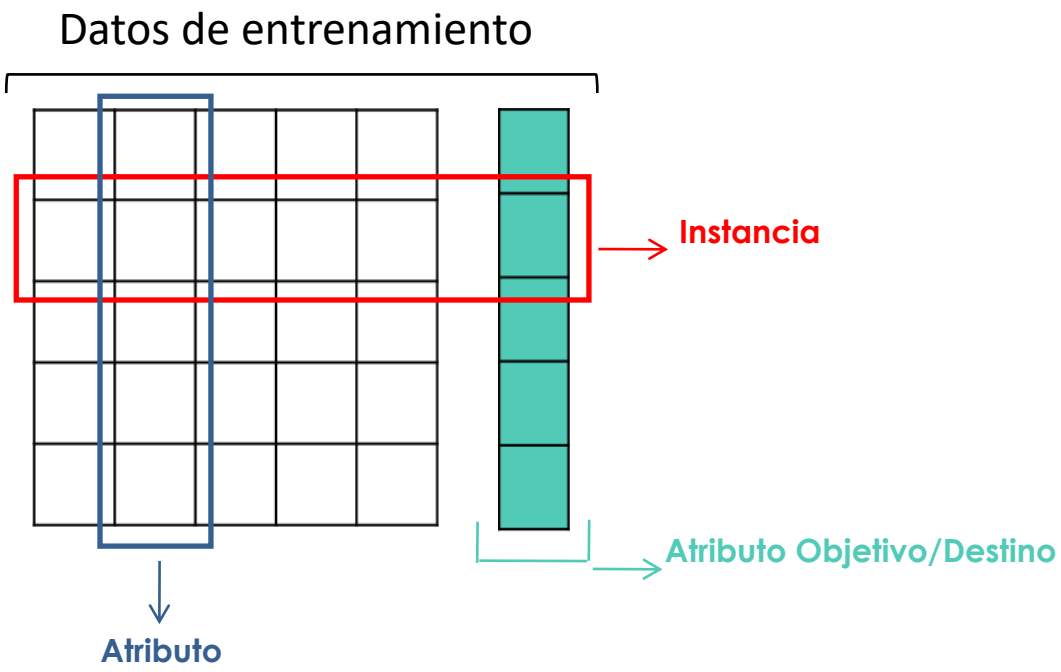


Figura 2.3 Representación del formato de los datos de entrenamiento

A su vez, en el caso de los modelos supervisados, para evaluar la capacidad predictiva del modelo, existen diferentes mecanismos. Realizar la evaluación sobre el mismo conjunto de datos de entrenamiento es inadecuado, ya que se están utilizando para evaluar los mismos datos con los que el modelo aprendió. Es por esto que para la evaluación de los modelos se utiliza un conjunto de datos (comúnmente llamado *conjunto de testeo*) distinto e independiente a los datos de entrenamiento. De esta manera se busca obtener un resultado no sesgado de la capacidad predictiva de un método particular. En la Figura 2.4 se puede observar un esquema básico de los pasos para un obtener un modelo y evaluarlo, utilizando métodos de AA.

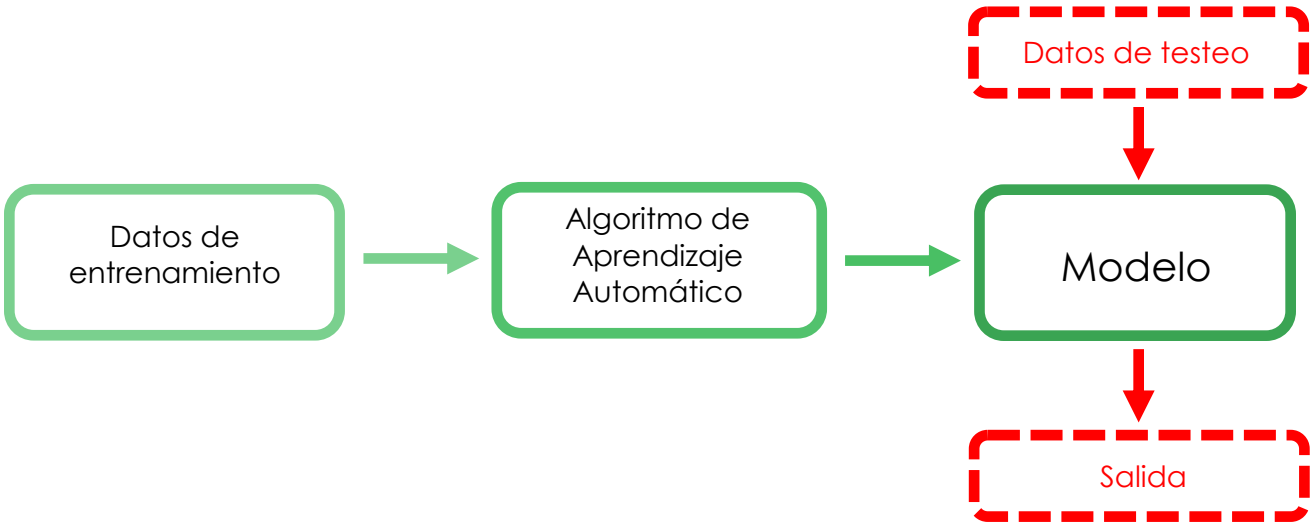


Figura 2.4 Esquema básico para la obtención de un modelo. El usuario interviene en la definición de los datos (entrenamiento y testeo), la definición de los parámetros del algoritmo y el análisis del rendimiento de la salida del modelo.

Dependiendo del tipo de salida que tenga el método de aprendizaje automático y de la forma en qué son tratados los ejemplos, se presentan dos enfoques bien diferenciados: *Supervisado*, dentro del cual se encuentran los algoritmos de regresión y clasificación y *No Supervisado*, que comprende los algoritmos de agrupamiento y reglas de asociación. También puede encontrarse el enfoque *Semi Supervisado*, que se plantea como una hibridización entre los dos enfoques nombrados anteriormente, en la que una parte de los datos puede tener asociada un atributo objetivo y otra parte no. En las siguientes secciones se abordarán los dos principales enfoques aquí mencionados.

2.2.2 Enfoque Supervisado

El objetivo del aprendizaje supervisado es el de crear una función capaz de predecir el valor que le corresponda a cualquier objeto de entrada válido después de haber "aprendido" de los datos de entrenamiento. La salida de la

función puede ser un valor numérico (por ejemplo en los problemas de regresión) o una etiqueta de clase (por ejemplo en los problemas de clasificación). Para esto, en el proceso de aprendizaje se debe poder generalizar, a partir de los datos de entrada, las situaciones no vistas previamente.

Antes de describir cada método, haremos una breve descripción de las diferencias entre regresión y clasificación, y las métricas estadísticas que se utilizan para evaluar el comportamiento en cada caso.

2.2.2.1 Regresión versus Clasificación

La diferencia concreta entre estos dos tipos de algoritmos reside en el tipo de objetos que intentan predecir.

En estadística, cuando se habla de *regresión* se hace referencia al proceso mediante el cual se estima la relación entre variables. Los algoritmos de regresión intentar armar un modelo para predecir un valor real.

En contrapartida a la regresión, la *clasificación* intenta predecir a qué clase pertenecerá un nuevo objeto, a partir de un conjunto de clases prefijadas. Estas clases pueden ser dos, en este caso estamos frente a una clasificación binaria; o más, en el caso de una clasificación múltiple.

Generalmente, cuando se va abordar un nuevo problema de AA, una de las primeras cosas que se debe determinar es en qué categoría de estas dos se va a enmarcar. A continuación, se detallan algunas de las métricas más utilizadas para evaluar el rendimiento de estos modelos.

2.2.2.2 Métricas Estadísticas

Luego de aplicar un algoritmo de AA se obtienen como resultado distintas métricas estadísticas.

Para el caso de *regresión*, las métricas más comunes son las siguientes:

- Coeficiente de correlación de Pearson (*r*): es una medida del grado de correlación o asociación lineal que tienen dos o más variables cuantitativas. Se expresa como un número entre -1 y +1. Los valores cercanos a +1 indican una correlación lineal positiva fuerte y los valores cercanos a -1 indican una correlación lineal negativa fuerte. Los valores que están cercanos a cero indican una no correlación entre las variables.

- Coeficiente de determinación (r^2): es una medida de la bondad de ajuste de los datos. Indica la proporción de la varianza que puede ser predicha. Un valor cercano a 1, indica mejores predicciones. Se define como la proporción de la variación total en y que puede explicarse usando un modelo lineal $\hat{y} = b_0 + b_1 x$.

- Error absoluto medio: es el promedio de los errores absolutos de cada instancia. Cuantifica cuán cercana están las predicciones de los valores reales.

- Error cuadrado medio: es el promedio de los errores al cuadrado y una de las medidas más utilizadas para evaluar una predicción.

En los casos de *clasificación*, se suelen reportar las siguientes métricas:

- Porcentaje de casos correctamente clasificados: es decir, el porcentaje de casos para los que la predicción fue realizada correctamente.
- Área de la Curva de Característica Operativa del Receptor: conocida como área de la curva de ROC (del inglés, *Receiver Operating Characteristic*) (Hanley, 1982). Es utilizada para comparar el rendimiento de un número de modelos. La curva de ROC se dibuja en el espacio definido por la tasa de verdaderos positivos (es decir, la tasa de aquellas instancias que han sido clasificadas correctamente) como eje de las abscisas, y la tasa de falsos positivos (es decir, la tasa de aquellas instancias que han sido clasificadas erróneamente) como eje de las ordenadas. El mejor modelo

sería aquel que tiene una tasa de verdaderos positivos alta y una tasa de falsos positivos baja. Un índice de bondad del modelo es el área bajo la curva de ROC. En este sentido, cuanto mayor sea el área mejor será la predicción.

Por otro lado, podemos hacer mención a la matriz de confusión, que más que una métrica podría considerarse una representación gráfica que permite la visualización del rendimiento de un clasificador discriminado por clase. Cada columna de la matriz representa las instancias en la clase "real" y cada fila representa las instancias en la clase "predicha" (o viceversa). Los elementos que aparecen en la diagonal representan el número de instancias que fueron correctamente predichas, mientras que los elementos fuera de la diagonal representan el número de instancias clasificadas de manera errónea. Cuanto más alto sean los valores de la diagonal, mejor será el rendimiento del modelo.

2.2.2.3 Aspectos a tener en cuenta en el modelado

La evaluación de un modelo es una etapa muy importante, ya que define la confiabilidad y significancia estadística del modelo (Gramatica, 2006). Existen distintas alternativas dependiendo de cómo se dividen los datos en entrenamiento y testeo:

- Validación Interna: es una técnica que se aplica cuando el conjunto de datos es pequeño. Se parte de un conjunto de datos inicial del cual se separan algunos ejemplos que luego serán utilizados para evaluar el modelo. La técnica más conocida y utilizada es la validación cruzada de *k* iteraciones (*k* – fold *cross validation*). Consiste en dividir el conjunto de datos (de manera aleatorio o estratificada) en *k* grupos. Se utilizan *k*-1 grupos de datos para entrenar el modelo y el grupo restante se utiliza

como conjunto de testeo. Este procedimiento se repite k veces, obteniéndose *k* medidas diferentes de rendimiento, las cuáles pueden utilizarse para calcular la media (indicador del rendimiento del modelo) y la varianza (indicador de la capacidad de generalización del modelo). Este procedimiento es repetido N veces. De esta manera, el modelo es entrenado con *k* conjuntos de entrenamiento distintos, lo que permite hacer una buena estimación de su capacidad predictiva y generalización. Una variante de esta alternativa es "*dejar uno afuera*" (*leave one out cross validation*). Es una técnica muy poco usada actualmente ya que surgió en un momento en que los conjuntos de datos eran muy pequeños. Constituye un caso especial de la validación cruzada de k-folds, donde k=n (n cantidad de ejemplos). Consiste en separar el primer ejemplo para testeo y dejar el resto como conjunto de entrenamiento. Luego, separar el segundo ejemplo para testeo y continuar de la misma manera hasta que todos los ejemplos se hayan separado una vez para utilizarse para el testeo. De esta manera se van a obtener tantas medidas de rendimiento como ejemplos haya en el conjunto de datos. De la misma manera que en la validación cruzada de *k* iteraciones, estos datos son combinados para obtener un resultado global del rendimiento del modelo.

- Validación Externa: en este tipo de validación, el conjunto inicial es dividido en tres conjuntos de datos: *conjunto de entrenamiento*, *conjunto de validación* y *conjunto de testeo*. El primero se utiliza para entrenar el modelo, el segundo es utilizado para estimar el error de predicción con el fin de comparar modelos y ajustar parámetros. Por último, el tercer conjunto es usado para evaluar la generalidad del modelo final. El conjunto de testeo debería ser lo suficientemente grande (lo que implica que el conjunto inicial de datos debe ser grande) para que la estimación de la calidad predictiva del modelo sea representativa, sobre todo si la construcción de este conjunto se ha realizado de manera aleatoria.

Si bien la opción de utilizar tres conjuntos para entrenar un modelo es la adecuada, existe otro enfoque en el que se divide el conjunto inicial en *conjunto de entrenamiento* y *conjunto de testeo*. El primer conjunto se utiliza para entrenar al modelo que luego es evaluado con el *conjunto de testeo* (Alexander et al., 2015).

Otro de los aspectos a tener en cuenta al entrenar un modelo, es el sobre-ajuste de los datos. Como se ha mencionado anteriormente, un algoritmo de AA debe alcanzar un estado en el que será capaz de predecir nuevos datos a partir de lo aprendido con los datos de entrenamiento. Sin embargo, cuando un modelo es sobre-ajustado suele amoldarse de manera precisa al conjunto de datos con el que ha sido entrenado, pero no es generalizable ante nuevos datos. Existen diferentes alternativas para evitar el sobre-ajuste, entre ellas podemos mencionar a modo de ejemplo, el *dropout* presentado en (Srivastava et al., 2014).

En las siguientes subsecciones serán explicadas las técnicas de aprendizaje automático utilizadas en esta tesis.

2.2.2.4 Regresión Lineal

La regresión lineal es un método matemático que modela la relación entre una variable de salida (y), las variables independientes (x_i) y un término de error aleatorio (ε). En el caso de la regresión simple, se tiene una única variable independiente x, $x \in R$. La regresión lineal múltiple, es una extensión de la simple donde la variable independiente x es un vector, tal que $x \in R^n$. El modelo lineal se puede expresar entonces de la siguiente manera:

$$y = \beta_o + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n + \varepsilon$$

Los coeficientes β_i , $i \in \{1..n\}$ asociados a cada variable independiente se determinan de manera de que se minimice el cuadrado del error ε. Si bien es una manera rápida y sencilla de modelar datos, muchas veces los datos (como es el

caso de las estructuras químicas y distintas propiedades) no tienen un comportamiento lineal. Se debe tener en cuenta que para que se pueda utilizar adecuadamente la regresión lineal, las variables deben ser linealmente independientes entre sí.

2.2.2.5 Árboles de Decisión

En un árbol de decisión, los datos son divididos recursivamente en conjuntos más pequeños con particiones binarias. En cada iteración del método, se evalúan distintas particiones (evaluando todo el conjunto los datos) y se elige la mejor. Para medir que tan buenas son las particiones existen distintos enfoques, como métodos basados en ganancia de la información (Quinlan, 1986), o métodos que utilizan el índice de Gini (Breiman et al., 1984). La división de los datos genera como salida del método una estructura de árbol, donde cada nodo representa a una de las variables de entradas. Cada rama refleja el resultado de una prueba e indica el camino a recorrer dependiendo del valor de la variable. Cada nodo hoja del árbol representa un valor de la variable destino. Es decir que el valor predicho de la variable destino es obtenido por el camino recorrido desde la raíz hasta una hoja del árbol.

Estos modelos son muy fáciles de interpretar, gracias a su construcción lógica y tienen la capacidad de modelar relaciones no lineales. Una revisión y análisis de los algoritmos más ampliamente disponibles puede encontrarse en el trabajo de Loh (2011).

2.2.2.6 Bosques Aleatorios (Random Forest)

Este método fue planteado por Breiman (2001) y surge como una mejora a los árboles de decisión. Básicamente, el algoritmo de Bosques Aleatorias (BA) genera un conjunto arbitrario de árboles de decisión independientes que son testeados

sobre conjuntos de datos aleatorios que tienen el mismo número de variables seleccionadas al azar. Esta aleatoriedad, le permite mejorar la precisión de las predicciones. Más específicamente, en cada punto de división del árbol, la búsqueda de la mejor variable para dividir los datos no es realizada sobre todo el conjunto de variables, sino sobre un subconjunto de las mismas generado de manera aleatoria. Luego, se busca la mejor división de los datos teniendo en cuenta sólo ese subconjunto de variables.

BA es eficiente en términos de tiempo de ejecución para ser utilizado con conjuntos de datos muy grandes. Puesto que cada árbol se construye de manera independiente, BA ofrece la opción de ser paralelizado eficazmente. Tiene la capacidad de manejar un número muy grande de variables teniendo en la cuenta la contribución de cada una de ellas. En comparación con los árboles de decisión, el modelo es más compleja de interpretar.

2.2.2.7 Redes Neuronales Artificiales

Las redes neuronales artificiales (del inglés, *Artificial Neural Networks*, ANN) constituyen una familia de técnicas de modelado no lineal, inspiradas en el funcionamiento neuronal del ser humano y en sus capacidades para memorizar y asociar hechos. Una red neuronal consiste de unidades llamadas neuronas que se encuentran interconectadas. Cada una de esas neuronas recibe una serie de entradas y reacciona a ellas produciendo una salida. Las neuronas de la red intercambian datos y tienen la capacidad de aprender y mejorar su funcionamiento a partir de lo aprendido. Las neuronas de la red reciben datos desde el exterior, los procesan utilizando funciones de transformación y producen una señal de salida. Una arquitectura típica de una red neuronal está formada por una *capa de entrada* donde las neuronas definen todos los valores de las variables de entrada para el modelo; *una o más capas ocultas* de neuronas que reciben información de la *capa de entrada* y una *capa de salida* que presenta finalmente el valor de la predicción.

Uno de los algoritmos más conocidos para entrenar redes neuronales es *back-propagation (propagación hacia atrás)*(Hecht-Nielsen, 1988). Este algoritmo repite un ciclo de dos fases: propagación y actualización de pesos. Funciona de la siguiente manera: un vector de entrada se propaga hacia adelante a través de la red, capa por capa, hasta que alcanzar la capa de salida. La salida obtenida se compara entonces con la salida deseada, y se calcula un valor de error para cada una de las neuronas de la capa de salida. Una vez realizado esto, los valores de error son propagados hacia atrás, comenzando desde la salida, hasta que cada neurona tiene un valor de error asociado que representa aproximadamente su contribución a la salida original. Además, los pesos son actualizados de manera tal que, en el siguiente ciclo, las predicciones se acerquen cada vez más a la salida deseada.

Por otro lado, otro algoritmo utilizado con frecuencia es el *perceptrón multicapa (multilayer perceptron)* que hace uso de la *propagación hacia atrás* para entrenar las redes (Riedmiller, 1994).

2.2.3 Enfoque No Supervisado

En este enfoque el modelo es ajustado a las observaciones. La principal diferencia con respecto al enfoque *Supervisado* es que sólo requiere instancias y no se tienen un valor de atributo objetivo, es decir, no hay datos etiquetados.

2.2.3.1 Agrupamiento o Clustering

El agrupamiento es una técnica de asociación que tiene como objetivo agrupar datos (generalmente instancias), de manera tal que los elementos en un mismo grupo (cluster) sean similares entre sí y diferentes de los elementos en otros grupos. Existen dos tipos de métodos: los de partición y los jerárquicos. Los de partición intentan clasificar los datos definiendo primeramente una cantidad de grupos y

luego determinando de qué manera quedarán conformados esos grupos. No existe una fórmula que nos indique de manera general el número de grupos que se necesitan, sino que depende de la naturaleza de los datos y de los resultados que se quieran obtener. Por otro lado, en los métodos jerárquicos se busca construir grupos anidados, y esto puede realizarse tanto de manera aglomerativa o divisiva. Para un análisis más profundo de estos temas se recomienda la lectura de (Kaufman y Rousseeuw, 2009).

Una alternativa al agrupamiento convencional es el agrupamiento difuso (Bezdek et al., 1984). Este surge de la necesidad de resolver una deficiencia en el enfoque de agrupamiento fijo: en éste, cada instancia puede agruparse únicamente con los elementos de su grupo y, por lo tanto, no puede ser similar al resto de los elementos. En el agrupamiento difuso esto cambia radicalmente, ya que una instancia puede pertenecer a más de un grupo. Para esto, cada instancia tiene asociada un grado de pertenencia a un grupo. Esto se logra aplicando una función de pertenencia, que representa la similitud entre un elemento y un grupo. Un grado de pertenencia alto indica mayor similitud entre el elemento y el resto de los elementos del grupo, mientras que valores bajos indican poco grado de similitud entre ellos. Más detalles sobre las características de este método de enfoque difuso serán brindados en el capítulo 6.

2.2.3.2 Mapas auto-organizativos (SOM)

Un mapa auto-organizativo (*Self-Organizing Map*, SOM) es un tipo de red neuronal artificial, la cuál es entrenada utilizando un enfoque no supervisado (Kohonen, 1990). Un SOM está formado por dos capas de neuronas. La capa de entrada) es la encargada de recibir y transmitir a la capa de salida la información que proviene del exterior. La capa de salida genera un mapa, que es una representación discreta y bidimensional del espacio muestral de los datos de entrada. Este mapa está formado por nodos y, asociado con cada nodo existe un vector de pesos (de la misma dimensión que las variables de entrada) y una

posición dentro del mapa. Los SOM utilizan también una función de vecindad para preservar las propiedades topológicas del espacio de entrada. Para ubicar un vector de datos en el mapa, se debe encontrar el nodo que tenga el vector de pesos más cercano (menor distancia) a este vector de entrada en el espacio de datos. De esta manera un SOM define una proyección desde un espacio de datos de alta dimensionalidad a un mapa bidimensional de neuronas.

2.3 Sumario

Este capítulo fue dedicado a introducir conceptos básicos en las áreas de la Analítica Visual y el Aprendizaje Automático. Se describió en qué consiste la analítica visual, como así también su definición multidisciplinaria, además de presentar la importancia de aplicarla en las áreas de Informática Molecular y Bioinformática.

En relación al Aprendizaje Automático, se definieron distintos enfoques para el modelado de datos. Además, se analizaron distintas cuestiones de relevancia para los modelos, como es el caso de su evaluación.

Capítulo 3: Conceptos de Modelado QSAR/QSPR

En el presente capítulo se introducirán conceptos de modelado QSAR/QSPR (del inglés, *Quantitative Structure Property-Activity Relationship*) que luego serán utilizados en los capítulos siguientes. Para una lectura más amplia sobre este tema se recomiendan los siguientes libros de Puzyn et al. (2010) y Roy et al. (2015).

3.1 Modelos cuantitativos de relación estructura/propiedad o estructura/actividad

La inferencia de modelos QSAR/QRPR fue puesta en práctica por primera vez hace alrededor de cincuenta años (Hansch, 1969). A partir de ese momento, este enfoque viene mejorando sin pausa, con avances interdisciplinarios y desarrollos impulsados por la comunidad científica. Esto hace que sea una de las técnicas más comúnmente empleada para modelar propiedades físico-químicas o de actividad biológica de distintos químicos o sistemas biológicos. El modelado QSAR/QSPR ha evolucionado comenzando con su aplicación sobre pequeños conjuntos de moléculas y utilizando métodos de regresión relativamente simples para su análisis (en sus comienzos), hasta el análisis de grandes cantidades de compuestos a través de diferentes técnicas estadísticas y de aprendizaje automático. (Cherkasov et al., 2014).

Los modelos QSPR establecen relaciones entre características estructurales de un compuesto químico y una propiedad fisicoquímica. Comúnmente, QSPR se escribe en forma intercambiable como QSAR ya que, si bien los métodos basados

en QSAR relacionan la estructura química con la actividad biológica, la esencia de los métodos es la misma. En el transcurso de esta tesis haremos explícita esta diferencia cuando sea necesario, de lo contrario al leerse QSPR, se debe tener en cuenta que lo mismo puede ser aplicado para QSAR (o viceversa).

La aplicación de este tipo de modelado, surge de la necesidad de que todas las sustancias químicas sean probadas en términos de su toxicología y propiedades ambientales antes de su uso; de una manera rápida, a un muy bajo costo y reduciendo el número de animales usados para experimentos (Gajewicz et al., 2012). Este tipo de experimentación recibe el nombre de *in silico* y de ninguna manera pretende reemplazar a los clásicos experimentos *in vivo* e *in vitro*, sino que apunta a agilizar procesos, como por ejemplo analizar un conjunto de compuestos, sin necesidad de sintetizar las moléculas (Figura 3.1). Si bien el análisis presentado aquí está enfocado en el área química, es válido de la misma manera para el diseño de materiales (donde la etapa de experimentación *in vivo* no se lleva a cabo).

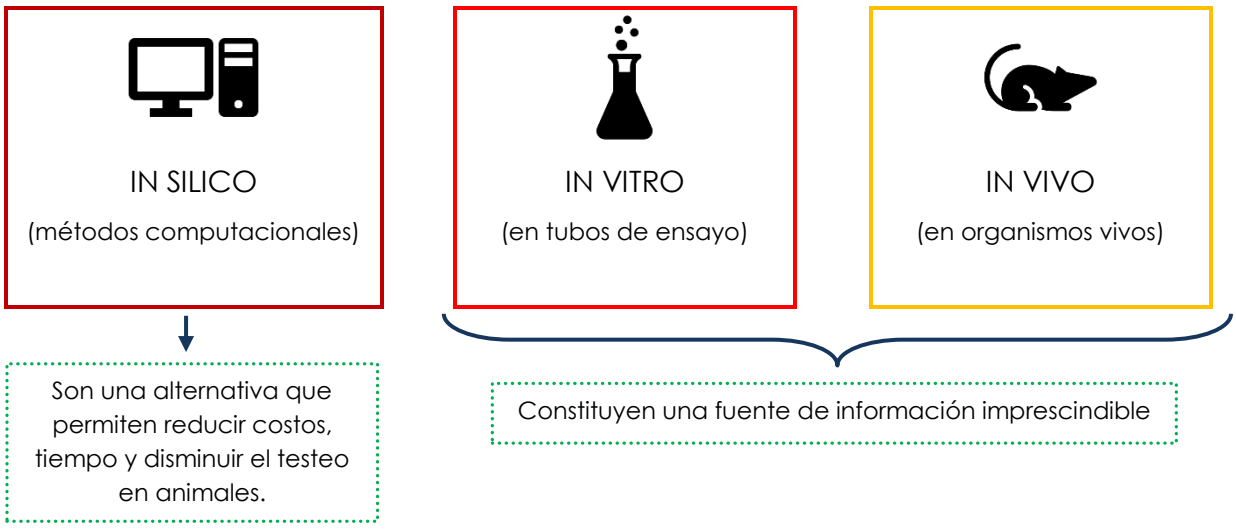


Figura 3.1 Esquema de las distintas alternativas para testear compuestos.

3.1.1 Pasos principales en el modelado QSAR/QSPR

Desarrollar modelos QSPR implica llevar a cabo una serie de acciones a partir de un conjunto de compuestos para los que se ha medido experimentalmente una propiedad/actividad. Es decir, que es un proceso que involucra varias etapas de desarrollo, las cuales serán detalladas en esta sección.

Existen algunos principios que deben cumplir los modelos QSPR (OECD, 2007): deben tener una propiedad o actividad a estudiar; un algoritmo determinado para tal fin; un dominio de aplicación definido; medidas apropiadas de bondad de ajuste, de robustez y de predicción; y en el caso de ser posible una interpretación mecanicista. En contrapartida, dentro de los errores más comunes en el modelado QSPR podemos encontrar: usar descriptores no informativos para el armado del modelo; contar con un proceso de selección de descriptores que no es el adecuado; modelar relaciones estructura-propiedad que no son lineales con modelos lineales; validar de manera incorrecta los modelos y no definir el dominio de aplicación de los mismos.

En la Figura 3.2 se grafica de manera esquemática los pasos pasa llevar adelante un proceso de modelado QSPR (Cherkasov et al., 2014). El esquema comienza con las estructuras moleculares, es decir, el conjunto de compuestos que será punto de partida para todo el desarrollo. Como paso siguiente, sobre la base de datos de compuestos, se calculan miles de descriptores moleculares, generalmente utilizando alguna herramienta específica para este fin (DRAGON, 2007; Yap, 2011). Estos descriptores, son valores numéricos que representan distintas características de las moléculas (por ejemplo, el peso molecular o número de carbonos). Luego, este conjunto de descriptores debe ser reducido, con el fin de quedarnos sólo con aquellos que son más representativos para la propiedad y eliminar aquellos que aportan información redundante o irrelevante para la misma. A partir de este conjunto reducido, se procede a armar el modelo predictivo utilizando distintas técnicas de aprendizaje automático (presentadas

en el capítulo 2). Estos modelos deben cumplir con ciertos niveles de rendimiento para poder ser tenidos en cuenta. En esta instancia es donde entra el juego la utilización de distintas técnicas para evaluar la confiabilidad del modelo.

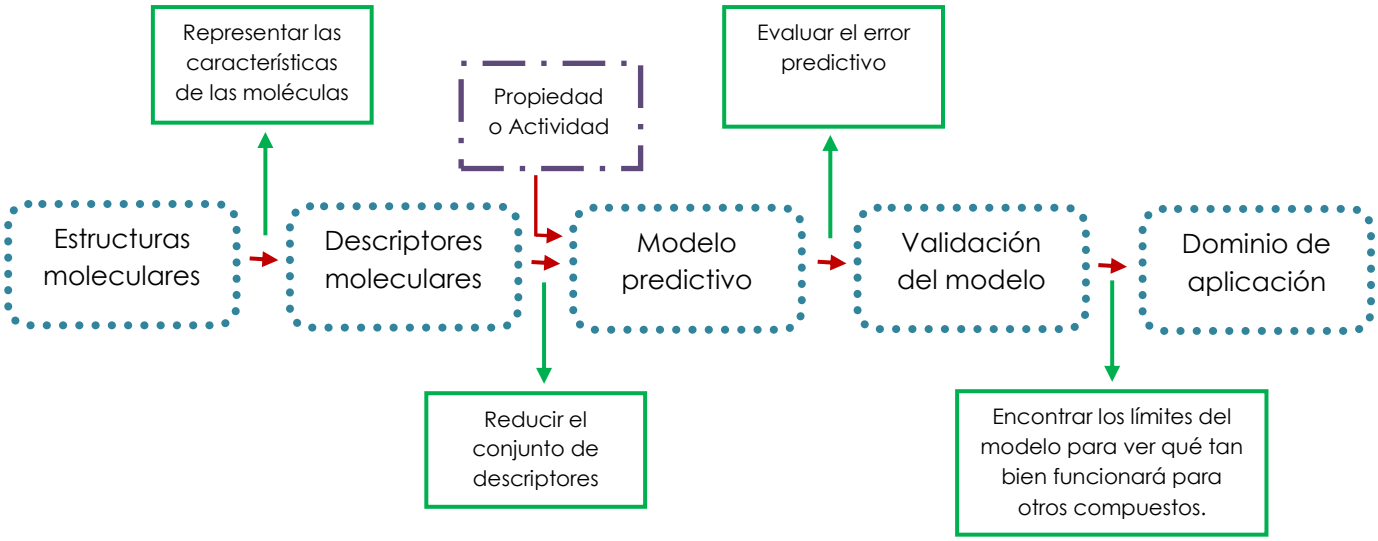


Figura 3.2 Esquema general de los pasos a seguir para llevar adelante un modelado QSPR.

Por último, una vez definido el modelo, se deben encontrar los límites de ese modelo para saber qué tan bueno será su rendimiento con compuestos nuevos, es decir se debe definir su dominio de aplicación.

A continuación, en las siguientes secciones, se explicará con más detalle cada uno de estos pasos.

3.2 Descriptores Moleculares

Los descriptores moleculares son valores numéricos que describen algún aspecto estructural de la molécula. Estos descriptores pueden ser: (i) mediciones experimentales, tales como log P, la polarizabilidad o el momento del dipolo, o (ii)

índices teóricos, derivados de la representación simbólica de la molécula a través de fórmulas matemáticas o algoritmos computacionales. La diferencia fundamental entre los descriptores teóricos y los medidos experimentalmente es que los descriptores teóricos no contienen error estadístico debido al ruido experimental, en comparación con las mediciones experimentales.

Los descriptores son de gran importancia para el modelado QSPR, ya que son los que aportan información relacionada con distintas características de los compuestos, que luego será utilizada para modelar las propiedades de interés. En las últimas décadas, se ha invertido mucho tiempo de investigación en establecer cómo capturar y convertir, siguiendo una ruta teórica, la información que se encuentra codificada en una estructura molecular en uno o más números usados para establecer los modelos QSPR (Todeschini y Consonni, 2008). Una definición más formal de un descriptor molecular es la siguiente: "*el descriptor molecular es el resultado final de un procedimiento lógico y matemático que transforma la información química codificada dentro de la representación simbólica de una molécula, en un número útil o el resultado de algún experimento estandarizado*"(Todeschini y Consonni, 2008).

Existe un amplio rango de descriptores moleculares, con diversas taxonomías para su agrupamiento. Existen distintas herramientas de software que permiten calcular hasta alrededor de 5000 descriptores derivados de distintas teorías y enfoques (DRAGON, 2007; Yap, 2011). El enfoque más común para clasificar los descriptores moleculares teóricos es de acuerdo a los diferentes tipos de representación molecular, es decir, por su dimensionalidad:

Cerodimensionales (0D): la representación molecular más simple es la fórmula química, que es la lista de los diferentes átomos, cada uno acompañado de un subíndice representando el número de ocurrencias de los átomos en las moléculas. Esta representación es independiente de cualquier conocimiento acerca de la estructura de la molécula.

Unidimensionales (1D): los descriptores derivados de este tipo de representación son vectores holográficos o cadenas de bits, y son calculados a partir de una representación de listas de fragmentos estructurales de una molécula. Estas listas son tan simples como una lista parcial de fragmentos, grupos funcionales o sustituyentes de interés presentes en la molécula. Por lo tanto, el cálculo de este tipo de descriptores no requiere de un conocimiento completo de la estructura molecular.

Los descriptores 0D y 1D siempre se pueden calcular fácilmente, se interpretan naturalmente, no requieren optimización de la estructura molecular y son independientes de cualquier problema conformacional. En contrapartida, generalmente muestran una degeneración muy alta, es decir, muchas moléculas tienen los mismos valores, por ejemplo, isómeros. El contenido de información que tienen es bajo, pero sin embargo pueden desempeñar un papel importante en el modelado de varias propiedades físico-químicas o pueden incluirse en modelos más complejos.

Bidimensionales (2D): son los descriptores moleculares derivados de la representación topológica de una molécula e incluyen a los llamados índices topológicos. La representación de una molécula en términos del grafo molecular se conoce comúnmente como la representación topológica y define la conectividad de los átomos del compuesto en términos de la presencia, y en última instancia, de la naturaleza de los enlaces químicos.

Tridimensionales (3D): la representación molecular tridimensional considera a una molécula como un objeto geométrico rígido en el espacio y permite una representación no sólo de la naturaleza y conectividad de los átomos, sino también de la configuración espacial global de los átomos de las moléculas. Esta representación de una molécula se denomina representación geométrica y define un compuesto en términos de los tipos de átomos la constituyen y el conjunto de coordenadas (x, y, z) asociadas a cada átomo. Los descriptores moleculares derivados de este tipo de representación son llamados descriptores

3D y muchos de ellos fueron propuestos para medir las propiedades estéricas y de tamaño de las moléculas.

Cuatridimensionales (4D): los descriptores en esta categoría, no sólo representan las características 3D de una molécula, sino que también pueden representar la disposición de los campos de fuerzas responsables de las distintas interacciones estéricas, hidrofóbicas y electroestáticas.

Actualmente existen una gran cantidad de herramientas informáticas para el cálculo de descriptores teóricos. Algunas de los más utilizados son CODESSA (2005), DRAGON (2007), PaDEL (Yap, 2011), entre otras. La más popular y la utilizada durante el trabajo de esta tesis es DRAGON. En el sitio web de DRAGON[1] se puede acceder a una lista con todos los descriptores que permite calcular.

3.3 Selección de Descriptores Moleculares

La selección del conjunto de descriptores más relevantes es un paso muy importante en el modelado, ya que se utiliza para seleccionar los descriptores más relevantes y representativos para la propiedad en estudio, dentro de un universo de miles de descriptores. El proceso llevado a cabo por los métodos de selección de características, consiste en seleccionar y evaluar distintos subconjuntos de descriptores con el fin de identificar un grupo reducido que se encuentre correlacionado con la propiedad objetivo. Muchos de estos enfoques requieren un gran esfuerzo computacional, debido a que se deben evaluar distintas combinaciones de descriptores moleculares.

Podemos enumerar algunas razones de la importancia de la selección de descriptores (Danishuddin y Khan, 2016): (i) el uso de pocos descriptores incrementa la comprensión de los modelos, (ii) puede reducir el riesgo de sobre-ajuste, eliminando descriptores que no aportan información o son redundantes,

[1] http://www.talete.mi.it/products/dragon_description.htm

(iii) permite entrenar modelos más rápidamente. Existen distintos enfoques y metodologías para la selección de descriptores que se explicarán con más en detalle en el capítulo 4.

3.4 Aprendizaje de Descriptores Moleculares

El aprendizaje de características (del término en inglés *feature learning*) hace referencia al proceso de inferir (o extraer) nuevas variables que caractericen un conjunto de datos en estudio. Una de las técnicas más populares es el *análisis de componentes principales* (*Principal Component Analysis*, PCA) (Wold et al., 1987) que es utilizada para extraer un conjunto reducido de características a partir de un conjunto de descriptores más grande.

Por otro lado, existen otras técnicas más específica para el modelado QSAR que permiten extraer un número reducido de nuevos descriptores calculados directamente de la estructura química de los compuestos (Dorronsoro et al., 2004). Una vez que los descriptores son extraídos, directamente pueden ser utilizados para inferir los modelos QSPR. Un punto débil de este tipo de técnicas es la interpretabilidad de los modelos, ya que generalmente los descriptores extraídos son difíciles de entender en términos químicos. En la sección 4.2 se presentará más en detalle esta metodología específica de aprendizaje de descriptores utilizada durante este trabajo, que permite extraer unos pocos descriptores directamente de la estructura de la molécula.

3.5 Dominio de Aplicación

La definición del dominio de aplicación es un punto fundamental en el modelado QSPR ya que incrementa la confianza en las predicciones y permite hacer un uso más práctico de los modelos. En la literatura referida a QSAR/QSPR no siempre es evidente si se ha aplicado (o en qué medida) el concepto de dominio de aplicación (Netzeva et al., 2005). En algunos casos se encuentra implícito, por

ejemplo, cuando se desarrolla un modelo a partir de un conjunto de datos perteneciente a una sola clase química. En otros casos, se define explícitamente y el enfoque más comúnmente adoptado es definir el dominio de aplicación del modelo utilizando reglas estructurales, es decir dividiendo los compuestos en familias químicas y determinando cuáles de esas familias no están siendo bien modeladas (Schultz et al., 2002). Si se usan descriptores continuos, es posible definirlo en términos de la cobertura del conjunto de entrenamiento en el espacio descriptor del modelo (Jaworska et al., 2005). Estas aproximaciones se basan en métodos estadísticos, en donde las estimaciones interpoladas se consideran más fiables que las extrapoladas. Otros enfoques se basan en el análisis de regresión lineal múltiple (MLR) en combinación con enfoques basados en distancias (Gramatica et al., 2003) o análisis de árboles de decisión (Tong et al., 2004). Muchos enfoques para definir el dominio de aplicación se basan en el análisis de similaridad (Nikolova y Jaworska, 2003). Todos estos enfoques se apoyan en la premisa que una predicción es confiable para un compuesto nuevo, si este compuesto es *similar* (estructuralmente) a los que componen el conjunto de entrenamiento. Si contamos con un conjunto nuevo de compuestos que son muy diferentes del conjunto de compuestos que se utilizaron para entrenar el modelo, por más que el modelo sea robusto y esté validado, no podemos asegurar que las predicciones realizadas vayan a ser confiables. Esto se debe a que no se puede predecir confiablemente la propiedad modelada para todo el universo de compuestos químicos. Por el contrario, se deben encontrar los límites del modelo, sus alcances y limitaciones, es decir, su dominio de aplicación. Mientras que el concepto es fácil de entender (véase Figura 3.3), en algunos casos resulta relativamente difícil (a veces imposible) definir un dominio de aplicación. Este tipo de enfoques son los que serán abordados en este trabajo. En el capítulo 6 se abordarán estos conceptos en más profundidad.

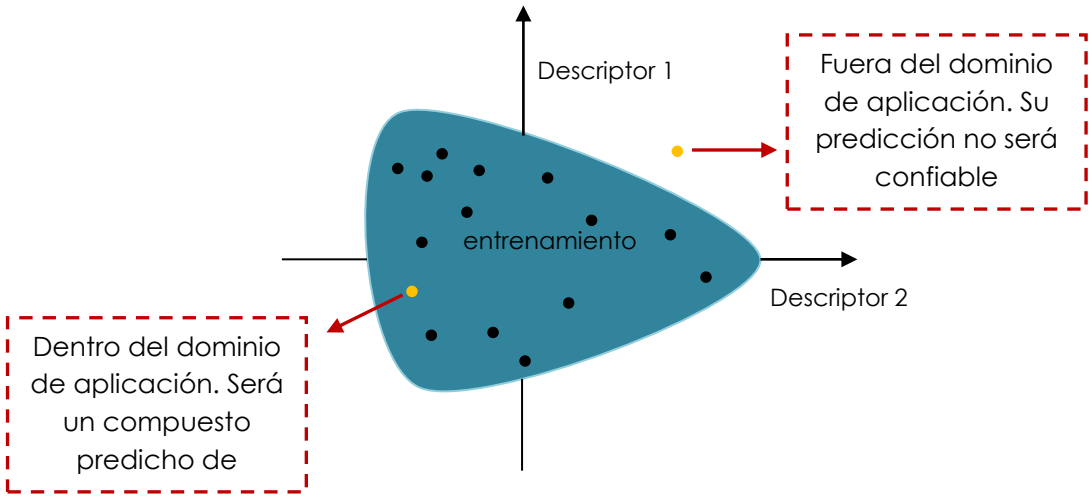


Figura 3.3 Descripción esquemática del dominio de aplicación

3.6 Campos de Aplicación del Modelado QSAR/QSPR

Existe un gran número de aplicaciones de modelos QSAR/QSPR en distintos tipos de industria y organizaciones gubernamentales. Estos modelos tienen un amplio uso en cuestiones relacionadas con: (i) la predicción de diferentes propiedades físico-químicas de las moléculas; (ii) el diseño racional de numerosos productos tales como fármacos, pesticidas, productos de química fina, entre otros; (iii) la identificación de compuestos peligrosos en etapas tempranas del desarrollo de un producto; o (iv) la predicción de la toxicidad de distintas sustancias para los seres humanos a través de distintos tipos de exposición; entre muchos otros.

A continuación, explicaremos con más detalle tres campos de aplicación donde se ha aplicado favorablemente el modelado QSPR: diseño racional de fármacos, las ciencias de los materiales y el medioambiente. Para cada una de estas áreas

se han estudiado diferentes propiedades/actividades durante el trabajo de esta tesis.

3.6.1 Diseño Racional de Fármacos

Los modelos QSAR/QSPR tienen un lugar fundamental en el diseño de fármacos, debido a que permiten realizar una evaluación *in silico* de propiedades relacionadas a la actividad, selectividad y toxicidad de compuestos candidatos, y de esta manera contribuyen a reducir tiempos y costos de desarrollo. El diseño racional de fármacos comienza en primera instancia con el conocimiento de las respuestas químicas específicas de una determinada propiedad/actividad biológica en el organismo.

En este trabajo, las propiedades que se han estudiado vinculadas al diseño de fármacos fueron la *absorción intestinal humana* (*Human Intestinal Absorption, HIA*) y el *pasaje de la barrera hemato-encefálica* (*Blood-Brain Barrier, BBB*). Ambas relacionadas al comportamiento ADMET (absorción, distribución, metabolismo, excreción y toxicidad) y esenciales para el desarrollo de nuevos fármacos. Para estas se han desarrollado diferentes modelos que permiten predecir el comportamiento de diferentes compuestos frente a estas propiedades.

3.6.2 Diseño de Materiales

Dentro de la ciencia de los materiales, los plásticos, o polímeros sintéticos, ocupan un lugar de preponderancia y su uso se ha incrementado considerablemente en los últimos 30 años, estando presentes en múltiples áreas de aplicación como medicina, materiales de construcción; diversas industrias tales como aeronáutica, petrolera, automotriz, etc. (Utracki y Wilkie, 2002). Ante un mercado cada vez más demandante, se trabaja en el desarrollo de nuevos polímeros sintéticos que satisfagan necesidades específicas según el campo de aplicación. La

metodología clásica consistía en sintetizar primero el material y luego testear sus propiedades, es decir se llegaba un nuevo material por prueba y error. La propuesta actual consiste en aplicar técnicas in silico, más precisamente el modelado QSPR, que se ha vuelto una alternativa de mucho valor para las primeras etapas de desarrollo (diseño molecular), ya que permite establecer relaciones entre las características estructurales de los polímeros y una determinada propiedad fisicoquímica sin necesidad de sintetizar muestras.

Dentro de las propiedades de los polímeros, las mecánicas son las que se han estudiado en este trabajo. Estas propiedades determinan la respuesta de un material cuando se lo somete a la acción de fuerzas mecánicas externas. Para los polímeros, las propiedades mecánicas presentan características particulares: dependen fuertemente de la temperatura y de la escala de tiempo, pueden desarrollar grandes niveles de deformaciones reversibles y/o irreversibles y son afectadas notablemente por cambios físicos e interacciones químicas con otros materiales. Más específicamente, las propiedades mecánicas sufren cambios profundos en el rango de temperaturas donde se manifiesta el cambio del material entre un estado rígido y uno frágil (Katritzky et al., 1998), condicionando esto el proceso de fabricación y el perfil de aplicación del nuevo material.

Existen numerosas propiedades mecánicas que definen el perfil de aplicabilidad de un polímero. En particular en este trabajo de tesis, se han estudiado propiedades mecánicas derivadas del ensayo de tensión en una dimensión, que brindan información relacionada a la ductilidad, resistencia y rigidez de un material polimérico, siendo respectivamente: elongación a la rotura (*Elongation at Break*), resistencia a la rotura (*Strength at Break*) y módulo elástico o de Young (*Tensile Modulus*).

3.6.3 Medioambiente

Los modelos QSPR constituyen una valiosa herramienta para la evaluación preliminar del impacto de distintos contaminantes sobre el medioambiente y la

salud humana. Los modelos generados pueden emplearse tanto en la evaluación de riesgos como en la toma de decisiones en políticas de salud pública. En esta tesis, se estudió particularmente el coeficiente de partición sangre-hígado (log P_{liver}) para compuestos orgánicos volátiles (VOCs, del inglés *Volatile Organic Compounds*). Los VOCs son gases que se emiten de ciertos sólidos o líquidos, y son ampliamente utilizados como ingredientes en productos para el hogar (pinturas, barnices, productos de limpieza, desinfección, cosmética, entre otros) y muchos de ellos son tóxicos según el nivel de exposición. En este contexto, el estudio del coeficiente de partición sangre-hígado es importante para evaluar el riesgo que puede tener inhalar VOCs, así como también determinar de qué manera se distribuyen en el organismo. Un análisis bien detallado de este estudio puede encontrarse los capítulos 4 y 5.

3.7 Sumario

En este capítulo se desarrollaron diferentes conceptos de importancia para el proceso de modelado QSAR/QSPR. Se explicaron los pasos para desarrollar un modelo, haciendo énfasis en temas importantes como la selección y aprendizaje de descriptores moleculares, además de explicar la relevancia de poder determinar el dominio de aplicación de un modelo. Por último, se realizó una introducción a los campos de aplicación de estos modelos junto con una explicación de las propiedades que se han estudiado en el marco de esta tesis.

Capítulo 4: Metodologías para la Selección y Aprendizaje de Descriptores en el Modelado QSAR/QSPR

En este capítulo se presentarán distintos enfoques para llevar a cabo una de las etapas fundamentales en el modelado QSPR: la selección de descriptores. Así mismo, se abordarán conceptos y un enfoque para el aprendizaje de descriptores. Se presentarán dos casos de estudio, uno focalizado en el proceso de selección de descriptores y otro como un primer acercamiento en la utilización de una técnica para el aprendizaje de características. En particular, el caso de estudio presentado en la sección 4.3.1 se trata de un método semi-automático para la selección de descriptores, en el que se trabajó con expertos en el dominio y que sirvió de inspiración para desarrollar una herramienta que pueda brindar soporte en esa tarea desde el punto de vista computacional. Esta herramienta será presentada en el capítulo siguiente.

4.1 Selección de Descriptores Moleculares

El diseño de modelos QSPR requiere tratar con varias cuestiones como hemos explicado anteriormente. Una de ellas es la selección del conjunto más relevante de descriptores moleculares para la propiedad o actividad que va a ser modelada. Las estructuras químicas son generalmente codificadas por una variedad de familias de descriptores tales como grupos funcionales, topológicos, constitucionales, termodinámicos, etc. Varios de ellos pueden contribuir con información redundante o pueden ser irrelevantes para la propiedad o actividad biológica en estudio y, por lo tanto, afectar el descubrimiento de la relación descriptor - propiedad/actividad. Por esta razón, la selección de los descriptores

más relevantes es considerada como una de las tareas más difíciles y cruciales para el modelado QSPR (Goodarzi et al. 2012; Palczewska et al. 2013).

4.1.1 Selección manual de descriptores

Hansch (Hansch, 1969) planteó el enfoque en el que se representa a una estructura molecular con sólo unos poco descriptores (normalmente *log P*, las constantes de Hammett[2], HOMO/LUMO[3], y algunos parámetros estéricos) que son seleccionados manualmente por el experto e insertados en la ecuación QSPR para modelar el valor de una propiedad objetivo. En este enfoque, Hansch plantea que la selección de descriptores es guiada por la convicción del experto al tener un conocimiento previo del mecanismo de la propiedad estudiada. Además, esta selección es conducida también por la premisa de atribuirle un significado *mecanicista* a cualquier descriptor molecular seleccionado a partir de un grupo reducido de variables potenciales para el modelado. Por ejemplo, el caso de HOMO/LUMO que siempre se selecciona para el modelado de la reactividad química, o *log* P que es utilizado en una gran cantidad de modelos de toxicidad. Este enfoque se ha aplicado recientemente en distintos artículos (Ali y Ali, 2015; Roberts y Costello, 2003).

4.1.2 Selección automática de descriptores

Existen tres familias típicas de métodos para abordar el problema de selección de descriptores: *filtros*, métodos *de envoltura* (*wrappers*) y métodos *embebidos* (Guyon y Elisseeff, 2003). Los *filtros* permiten establecer alguna medida del aporte individual de cada descriptor con respecto a la propiedad objetivo de manera independiente de cualquier predictor. Tradicionalmente, estos métodos de *filtros* se utilizaban en primera instancia, una vez calculados los descriptores para reducir

[2] Constantes definidas por Louis Hammett para realizar comparaciones cuantitativas entre sustituyentes.

[3] Son tipos de órbitas moleculares. HOMO (del inglés *Highest Occupied Molecular Orbital*) y LUMO (del inglés *Lowest Unoccupied Molecular Orbital)*.

la dimensionalidad del conjunto de datos y luego aplicar algún método más sofisticado. Se trata de técnicas que tienen un costo computacional bajo, ya que utilizan propiedades estadísticas con el fin de "filtrar" aquellos descriptores que resultan poco informativos, mirando sólo propiedades intrínsecas de los datos. Un enfoque que se puede aplicar es mantener el descriptor que tiene mayor correlación con la variable objetivo y eliminar el otro del conjunto de datos. Otra alternativa es eliminar el descriptor con la menor varianza y menor correlación con la variable objetivo y mantener aquellos con correlación más alta. Existen muchos métodos que se utilizan para filtrar datos, tales como ganancia de información, pruebas de chi-cuadrado, distancia euclídea, entre muchos otros (Duch, 2006).

Los métodos *de envoltura* suelen dar resultados más generalizables e implican la optimización de un predictor como parte del proceso de selección. El funcionamiento de este tipo de métodos *de envoltura* se realiza de manera interna en dos partes diferenciadas: una parte que se encarga de encontrar los posibles subconjuntos de descriptores y otra que se encarga de evaluar la relevancia de esos subconjuntos (Soto et al., 2009). Dentro de esta categoría pueden ser clasificados distintos métodos. Tres métodos muy populares son: (i) selección hacia adelante, donde se selecciona el descriptor que más relación tiene con la variable objetivo y luego se van agregando al modelo descriptores, de a uno a la vez. El proceso termina cuando la última variable que entra al modelo aporta una mejora en la predicción que es insignificante o cuando ya se incluyeron todas las variables; (ii) eliminación hacia atrás, comienza con todos los descriptores en el modelo y se van eliminando uno a la vez, eliminando el que tiene menor significancia, según una medida específica elegida para la evaluación. El proceso termina cuando todos los descriptores que quedan son significativos o se llega a un único descriptor;. (iii) regresión paso a paso, que utiliza los dos métodos anteriores, es decir que un descriptor que entró en el modelo en una etapa anterior, puede ser eliminado en una etapa posterior (Kohavi y John, 1997).

Por otro lado, otra técnica ampliamente utilizada para selección de descriptores son los *algoritmos genéticos*. Estos algoritmos se basan en el principio de la evolución biológica tales como, reproducción, mutación, recombinación, selección natural del más apto. Los descriptores toman el rol de los genes, y un conjunto de descriptores es llamado *cromosoma*. Cada objeto individual de una población es descrito por un cromosoma de valores binarios, donde el valor uno indica que ha sido seleccionado y el cero que no ha sido seleccionado. Los algoritmos genéticos se caracterizan por realizar búsquedas globales eficaces de manera rápida en grandes espacios de búsqueda (Yang y Honavar, 1998).

Por último, los métodos *integrados*, incorporan la selección de descriptores como parte del proceso de entrenamiento. Un ejemplo son los árboles de decisión, donde el mecanismo que selecciona los nodos internos que generan el mejor particionamiento de los datos intrínsecamente define un mejor conjunto de descriptores para el modelo.

4.1.2.1 Delphos

En este apartado haremos mención especial a una herramienta para la selección de descriptores que ha sido utilizada para distintos experimentos en el desarrollo de esta tesis.

Esta herramienta implementa un método de optimización multi-objetivo, basado en dos fases, con el fin de identificar subconjuntos de descriptores relevantes a la propiedad en estudio (Soto, Cecchini, et al., 2009). En la primera fase, se utiliza un método de envoltorio para realizar una exploración dentro del espacio de búsqueda y encontrar subconjuntos apropiados de descriptores, utilizando algoritmos evolutivos y diferentes métodos de aprendizaje automático. En la segunda fase, la selección final se ejecuta utilizando las selecciones de la primera fase y métricas de predicción más precisas. Como resultado, DELPHOS presenta múltiples conjuntos de descriptores que se correlacionan bien con la propiedad a predecir, teniendo en cuenta distintas métricas como el mínimo error absoluto medio (*MAE*) y error cuadrático medio (*MSE*).

4.1.3 Selección automática e interpretabilidad de los descriptores

A pesar de los numerosos enfoques estadísticos para llevar adelante el proceso de selección de descriptores, los usuarios expertos que desean extraer los descriptores más informativos para entender o predecir una propiedad/actividad todavía se enfrentan a un reto importante. Esto es debido a que ninguno de estos métodos puede ser declarado como el mejor enfoque para cualquier combinación posible de conjunto de datos y método de predicción. Por lo tanto, en un entorno real, los usuarios no tienen una forma certera de saber qué método de selección de descriptores funcionaría mejor.

Otro de los puntos difíciles de lograr, es que los descriptores que se seleccionen de manera automática, expliquen de una manera apropiada a la propiedad que se quiere predecir. Esto es, definir qué tan interpretables son, en términos físico-químicos, esos descriptores seleccionados, para lo cual es fundamental la intervención del experto en el dominio.

Por último, otra crítica común que reciben los métodos automáticos de selección descriptores, es que la mayoría de estos enfoques se consideran como "cajas negras" para los químicos. Esto se debe al hecho de que con el fin de mejorar los resultados o introducir el conocimiento de dominio en los criterios de selección, es necesario conocer el funcionamiento interno del método (Hewitt, Ellison, Enoch, Madden, y Cronin, 2010).

4.1.4 Enfoque híbrido de selección de descriptores

Como se ha mencionado anteriormente una de las características deseables de los modelos QSPR es su interpretabilidad, es decir, poder determinar qué aporte

en términos físico-químicos está teniendo cada descriptor del modelo en relación a la propiedad/actividad que queremos predecir. En otras palabras, se trata de ir un poco más allá de lo meramente estadístico, para encontrarle un significado a esos descriptores dentro del modelo. Para lograr eso, es indispensable la intervención del experto en el dominio durante el proceso de selección de descriptores, que seleccionará manualmente los descriptores que considere, una vez que la selección automática haya sido realizada. De esto se trata el enfoque híbrido para la selección de descriptores que presentaremos aquí.

En esta sección abordaremos uno de los trabajos realizados, en el cuál estudiamos una propiedad asociada a las ciencias ambientales: el l*ogaritmo del coeficiente de partición sangre-hígado* (*log* P_{liver}). Se desarrollaron distintos modelos de regresión para esta propiedad aplicados sobre compuestos orgánicos volátiles (VOCs). El objetivo aquí fue lograr obtener modelos que sean de buena calidad predictiva, de baja cardinalidad e interpretables en términos físico-químicos. Para lograr esto se propuso una metodología híbrida que combina un método de aprendizaje automático con una selección manual de descriptores basada en el conocimiento experto.

4.2 Aprendizaje de Descriptores Moleculares

En el capítulo anterior se presentó el concepto de aprendizaje de descriptores que hace referencia al proceso de inferir (o extraer) nuevas variables que caractericen a una propiedad en estudio. Existen muchas técnicas, desde la más antiguamente utilizada y conocida PCA (Wold et al., 1987) para transformaciones lineales, hasta por ejemplo los auto-encoders, un tipo de red neuronal para realizar transformaciones más sofisticadas (Hinton y Zemel, 1994).

4.2.1 CODES-TSAR

En este trabajo de tesis se incursionó en una técnica para el aprendizaje de descriptores que está basada en la combinación de los métodos CODES y TSAR (Dorronsoro et al., 2004), diseñados específicamente para el modelado QSPR. Esta estrategia extrae un conjunto reducido de nuevos descriptores, utilizando redes neuronales para el procedimiento de aprendizaje. Estos nuevos descriptores constituyen un nuevo espacio de variables, que tiene una baja dimensionalidad, donde las variables representan información derivada de toda la estructura molecular de los compuestos. Esta técnica ha sido utilizada para inferir modelos QSPR aplicados en el de diseño de fármacos (Castaño et al., 2008; Guerra, Campillo, y Páez, 2010; Guerra, Páez, y Campillo, 2008).

Para utilizar CODES, se debe describir todas las moléculas de la base de datos usando la notación SMILES (*Simplified Molecular Input Line System*),la cual se basa en la teoría de grafos y consiste en una serie de caracteres que no contienen espacios entre ellos (Weininger, 1988). Una vez realizado esto, CODES procesa uno a uno los SMILES y devuelve como resultado una matriz dinámica por cada molécula. Cada una de estas matrices dinámicas tiene una dimensión $n\,x\,m$ que depende de cada molécula, ya que n es el número de átomos de cada estructura y m es el número de iteraciones necesarias para lograr la convergencia en el proceso de entrenamiento. Cada matriz dinámica generada representa un conjunto de descriptores.

Una vez que se tienen los descriptores calculados, se utiliza la metodología TSAR con el fin de obtener un conjunto reducido. Más específicamente, esta metodología procesa cada matriz dinámica (descriptores) para obtener un número razonablemente pequeño de descriptores para cada molécula. Este método, que está basado en el trabajo de (Livingstone et al., 1991), utiliza un algoritmo auto-codificador (*auto-encoder*), con la premisa que si el patrón de

entrada se reproduce bien en la capa de salida de la red, la capa oculta debería representar la misma información, aunque esta capa tenga un número menor de nodos que las capas de entrada y salida. En TSAR, el número de nodos de la capa oculta determina el número de descriptores generados luego del proceso de reducción.

En esta instancia, es válido plantear un paralelismo con la selección de descriptores en el sentido del uso de las herramientas para cada caso, durante el trabajo de tesis. Es decir, el paquete CODES-TSAR para el aprendizaje sería equivalente en términos conceptuales a utilizar DRAGON-DELPHOS para el caso de selección. En el marco de esta tesis, CODES y DRAGON se utilizan para el cálculo de descriptores mientras que DELPHOS y TSAR se utilizan para la selección y aprendizaje de descriptores, respectivamente. A continuación, se presentarán dos escenarios en los que se detallarán el uso de estas metodologías.

4.3 Casos de estudio

En esta sección se detallarán dos casos de estudio. En el primero que presentaremos, se generan modelos para predecir una propiedad con la intervención del experto en el dominio en el proceso de selección de descriptores (Palomba et al., 2012). Este trabajo sirvió de inspiración para luego desarrollar una herramienta que permita brindar soporte al exporto en el proceso de la selección de los descriptores más relevantes para la propiedad. En el segundo caso, se presenta un primer acercamiento en la utilización de la técnica de aprendizaje de descriptores presentada anteriormente.

4.3.1 Selección de descriptores para la propiedad log P_{liver}

Los compuestos orgánicos volátiles (VOCs, del inglés *Volatile Organic Compounds)* son gases que se emiten desde ciertos sólidos o líquidos e incluyen

una variedad de productos químicos, algunos de los cuáles pueden tener efectos adversos para la salud a corto y largo plazo. La concentración de muchos VOCs es considerablemente más alta en interiores (hasta diez veces más) que en exteriores. Los compuestos orgánicos son ampliamente utilizados como ingredientes en productos para el hogar. Las pinturas, los barnices y las ceras contienen todos los disolventes orgánicos, así como también muchos productos de limpieza, desinfección, cosmética, desengrasado y entre otros. Todos estos productos pueden liberar compuestos orgánicos mientras que se utilizan y, en cierto grado, cuando se almacenan. La principal preocupación es la posibilidad de que los VOCs tengan efectos nocivos para la salud de las personas que se exponen a ellos (Vallero, 2007; Williams y Ralf, 2007; Woodruff, Burke, y Zeise, 2011). En Woodruff et al. (2011) se describe la necesidad de mejores políticas de salud pública sobre sustancias químicas liberadas en el medio ambiente. Ellos propusieron la modernización de enfoques para evaluar los riesgos de salud y señalaron la importancia de la comprensión científica de la relación entre la exposición a contaminantes y los efectos adversos para la salud.

En este contexto, los modelos QSPR constituyen una valiosa herramienta para la predicción *in silico* de propiedades vinculadas a la distribución en el organismo de los VOCs que se inhalan. Esta metodología se ha aplicado en estudios de inhalación de VOCs (Dashtbozorgi y Golmohammadi, 2010; Katritzky et al., 2005) y se relaciona con el análisis de los modelos fármaco-cinéticos basados en la fisiología (PBPK, del inglés *Physiologically Based Pharmacokinetic*). El modelado PBPK es una técnica de modelado matemático para la predicción de la absorción, distribución, metabolismo y excreción (ADME) de sustancias químicas en los seres humanos y otras especies animales (Buist et al., 2012). El desarrollo de modelos PBPK es una tarea compleja debido a que requiere de la definición de sub-modelos basados en una gran cantidad de datos específicos (Sager et al., 2015; Vork et al., 2013).

Para el estudio de los VOCs, es importante considerar la ruta de inhalación y, en consecuencia, se han desarrollado muchos modelos respiratorios PBPK durante la

última década. En este contexto, distintos coeficientes tales como partición sangre-aire, aire-hígado e hígado-sangre de los VOCs son importantes para la evaluación del riesgo y estimación de la biodisponibilidad (Abraham, Ibrahim, y Acree, 2007). Se han realizado varios intentos para modelar la relación entre la estructura química (o ciertas propiedades moleculares) y la distribución sangre-hígado, denotada habitualmente como *log* P_{liver}, de los VOCs y drogas. En Abraham and Weathersby (1994) utilizaron los descriptores de Abraham para estimar los valores de *log* P_{liver} de los VOCs. En Balaz and Lukacova (1999) correlacionaron valores de *log* P_{liver} para 28 compuestos utilizando cuatro variables. Poulin and Theil (Poulin y Theil, 2000) desarrollaron una ecuación para la predicción de los coeficientes de partición plasma-tejidos *in vivo* de fármacos. Zhang (2004) construyó un modelo no lineal para calcular *log* P_{liver} en VOCs. En Liu et al. (2005) se obtuvieron un modelo no lineal para predecir la partición de tejido-sangre de compuestos orgánicos utilizando máquinas de soporte vectorial de cuadrados mínimos. En Rodgers et al. (2005) se obtuvieron ecuaciones para la predicción de la distribución agua-plasma-tejido. Zhang and Zhang (2006) generaron un modelo de entrenamiento general para predecir la distribución in vivo sangre-hígado (entre otros tejidos) de medicamentos. En Abraham et al. (2007) aplicaron ecuaciones de solvatación para correlacionar los coeficientes de partición in vitro de sangre con el hígado de VOCs y las drogas. En Martin-Biosca et al. (2009) emplearon cromatografía micelar de biopartición (CMB) para la predicción de coeficientes de partición sangre-tejidos de medicamentos y propusieron modelos de regresión lineal múltiple y PLS2 basados en los datos de retención de CMB.

Mientras que la mayoría de estos estudios hacen interesantes aportes al estudio de la propiedad log P_{liver}, en general, la precisión de predicción o interpretación en términos físico químicos de los modelos presentados, no son lo suficientemente buenas para el uso generalizado en una escala industrial. En particular, y como hemos mencionado anteriormente en este capítulo, la cuestión clave para las metodologías QSPR basadas en datos es cómo el conocimiento del experto

puede ser incorporado en el proceso de modelado con el fin de obtener predictores interpretables. Teniendo en cuenta estas premisas, en el trabajo desarrollado se presentan modelos estadísticos QSPR para log P_{liver}. La metodología propuesta combina el uso de métodos de aprendizaje automático con el análisis de expertos para la identificación de los descriptores moleculares más relevantes para la definición del modelo QSPR. Esta integración se logra por medio de un cuidadoso análisis, en el que el reducido número de descriptores seleccionados por métodos automatizados son evaluados por los expertos en cuanto a su significado químico y su contribución a un modelo estadístico candidato QSPR. A partir de este análisis semi-automático se elige un nuevo conjunto de descriptores y finalmente se obtiene el modelo QSPR estadístico asociado. De esta manera, se pretende hacer una doble contribución con este trabajo. En primer lugar, el diseño de nuevos modelos para log P_{liver} que tengan una alta capacidad predictiva y una buena interpretabilidad. En segundo lugar, la aplicación de nuestra metodología específica de diseño que integra un método de aprendizaje automático con el conocimiento experto humano, y por lo tanto la recomendación de sus aplicaciones análogas para la predicción de otras propiedades químicas.

4.3.1.1 Metodología propuesta

Para la selección de los descriptores más relevantes, se empleó un esquema mixto en el cuál se combinó el proceso automático con el conocimiento del experto químico. Como primer paso, se aplicó un enfoque de aprendizaje automático basado en una validación cruzada (cross-fold) con una selección de características de *k*-grupos (*k*-fold) (Picard y Cook, 1984). Este enfoque consiste en dividir el conjunto de datos en k grupos. La selección de características utiliza un algoritmo de aprendizaje que se aplica para predecir cada grupo usando los datos de los *k*-1 grupos restantes. Dado que se pueden seleccionar k diferentes

conjuntos de características, se empleó un esquema de votación, donde los descriptores seleccionados con más frecuencia se mantienen para el conjunto final de descriptores relevantes. Esta técnica asegura que las predicciones particulares no estén sesgadas por la sobre-selección o sobre-ajuste, ya que cada predicción se realiza sin necesidad de usar las muestras de testeo durante la selección de características ni durante el proceso de construcción del clasificador. A partir de estos experimentos, los descriptores seleccionados con mayor frecuencia se mantuvieron para el conjunto inicial de descriptores relevantes.

Como segundo paso, se empleó el conocimiento químico con el fin de evaluar el mérito de cada descriptor seleccionado automáticamente. Dado que la mayoría de ellos no exhiben una explicación físico-química clara, un pequeño número de estos descriptores fueron elegidos para los modelos QSPR finales, mientras que otros pocos descriptores fueron incorporados en base a la experiencia química. En la Figura 4.1 se esquematiza nuestra metodología general y en los siguientes apartados se darán explicaciones detalladas de esos pasos.

4.3.1.2 Cálculo y selección de descriptores moleculares

Los valores de los coeficientes de partición sangre-hígado *in vitro*, *log* P_{liver} (humano/rata), fueron tomados de Abraham et al. (2007). El conjunto de datos tiene 122 VOCs entre los que se destacan los hidrocarburos, haluros de alquilo, alcoholes, éteres, ésteres, cetonas, epóxidos, nitrilos, halobencenos, hidrocarburos policíclicos y los derivados del benceno. El rango de valores de log P_{liver} va desde -0.56 a 1.17.

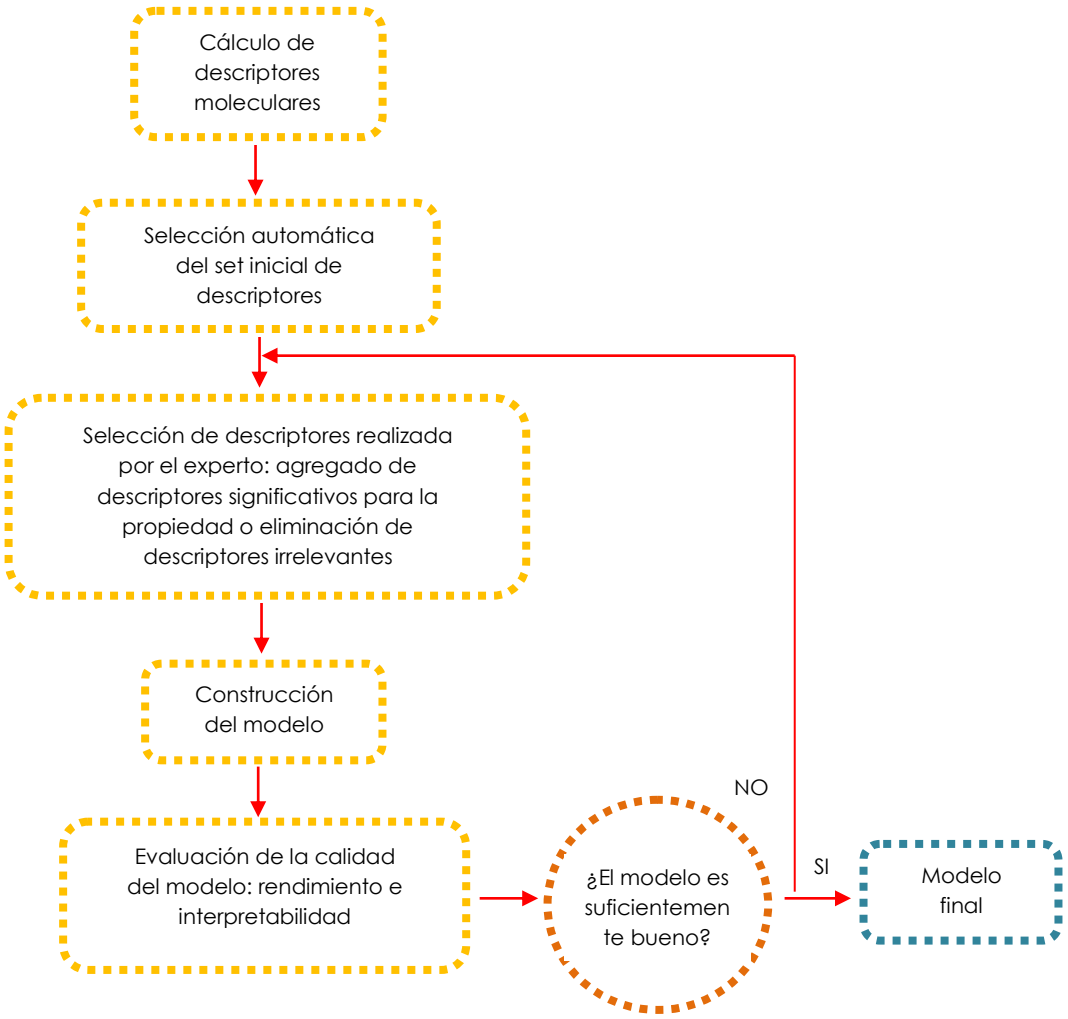


Figura 4.1 Esquema de la metodología combinada (semi-automática) propuesta para el desarrollo de un modelo QSPR.

Como hemos mencionado en otras secciones, un paso crítico en el desarrollo de modelos QSPR es el cálculo de los descriptores moleculares, ya que el rendimiento del modelo y los resultados dependen en gran medida en la forma en que se calculan los descriptores. En este caso, el proceso de cálculo de los descriptores moleculares se describe como sigue: todas las estructuras de VOCs se extrajeron utilizando HyperChem 8.0.7 (HyperChem, 2006). Las moléculas se optimizaron con el mismo software, a fin de encontrar conformaciones energéticamente estables.

Las estructuras fueron pre-optimizadas utilizando el campo de fuerza (MM+) de mecánica molecular. Luego, las geometrías resultantes se refinaron aún más utilizando el método semiempírico Orbital Molecular AM 1 (Modelo Austin 1) utilizando el algoritmo Polak-Ribiere y un valor límite de gradiente de 0.01 *kcal/(mol Å)*. Como paso siguiente, los archivos de salida de HyperChem fueron utilizados por Dragón (DRAGON, 2007) para calcular varias clases de descriptores tales como: constitucionales, geométricos, topológicos y electrostáticos. Por último, se han eliminado los descriptores constantes (es decir, las variables que tienen el mismo valor para todas las muestras en el conjunto de datos) y los que son casi-constantes (es decir, las variables que tienen el mismo valor, pero permiten que un pequeño número predeterminado de muestras tomen otros valores).

El conjunto final de descriptores fue elegido mediante el uso de un enfoque híbrido que combina la selección automática de características con el conocimiento del experto físico-químico. El método de selección de características que hemos utilizado aquí se basa en una validación cruzada de 5 grupos, con una selección de características *in-fold* sobre el conjunto de entrenamiento, que seleccionó los siguientes descriptores: RTU+, Mor29u, AMW, ZM2V, Jhetv, PW4, Ss, Ms, Me, Mv, NCIC, AAC, GATS2m, S1K, PW3, EEig07x, IC1, Qindex, RBN, Mor04m, Mor11v, ATS1v y MAXDN (los nombres completos de los descriptores se pueden encontrar en el sitio web de DRAGON[4]. Luego fue realizada la selección manual guiada físico-químicamente por expertos de dominio, la cual tenía por objeto incluir en el modelo aspectos ortogonales de las moléculas. De este modo, se busca considerar descriptores con características importantes e interpretables y que mantengan la redundancia mínima. Estos descriptores seleccionados manualmente son: AMW, Mor29u, ALOGP, Pol y Se. Una breve descripción de cada uno de ellos se incluye en la Tabla 4.1. Los dos primeros descriptores fueron tomados de los resultados de los algoritmos de selección de características, y los tres siguientes se añadieron sobre la base de

[4] http://www.talete.mi.it/products/dragon_description.htm

criterios de los expertos. Los fundamentos físico-químicos de esta selección son desarrollados en detalle más adelante en el texto.

Descriptor	Significado	Familia
AMW	Peso molecular promedio	Constitucional
Mor29u	Representación Molecular-3-D de estructuras basadas en difracción de electrones - no ponderado/señal 29	3D-MoRSE
ALOGP	Coeficiente de partición octanol-agua de Ghose-Crippen	Propiedad molecular
Pol	Número de polaridad	Topológico
Se	Suma de electronegatividades atómicas de Sanderson	Constitucional

Tabla 4.1 Conjunto final de descriptores

4.3.1.3 Diseño de la experimentación

Para los experimentos, separamos un conjunto de compuestos de testeo, que sólo es usado una vez para estimar el rendimiento imparcial de nuestro método de predicción. Hemos aplicado esta estrategia de validación en dos experimentos. En ambos casos los compuestos seleccionados para la prueba fueron escogidos mediante el uso de una selección estratificada para asegurar que los compuestos de los conjuntos de entrenamiento y testeo estén distribuidos de manera similar.

Se aplicaron dos metodologías como algoritmos de regresión: árboles de decisión y un ensamble de redes neuronales. El modelo de árbol de decisión utilizado fue M5P (Wang y Witten, 1996). Esta es una extensión del algoritmo M5 de Quinlan que permite el uso de árboles de decisión para los problemas de regresión, es decir, los atributos y variables de destino se puede definir de forma continua sobre el conjunto de los números reales. El aspecto clave de este algoritmo de árbol de decisión es que hace uso de un modelo de regresión lineal para cada hoja del árbol. También proporciona un mecanismo para la poda (es decir, manteniendo

la altura del mínimo árbol para evitar el sobreajuste) y el proceso de suavizado que permite compensar las discontinuidades entre los modelos lineales adyacentes en las hojas del árbol. Para nuestros experimentos hemos ajustado a 4 el número mínimo de compuestos permitido por hoja.

Las redes neuronales utilizadas en nuestros experimentos hacen uso del algoritmo de retro-propagación tradicional (*back propagation*), el cual ha sido utilizado con anterioridad en la literatura QSPR (Niculescu, 2003). Un total de cincuenta redes se utilizaron para definir el ensamble. La arquitectura de cada red es una sola capa oculta con tres nodos y todas las funciones de activación de los nodos internos de la red son sigmoides. Las redes se inicializaron con diferentes pesos al azar. Para facilitar la optimización de los parámetros de gradiente todos los descriptores se normalizaron antes del entrenamiento. La tasa de aprendizaje y el momento se establecieron en 0.3 y 0.2 respectivamente.

Las redes neuronales y árboles de decisión constituyen técnicas muy diferentes de modelado en su naturaleza. Por una parte, las redes neuronales son una de las técnicas más populares para el modelado QSPR (Basant, Gupta, y Singh, 2016; Jalali-Heravi y Parastar, 2000; Mosier y Jurs, 2002). Por otro lado, al poder modelar cualquier tipo de relación no lineal entre los datos, son hace propensas a generar un sobre-ajuste en los datos de entrenamiento (en ausencia de cualquier mecanismo para evitarlo). Por otro lado, los árboles de decisión son bien aceptados por los usuarios que son capaces de interpretar el significado del modelo con mucha facilidad (Quinlan, 1987). Por lo tanto, la decisión sobre cuál de estos modelos debería ser utilizado debería tener en cuenta de qué manera ha sido entrenado el modelo y, por otro lado, determinar cuánto importa la comprensión de ese modelo.

A continuación, se brindarán detalles de los dos experimentos llevados a cabo. Todos ellos se realizaron utilizando la herramienta de minería de datos Weka (Hall et al. 2009).

Para el primer experimento, se reporta el rendimiento de nuestros modelos usando una sexta parte del conjunto de datos (20 compuestos) como conjunto de testeo. Al utilizar árboles de decisión, el error absoluto medio (MAE) es 0.15 ± 0.04 ("±valor" corresponde a los intervalos de confianza calculados a un nivel del 95%). La raíz del error cuadrado medio (RMSE) es 0.18 y el coeficiente de determinación (r^2) es de 0.73. El ensamble de redes neuronales en esta misma partición de datos reporta un ligero decaimiento en la exactitud de regresión en comparación con el modelo anterior: MAE = 0.17 ± 0:04, RMSE = 0.19 y r^2 = 0.66. A continuación, haremos foco en analizar el modelo obtenido con árboles de decisión ya que, en este caso, fue el de mejor rendimiento.

La Figura 4.2 muestra las condiciones en los nodos internos del árbol de decisión y las regresiones lineales utilizadas en las hojas, mientras que La Figura 4.3 muestra la gráfica con la predicción individual de cada compuesto de testeo con el mejor ajuste lineal de nuestro modelo. El análisis de la estructura de árbol nos permite comprender el modelo utilizado para la predicción (Figura 4.2). La primera decisión del árbol se basa en el valor de Se; si es inferior a 16.025, esto conduce a una hoja con una regresión simple utilizando sólo tres de los cinco descriptores disponibles, a saber: ALOGP, Mor29u y Se. Realizando una inspección de la estructura de los compuestos que están asociados a esta hoja, se puede apreciar que la mayoría de ellos tienen una cadena de carbono corta y halógenos con valores bajos de log P_{liver}.

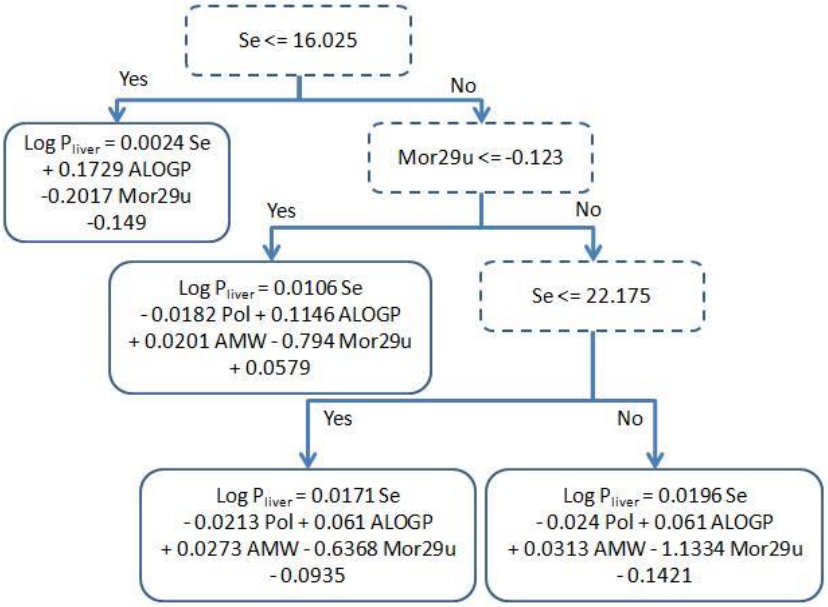


Figura 4 .2 Modelo de árbol de decisión obtenido después de separar para testeo el 16,6% de los datos y utilizando el algoritmo M5p.

Esta separación es coherente desde un punto de vista físico-químico: pequeñas moléculas polares tienen una mayor afinidad con los medios de sangre que las más largas.

Otra observación es que AMW y Pol presentan un elevado coeficiente de correlación de Pearson (≈ 0.56) con Se (Tabla 4.2) y por lo tanto sus contribuciones pueden ser explicadas principalmente por Se.

Descriptor	**r (coeficiente de correlación de *Se* versus descriptor)**	
	$Se \leq 16.025$	$Se > 16.025$
AMW	-0.55	-0.33
Pol	0.57	0.32
ALOGP	0.12	0.75
Mor29u	0.09	-0.15

Tabla 4.2 Coeficiente de correlación de Se versus AMW, Pol, ALOGP y Mor29u.

Cuando el valor de Se es mayor que 16.025, hay tres regresiones lineales diferentes utilizando los cinco descriptores. A partir de la Tabla 4.2, podemos ver que la correlación de AMW y Pol con Se son muy inferiores a lo que sucede en la rama izquierda, y de ahí que ahora se hacen necesarios en el modelo. Se puede notar que todos los coeficientes conservan el mismo signo, lo que está indicando que la contribución de los descriptores en el modelo es siempre la misma, y las diferencias en los coeficientes provienen de la producción de un mejor ajuste a los compuestos asignados a una hoja específica.

Podemos comparar también en Tabla 4.2 que ALOGP se vuelve más correlacionado con Se en la rama derecha que en la izquierda. De esta manera, podemos ver en la Figura 4.2 que hay una caída en el valor absoluto del coeficiente asignado al descriptor ALOGP (0.061 y 0.1146) en la rama derecha en comparación con la de la rama izquierda (0.1729). Un análisis más a fondo de la relevancia de los descriptores fisicoquímicos se puede encontrar en la próxima sección.

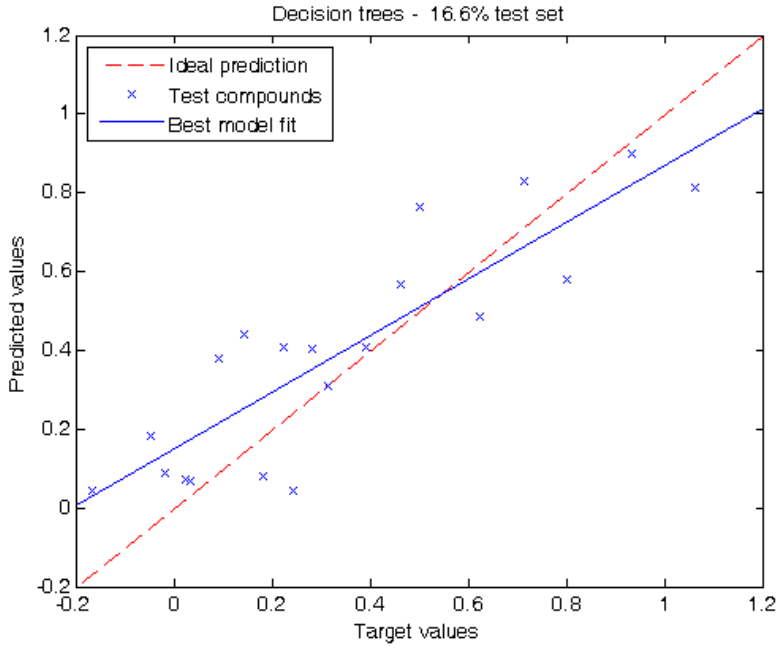


Figura 4.3 Valores reales versus valores predichos usando el 16.6% de los compuestos para testeo con el árbol de decisión mostrado en la Figura 4.2.

Por otro lado, en nuestro segundo experimento se evaluaron los resultados utilizando la mitad de los compuestos de la base de datos (61 compuestos) para testeo. Al utilizar árboles de decisión se obtuvieron los siguientes indicadores: MAE = 0.15 ± 0:04, RMSE = 0.21 y r^2 = 0.62. Utilizando un ensamble de redes neuronales, los resultados mostraron un mejor rendimiento de predicción con MAE = 0.16 ± 0:03, RMSE = 0.20 y r^2 = 0.66. La Figura 4.4 muestra los valores de predicción para cada compuesto de testeo usando el ensamble de redes neuronales.

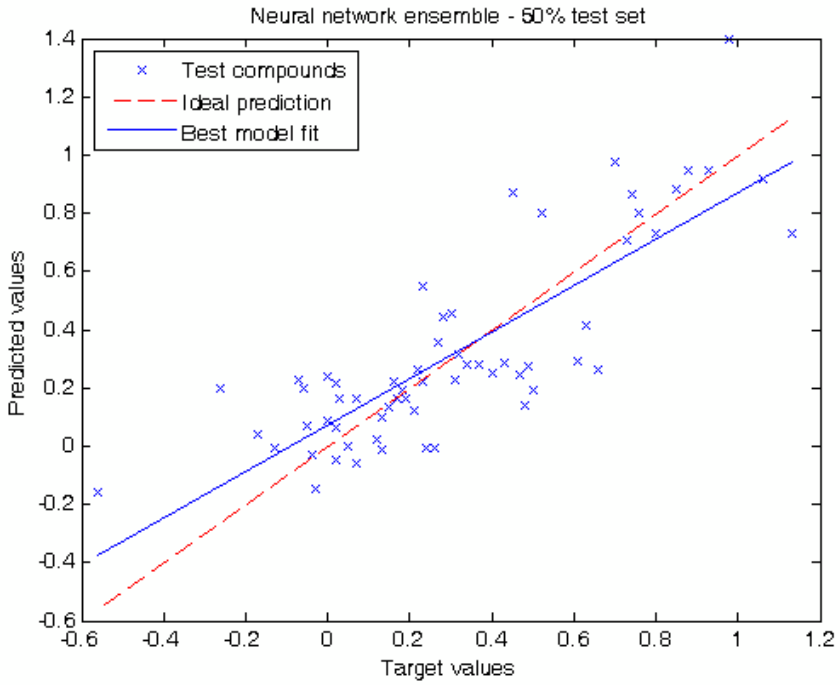


Figura 4.4 Valores reales versus valores predichos utilizando el 50% de los compuestos para testeo con ensamble de redes neuronales.

Los resultados obtenidos muestran una mejora sobre los resultados publicados por Abraham et al. (2007) ya que sus experimentos utilizando el mismo conjunto de datos y el mismo tamaño del conjunto de testeo arrojaron un RMSE = 0.221 y r^2 = 0.481.

4.3.1.4 Relevancia fisicoquímica de los descriptores seleccionados

El objetivo de este apartado es analizar la relación entre descriptores moleculares y la propiedad objetivo a fin de proporcionar la justificación fisicoquímica del modelo resultante. Cuando la interpretación del modelo QSPR es consistente con las teorías y el conocimiento de los mecanismos existentes, el modelo se vuelve más atractivo para los expertos (OECD, 2007). A pesar de que no siempre es posible encontrar una interpretación global, es deseable proveer una explicación para el modelo de una manera "mecanicista" (Gramatica, 2010).

Los cinco descriptores elegidos para el modelo proporcionan en mayor o menor medida información importante acerca de las propiedades moleculares relacionadas con la capacidad de la molécula para distribuirse entre los dos medios en estudio: el tejido del hígado y la sangre. En la Figura 4.5 se muestran las relaciones entre los valores de descriptores y valores de log P_{liver}. Nuestro análisis se centró en algunas familias químicas representativas destacadas con colores (alcanos, alcoholes, compuestos aromáticos y algunos hidrocarburos halogenados estructuralmente similares).

El descriptor AMW (peso molecular dividido por el número de átomos) (Figura 4.5a) discrimina las moléculas teniendo en cuenta su composición atómica (tipo y cantidad). Para dar un ejemplo, los alcanos (C_n H_{2n+2}) y los compuestos aromáticos (C_n H_n) sólo están constituidos por carbonos e hidrógenos. Dado que cada familia tiene una tasa distinta de C/H, presentan un valor específico de AMW, aun cuando los compuestos que pertenecen a cada una de las familias son ligeramente diferentes.

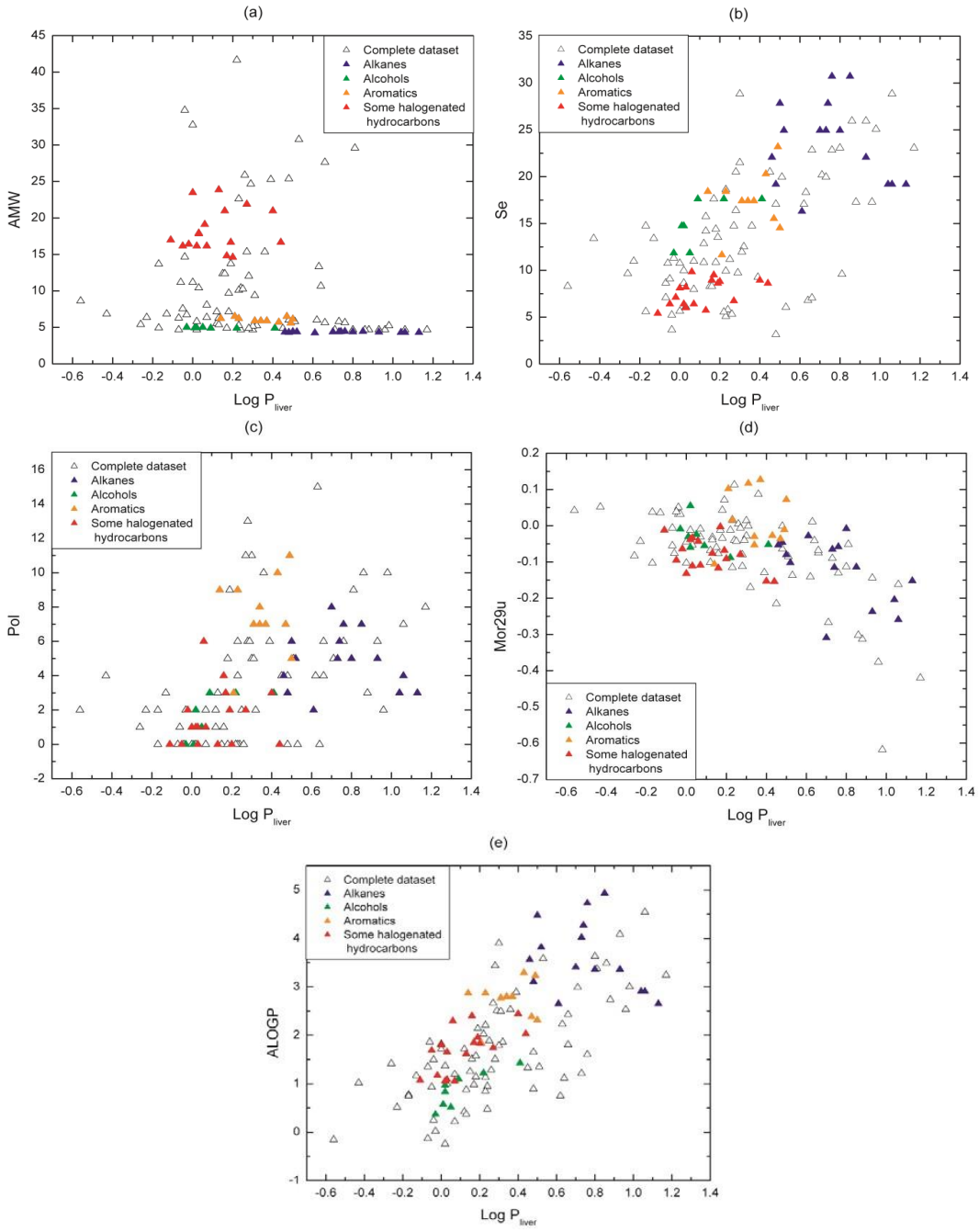


Figura 4.5 Gráfico de valores de los descriptores versus los valores de log P_{liver} para el conjunto de datos completo. **a**) AMW; (**b**) Se; (**c**) Pol; (**d**) Mor29u and (**e**) ALOGP.

Cuando estas familias pueden separarse de los datos de todo el conjunto en el gráfico, las diferencias en sus propiedades fisicoquímicas se hacen más evidentes, por ejemplo, su polaridad (que está relacionada con la afinidad de una molécula por un medio acuoso o por uno no polar). En esta figura, se puede observar también el comportamiento de las familias no polares como los alcanos, que tienden a tener valores altos de log P_{liver}, mientras que las familias polares como los alcoholes tiene valores inferiores de $logP_{liver}$. El mismo análisis puede aplicarse a los compuestos aromáticos y los hidrocarburos halogenados.

Algo similar sucede con el descriptor Se (Figura 4.5b) que logra la discriminación de las familias de VOCs con la suma de electronegatividades atómicas Sanderson (escaladas en el átomo de carbono).

El descriptor Pol (número de polaridad) y Mor29u (Representación Molecular 3D de las estructuras basadas en la difracción de electrones - señal 29/no ponderado) resaltan las propiedades estructurales 2D y 3D respectivamente y se representan gráficamente en la Figura 4.5c y 4.5d. Pol se relaciona con las propiedades estéricas de moléculas y se calcula sobre la matriz de distancia como el número de pares de vértices a una distancia topológica igual a tres (es decir, número de terceros vecinos) (Platt, 1947). En la figura 4.5c, se puede ver que Pol presenta valores bajos o cero para cadenas de carbono cortas mientras que toma valores más altos (entre 4 y 16) para las estructuras más largas (por ejemplo, la mayor parte de los hidrocarburos y alcanos halogenados largo respectivamente). En otras palabras, Pol es bajo o igual a cero para los compuestos con pocos átomos porque tienen un pequeño número de terceros vecinos y ocurre lo contrario para las moléculas largas. Por lo tanto, este descriptor funciona como un filtro específico que discrimina moléculas por su longitud de cadena.

Mor29u (3D-Morse - señal 29/no ponderado) pertenece a la familia de descriptores 3D Morse (Representación Molecular 3-D de estructuras basadas en la difracción de electrones). Se basan en la idea de obtener información de las

coordenadas atómicas 3D mediante la transformación utilizada en los estudios de difracción de electrones para la preparación de las curvas de dispersión teóricas (Schuur, Selzer, y Gasteiger, 1996). Los descriptores 3D MORSE se derivan de las proyecciones de átomos de moléculas a lo largo de diferentes ángulos, como en la difracción de electrones. Ellos representan diferentes vistas de toda la estructura de la molécula, aunque su significado sigue siendo todavía poco claro (Saíz-Urra, Pérez González, y Teijeira, 2006). Mientras que su influencia no parece tener una explicación precisa, su inclusión se justifica principalmente por la regresión: fue seleccionado por el método de selección de características y en todos nuestros experimentos. La eliminación de este descriptor de nuestras ecuaciones conduce a un descenso notable en la calidad de predicción, tanto en el entrenamiento como en la validación. A pesar de esto, podemos analizar parcialmente su contribución. Se puede ver en la Figura 4.5d que Mor29u toma valores positivos y negativos, porque la ecuación original incluye el término $\sin(s * r_{ij})/ s * r_{ij}$ (Schuur et al., 1996), donde s mide el ángulo de dispersión y r_{ij} representa las distancias interatómicas entre los átomos i y j. Entonces, el signo del descriptor por sí solo no es determinante para la relación con el objetivo. Otra observación de la Figura 4.5d es que las familias químicas que no son separadas en regiones como ocurre con AWM y Se (Figura 4.5a, 4.5b). Esto parece ser coherente debido a que los componentes de una familia química comparten muchas propiedades fisicoquímicas (polaridad, la movilidad, enlaces de hidrógeno, etc.) al margen de la estructura 3D.

Por último, ALOGP (coeficiente de partición octanol-agua Ghose-Crippen) da información relevante acerca de la afinidad molecular por un medio octanol-agua. De hecho, ALOGP es un descriptor que aparece comúnmente en los modelos sobre los coeficientes de partición (Balaz y Lukacova, 1999; H. Zhang y Zhang, 2006). Se calcula a partir de un modelo que consta de una ecuación de regresión basado en la contribución de hidrofobicidad de 120 tipos de átomos (Ghose, Viswanandhan, y Wendoloski, 1998; Viswanandhan, Ghose, Revankar, y Robins, 1989; Viswanandhan, Reddy, Bacquet, y Erion, 1993). Cada átomo en

cada estructura se clasifica en uno de los 120 tipos de átomos. Así, un valor de log P estimado para cualquier compuesto está dado por ALOGP = $\sum_i n_i * a_i$, donde n_i es el número de átomos de tipo i y a_i es la constante de hidrofobicidad correspondiente.

Se puede observar en la Figura 4.5e, que cada VOC tiene su propio valor ALOGP independientemente de su familia química. Es decir, este descriptor es sensible a diferencias mínimas en la estructura molecular. Como era de esperar, se puede observar la correlación entre este descriptor y log P_{liver} (Figura 4.5e), ya que las moléculas polares tienen valores bajo de ALOGP y log P_{liver} (por ejemplo, alcoholes e hidrocarburos halogenados) y las no polares tienen valores altos (por ejemplo, alcanos y compuestos aromáticos).

4.3.1.5 Conclusiones

Se presentaron nuevos modelos para la predicción de los coeficientes de partición sangre-hígado para compuestos orgánicos volátiles (VOCs) siguiendo el enfoque de modelado QSPR. Los modelos generados han demostrado una capacidad de predicción similar entre sí, y han superado significativamente los resultados obtenidos por Abraham et al. (Abraham et al., 2007), que es el único trabajo en esta área que utiliza el mismo conjunto de datos. Según nuestro conocimiento, este es el mayor conjunto de datos de VOCs con sus valores log P_{liver} asociados.

Un aspecto clave de los buenos resultados de nuestros enfoques se basa en la cuidadosa selección semi-automática de los descriptores (enfoque híbrido) utilizados para construir nuestros modelos. Este enfoque semi-automático se puede aplicar para modelar también otras propiedades y otros compuestos, siempre que estén disponibles los métodos estadísticos y el conocimiento del experto. Sin embargo, es importante ser siempre cauteloso en el uso de enfoques

QSPR. Mientras que la exactitud en las predicciones de los compuestos no vistos es estimada por el uso de un conjunto de testeo, es difícil evaluar la precisión de la predicción de los compuestos que quedan fuera del dominio de aplicabilidad del modelo. El dominio de aplicabilidad del modelo se ve afectado por la diversidad del conjunto de entrenamiento, la complejidad o la dimensionalidad de su representación y el modelo de predicción (Dragos, Gilles, Alexandre, Marcou, y Varnek, 2009; Soto, Vazquez, Strickert, y Ponzoni, 2011).

Por estas razones, nuestro modelo no predeciría con la misma precisión para compuestos de una naturaleza diferente a los presentes en el conjunto de entrenamiento. Sin embargo, el uso de estrategias que incluye el conocimiento del experto durante la fase de modelado conduce a modelos más admisibles que son más fáciles de interpretar y más probable que vinculen mejor a compuestos que no se ven.

Finalmente, este trabajo contribuye técnicas fiables para predecir métricas relacionadas con la exposición a sustancias químicas en el medio ambiente, que pueden ser aplicadas a la evaluación de riesgos y la toma de decisiones en políticas de salud pública.

4.3.2 Aprendizaje de descriptores en el diseño de fármacos

En esta sección presentaremos modelos QSAR de regresión y clasificación cuyos descriptores fueron obtenidos utilizando la combinación CODES–TSAR presentada anteriormente. Se estudiaron dos propiedades de relevancia para el diseño de fármacos: *HIA (Human Intestinal Absorption, HIA)* y *BBB (Blood-Brain Barrier)*. Es necesario aclarar que el estudio que se analizará brevemente aquí, forma parte de un análisis completo que será presentado en el capítulo 5. El objetivo allí, fue hacer una contrastación de estos resultados con los obtenidos a partir de

DELPHOS, además de analizar también modelos entrenados con descriptores obtenidos por medio de ambas técnicas.

4.3.2.1 Experimentos realizados

Para HIA, se utilizó un conjunto de datos de 202 compuestos previamente publicado en Guerra et al. (2010), mientras que para BBB, el conjunto de datos utilizado tiene 108 compuestos previamente publicado en Guerra et al., (2008). Para ambas propiedades se entrenaron diferentes modelos utilizando WEKA: Regresión Lineal (LR), Árboles de Decisión (DT), Redes Neuronales (NN), Bosque Aleatorio (BA, *random forest*) y Comité Aleatorio (CA, *random committee*) y se utilizó en cada caso la configuración de parámetros por defecto. El rendimiento fue evaluado utilizando distintas métricas en WEKA. Para el caso de regresión se reportan el coeficiente de correlación (CC), el error absoluto relativo (RAE) y la raíz del error cuadrático relativo (RRSE). Para clasificación se reportaron el porcentaje de casos correctamente clasificados (%CC), el área de la curva de ROC junto con la matriz de confusión y la raíz del error cuadrático relativo (RRSE). Para la deducción de los modelos, se probaron con distintas configuraciones para la división del conjunto de datos en entrenamiento y testeo (50-50, 66-34, 75-25).

En el caso de HIA, obtuvimos un conjunto de tres descriptores, CT_{HIA}. Los mejores modelos obtenidos para regresión y clasificación se muestran en la Tabla 4.3.

	Mejores modelos QSAR de regresión			**Mejores modelos QSAR de clasificación**			
Conjunto de descriptores	CC	**% de datos en el conjunto de entrenamiento**	**Método**	**%CC**	**ROC**	**% de datos en el conjunto de entrenamiento**	**Método**
CT_{HIA} (CODES-T1, CODES-T2, CODES-T3)	0.23	75%	Redes Neuronales	72%	0.517	75%	Bosque Aleatorio

Tabla 4.3 Métricas estadísticas de los mejores modelos obtenidos para CT_{HIA}.

Por otro lado, para BBB se obtuvo otro subconjunto de tres descriptores, CT_{BBB}. Los resultados pueden verse en la Tabla 4.4.

	Mejores modelos QSAR de regresión			**Mejores modelos QSAR de clasificación**			
Conjunto de descriptores	CC	**% de datos en el conjunto de entrenamiento**	**Método**	**%CC**	**ROC**	**% de datos en el conjunto de entrenamiento**	**Método**
CT_{BBB} (CODES-T1, CODES-T2, CODES-T3)	0.41	75%	Comité Aleatorio	72.97%	0.602	66%	Comité Aleatorio

Tabla 4.4 Métricas estadísticas de los mejores modelos obtenidos para CT_{BBB}.

4.3.2.2 Conclusiones

Para ambos casos, podemos observar que los modelos entrenados con los descriptores generados por CODES-TSAR tienen un rendimiento relativamente bajo. Se debe tener en cuenta, como se mencionó al comienzo de esta sección, que este análisis formó parte de un estudio más profundo que se presentará en el capítulo siguiente, por lo que las conclusiones que podemos brindar aquí son un poco acotadas. Adelantando de alguna manera los resultados del análisis completo, los resultados presentados aquí con los subconjuntos de descriptores CT_{HIA} y CT_{BBB} no logran superar en rendimiento a los modelos entrenados con descriptores obtenidos a partir de DELPHOS que se detallarán en el capítulo 5. Sin embargo, veremos que para el caso de HIA la combinación de los descriptores de DELPHOS y CODES-TSAR en un mismo modelo permite superar el rendimiento individual para cada técnica.

4.4 Sumario

En este capítulo se ha profundizado un poco más en conceptos relacionados con los procesos para la selección y aprendizaje de descriptores. Se presentaron estudios de dos propiedades: log P_{liver} (medioambiente) y *resistencia a la rotura* (diseño de materiales).

Para el caso de log P_{liver} el análisis estuvo enfocado en el proceso de selección de los descriptores, presentando una metodología en la que se incorpora al experto en el dominio durante el desarrollo. De esta manera, se logró obtener un modelo que presentó un buen rendimiento en términos estadísticos con el valor agregado de que cada descriptor puede explicarse respecto de su aporte en términos físico-químico con respecto a la propiedad. Este estudio representó un punto de partida para el desarrollo de una herramienta para dar soporte al experto en este proceso, que será presentada en el próximo capítulo.

Para la propiedad *resistencia a la rotura*, se realizó un primer acercamiento en la utilización de una técnica de aprendizaje de descriptores. En este caso los resultados en primera instancia no resultaron favorables, ya que para la propiedad estudiada los modelos con los descriptores generados no presentaron un rendimiento aceptable. En el capítulo siguiente, se continuará estudiando esta propiedad con el fin de mejorar el rendimiento de estos modelos, a través del uso de la herramienta de soporte desarrollada.

Capítulo 5: Analítica Visual Aplicada a la Selección de Descriptores

Muchas veces el modelador se enfrenta a situaciones en las que utiliza diferentes métodos (o herramientas) de selección de descriptores y obtiene de cada uno de ellos diferentes resultados. Por lo tanto, se enfrenta a la difícil tarea de determinar cuál de todos ellos usar. En el presente capítulo se explicará una parte importante de la contribución de esta tesis que pretende dar soporte a ese desafío: el desarrollo y aplicación de una herramienta de software que denominamos VIDEAN (por sus siglas en inglés "**V**isual and **I**nteractive **De**scriptor **AN**alysis"), que combina métodos estadísticos con visualizaciones interactivas con el fin de permitirle al usuario elegir un conjunto de descriptores para la predicción de una propiedad fisicoquímica o biológica determinada (Martínez et al., 2015). Además, se presentan múltiples casos de estudio. A través de ellos, se busca describir cómo un experto puede utilizar esta herramienta para seleccionar un subconjunto de descriptores entre un grupo de subconjuntos candidatos. Además de mostrar cómo es posible modificar subconjuntos de descriptores existentes e incluso incorporar nuevos descriptores de acuerdo a su propio conocimiento de la propiedad bajo estudio. La herramienta se encuentra disponible para su uso en http://lidecc.cs.uns.edu.ar/VIDEAN/.

5.1 Aspectos generales

Como se ha descripto en capítulos anteriores, el diseño de modelos QSPR es un problema difícil, donde la selección de los descriptores más relevantes constituye un paso clave del proceso. Varios métodos de selección de características que abordan este paso se concentran en las asociaciones estadísticas entre los

descriptores y la propiedad objetivo, mientras que el conocimiento químico se deja fuera del análisis. Por esta razón, la interpretabilidad y la generalidad de los modelos QSPR obtenidos por estos métodos de selección pueden ser afectados negativamente. Por lo tanto, es necesario un enfoque que integre los conocimientos del experto al proceso de selección y de esta manera aumentar la confiabilidad en el conjunto final de descriptores y por consiguiente en el modelo predictivo.

Así es como uno de los objetivos del desarrollo de esta herramienta fue incorporar el conocimiento del experto al proceso de selección de características por medio de una exploración visual interactiva de los datos y la ayuda de herramientas estadísticas. Para esto se presentan visualizaciones coordinadas (Sadana y Stasko, 2016) que capturan diferentes relaciones e interacciones entre los descriptores, propiedad objetivo y subconjuntos de descriptores candidatos.

Hay dos preguntas de investigación que pretendimos responder llevando a cabo este trabajo. La primera de ellas es si podemos aprovechar los resultados de diferentes enfoques de selección de descriptores para llegar a una decisión con más fundamento en el subconjunto de descriptores seleccionados. Si bien existen varios métodos basados en conjuntos y sistemas de votación (Cao, Xu, Liang, Chen, y Li, 2010; Q. Zhang, Hughes-Oliver, y Ng, 2009) estos enfoques dejan el conocimiento químico fuera del proceso de selección. La segunda pregunta de investigación es cómo podemos implicar al experto del dominio (por ejemplo, un químico), de modo que él o ella puedan incorporar su conocimiento y experiencia durante el proceso de selección de características de una manera semi-automática.

Con la implementación de VIDEAN, se propone responder estas dos preguntas combinando métodos estadísticos con visualizaciones interactivas. Este tipo de enfoque pertenece al área emergente de la analítica visual, presentada en el capítulo 2.

La herramienta propone un enfoque que se centra en la difícil tarea del análisis y selección de subconjuntos de descriptores para modelos QSPR. A través de técnicas de análisis exploratorios y visuales, el software facilita la comparación de múltiples descriptores y subconjuntos de descriptores procedentes de diferentes métodos de selección automáticos. La idea principal es utilizar a VIDEAN como un sistema de soporte de decisiones para analizar diferentes subconjuntos de descriptores que han sido obtenido previamente mediante otra metodología. Sin lugar a dudas, la tarea del analista es compleja ya que hay múltiples aspectos que están involucrados en las decisiones sobre el modelo. Por lo tanto, se requieren diferentes estrategias de exploración de datos con el fin de integrar estas múltiples piezas de información. Es importante aclarar que un modelo QSAR está constituido por un subconjunto de descriptores y la relación que asocia estos descriptores con una propiedad objetivo. Sin embargo, en lo subsiguiente cuando hablemos de "modelo" en el contexto de VIDEAN, haremos referencia sólo a un subconjunto candidato de descriptores. Por lo tanto, el uso de la palabra "modelo" se utilizará con frecuencia aquí como una simplificación de "subconjunto candidato de descriptores".

En la siguiente sección se describirá la implementación de la herramienta. Luego, se presentarán y analizaran múltiples casos de estudios en los que se ha utilizado VIDEAN.

5.2 Arquitectura de la herramienta

Con VIDEAN nos proponemos visualizar diferentes aspectos relacionados con la información requerida para el modelado, el descubrimiento de relaciones ocultas entre los descriptores y sus relaciones con la propiedad objetivo. De esta manera, los objetivos específicos que guiaron el diseño de estas visualizaciones pueden ser definidos como sigue:

1. Evitar la selección de descriptores redundantes en modelos QSPR. Esto significa que, si dos o más descriptores están aportando una información similar, la

herramienta debe ayudar a elegir los descriptores más significativos o adecuados para el modelo, y de esta manera mantener el subconjunto de descriptores con la menor cardinalidad posible.

2. Que la información aportada entre los descriptores seleccionados sea complementaria. Esto significa que cada descriptor del modelo debe ser relevante para la predicción de la propiedad objetivo. En otras palabras, si se suprimiera un descriptor, la precisión de la predicción del modelo debería empeorar.

5.2.1 Visualizaciones interactivas propuestas

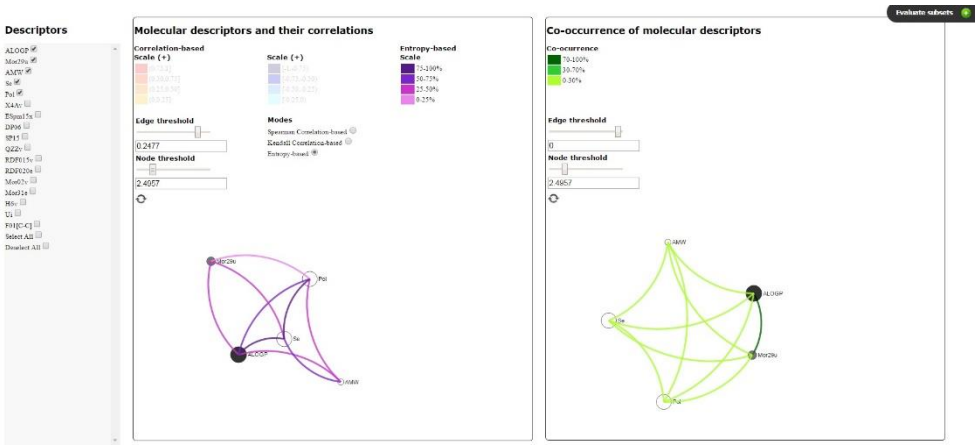
En esta sección se detallarán las visualizaciones que ofrece VIDEAN. La interfaz se organiza en torno a los cuatro gráficos representados en la Figura 5.1. En la parte superior de la pantalla se pueden observar dos grafos no dirigidos que representan asociaciones entre pares de descriptores. La sección inferior de la pantalla contiene un grafo bipartito, que representa la relación entre los subconjuntos candidatos de descriptores (llamados "modelos" en esta herramienta) y los descriptores individuales, y un área de gráficos interactivos, que muestra las diferentes relaciones entre los descriptores y la propiedad objetivo. De esta manera, el modelador puede analizar múltiples aspectos que intervienen en el proceso de selección de descriptores de forma simultánea.

5.2.1.1 Grafos no dirigidos para el análisis de pares de descriptores

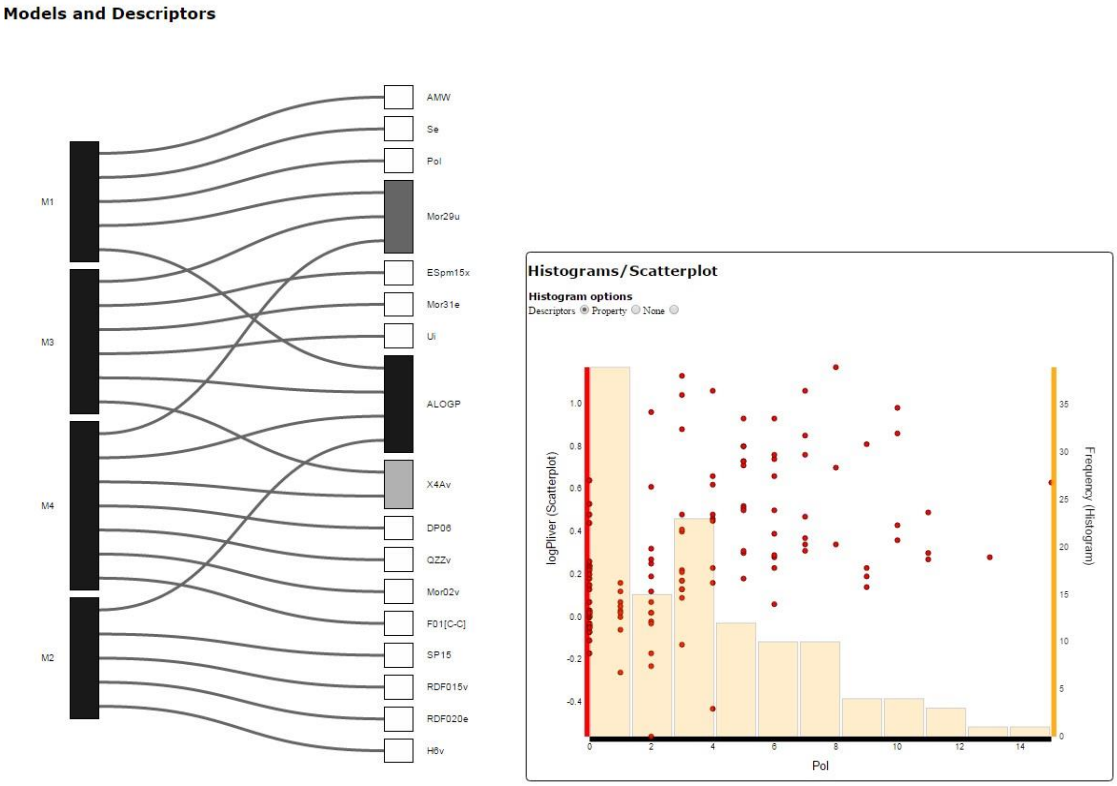
En estos gráficos la información se representa en forma de grafos no dirigidos, donde cada nodo (círculo) representa un descriptor seleccionado por al menos uno de los modelos (Figura 5.1-a). El gráfico de la izquierda tiene un papel central en el análisis y se llama grafo no dirigido primario (G_P), mientras que el gráfico de

la derecha juega un papel complementario y se llama grafo no dirigido secundario (G_s). En ambos gráficos, el color del nodo utiliza una escala de grises para indicar la proporción de modelos en los que el descriptor ha sido seleccionado: blanco, si se eligió únicamente por un único modelo, y negro, si fue elegido en todos los modelos. De esta manera, el consenso entre los diferentes modelos se incorpora en el análisis (Ganguly et al., 2006).

El grafo G_p admite distintos modos de visualización para representar diferentes tipos de relaciones entre los descriptores. Los dos modos principales son: basado en entropía y basado en correlación, y se puede seleccionar uno u otro dependiendo de lo que se quiera analizar. En el primer modo, los tamaños de los nodos están asociados con la entropía condicional del descriptor con respecto a la propiedad objetivo. Un tamaño más pequeño del nodo indica un mayor valor de la entropía condicional y un tamaño más grande del nodo indica un menor valor de la entropía condicional. En general, la entropía condicional es una medida de la cantidad de incertidumbre sobre una variable aleatoria cuando se conoce el valor de la otra variable aleatoria (Cover y Thomas, 1991). En este caso, se trata de una medida de cuánto se mantiene la incertidumbre sobre el valor de la propiedad objetivo cuando conocemos el valor del descriptor. En este modo, las aristas representan la información mutua entre los descriptores. La información mutua (IM) mide la cantidad de información que una variable aleatoria contiene con respecto de otra. Por lo tanto, la información mutua entre dos descriptores es la reducción en la incertidumbre de uno de ellos debido al conocimiento del otro, y viceversa (Cover y Thomas, 2012). Se debe tener en cuenta que la IM es una métrica, por lo que el valor obtenido de este cálculo será siempre no negativo y simétrico (Kojadinovic, 2005). El color de las aristas se utiliza para cuantificar pesos de las mismas, y la escala va de rosa a violeta e indica el valor de IM entre dos descriptores. El color rosa claro se utiliza cuando los descriptores son independientes (IM = 0). En el caso opuesto, el color violeta oscuro se utiliza cuando los descriptores son idénticos (altos valores de IM), es decir que la información derivada de uno de ellos puede ser usada para representar al otro.



a)



b) c)

Figura 5.1 Grafos primario (arriba a la izquierda izquierda) y secundario (arriba a la derecha) no dirigidos (G_p y G_s, respectivamente), que se enfocan en las asociaciones entre pares de descriptores. Abajo a la izquierda, se puede ver el grafo bipartito que representa los descriptores moleculares agrupados en cada modelo. Abajo a la derecha, se observa el diagrama de dispersión y el histograma con el fin de mostrar la relación entre los descriptores y la propiedad objetivo.

En el segundo modo, los tamaños de los nodos están asociados con la correlación de Spearman (o Kendall) entre los descriptores y la propiedad objetivo, mientras que las aristas representan la correlación de Spearman (o Kendall) entre descriptores. Aquí, un tamaño más pequeño de nodo indica una correlación más baja y un tamaño más grande de nodo indica una mayor correlación con la propiedad objetivo.

Las gamas de colores de las aristas van de rojo a amarillo para la correlación positiva (rojo = 1) y de azul a azul claro para la correlación negativa (azul = -1). Las correlaciones de Spearman y Kendall, a diferencia de la conocida correlación de Pearson, permiten la identificación de las relaciones no lineales.

En G_s (Figura 5.1-a, grafo de la derecha), las características del nodo (tamaño y color) se fijan automáticamente por el modo seleccionado para G_p. En este grafo las aristas representan el grado de co-ocurrencia de los descriptores en los modelos. En otras palabras, la co-ocurrencia de dos descriptores se calcula como el número de modelos en los que ambos descriptores aparecen simultáneamente sobre el número total de modelos. La gama de colores que se utilizan para las aristas varía de color verde oscuro (alta co-ocurrencia) a verde claro (baja co-ocurrencia). En este grafo el modelador puede reconocer pares de descriptores concurrentes relevantes, es decir aquellos que tienen gran importancia para la predicción de la propiedad como un par (no sólo de forma individual). Estas situaciones podrían estar relacionadas con la información físico-química complementaria entre dos descriptores. Por ejemplo, el peso molecular y polarizabilidad juntos puede ser útil para predecir una propiedad de partición entre dos disolventes, aunque la contribución individual de estos descriptores puede no ser tan relevante.

5.2.1.2 Grafo bipartito para el análisis de modelos versus descriptores moleculares

En este gráfico la información está representada por un grafo bipartito, donde todos los nodos son representados por rectángulos. Los nodos de la izquierda representan los modelos y los nodos de la derecha representan los descriptores de estos modelos. Las aristas indican ocurrencia de un descriptor en un modelo. De esta manera, podemos tener una visión general de los modelos bajo estudio (Figura 5.1-b).

5.2.1.3 Gráficos adicionales para el análisis de la propiedad versus descriptores

Una característica adicional asociada con los grafos consiste en mostrar la distribución de los valores del descriptor con respecto a la propiedad objetivo haciendo uso de las visualizaciones adicionales (Figura 5.1-c). Al hacer clic en un nodo de G_s se muestra un diagrama de dispersión con los valores del descriptor frente a los valores de la propiedad objetivo para todos los compuestos del conjunto de datos. Además, se pueden visualizar dos histogramas que indican la frecuencia de los valores de descriptores y de la propiedad objetivo. Estos histogramas pueden superponerse sobre el diagrama de dispersión. En este punto, el modelador puede tener una mejor comprensión de la contribución de cada descriptor a la modelización de la propiedad objetivo y el tipo de relación (lineal, cuadrática, cúbica, etc). Por ejemplo, en trabajos anteriores (Palomba et al., 2012; Soto et al., 2011), los mejores modelos se obtuvieron mediante la combinación de descriptores que cubren diferentes subregiones del dominio químico. En otras palabras, la exploración de esta visualización permite evaluar la contribución de los diferentes descriptores al modelo.

5.2.1.4 Lista de descriptores

Esta lista muestra todos los descriptores que participan en el análisis (Figura 5.1-a). Tiene un papel importante, ya que controla lo que puede ser visualizado por los grafos G_p y G_s. La selección (des-selección) de un descriptor de la lista implica la adición (eliminación) del nodo correspondiente en G_p y en G_s.

5.2.1.5 Modelos de predicción

VIDEAN permite modelar la propiedad de destino utilizando diferentes métodos estadísticos. Los descriptores que se utilizan para construir el modelo son los que se encuentran seleccionados en la lista de descriptores. Hay tres métodos disponibles: regresiones lineales, árboles de decisión y redes neuronales, que pueden ser parametrizados. Estos métodos de predicción se implementan utilizando las librerías de WEKA (M. Hall et al., 2009). Aunque el objetivo de VIDEAN no es sólo determinar el nivel predictivo desde el un punto de vista meramente estadístico, esta característica permite la comparación de la capacidad predictiva de varios subconjuntos de descriptores utilizando algunos métodos de referencia.

5.2.1.6 Interacción con las visualizaciones

Los grafos no dirigidos permiten varias interacciones con el usuario. Inicialmente, estos grafos se muestran reducidos en sus nodos y aristas en base a diferentes umbrales. Cuando se cambia el valor del umbral del nodo (arista), el grafo se actualiza de acuerdo con estos valores.

Cuando se pasa el puntero sobre un nodo, los nodos y aristas conectados a él son resaltados mientras que el resto del grafo es atenuado, dejando un claro contraste sin perder el contexto global. Cuando se pasa el puntero sobre una

arista, se muestra el valor numérico que está codificado por el color. Este número puede representar un valor de información mutua, una correlación, o el número de modelos, dependiendo del grafo y del modo en que se encuentre.

Haciendo doble clic sobre una escala de colores de los grafos, se podrán ver los enlaces dentro de esta gama de colores y todo lo demás se volverá opaco. Al hacer doble clic sobre un nodo, se abrirá una lista completa de descriptores, con la posibilidad de buscar alguno en particular. Se debe tener en cuenta que, mediante el uso de estas interacciones, un modelador podría descubrir descriptores interesantes. Por ejemplo, descriptores que no sean redundantes, que hayan sido seleccionados en la mayoría de los modelos y que estén fuertemente correlacionados con la propiedad objetivo pueden ser fácilmente identificados en el grafo G_p mediante la detección de los nodos más grandes y más oscuros conectados entre sí con enlaces de color claro.

Por último, las interacciones para el grafo bipartito son sutilmente diferentes. Cuando se pasa el puntero sobre un nodo que representa a un modelo, las aristas que conectan a este con sus descriptores se resaltan mientras que los otros descriptores se atenúan. Cuando se pasa el puntero sobre un nodo descriptor, las aristas de los modelos en los que el descriptor está presente se destacan, mientras que las aristas restantes se atenúan. Un punto a tener en cuenta es que los colores utilizados para los nodos que representan descriptores (escala del blanco al negro) son los mismos en los tres gráficos basados en grafos. Además, al hacer clic en un nodo modelo del grafo, sus descriptores se actualizan en la lista principal, y por lo tanto también en los grafos G_p y G_s. El modelador puede utilizar esta visualización para comprender mejor las diferencias entre los subconjuntos de descriptores, y para crear nuevos modelos a partir de la combinación de ellos.

En la siguiente sección, se ilustrarán los posibles usos de la herramienta a través del estudio de las distintas propiedades con las que se trabajaron en la presente tesis.

5.3 Casos de estudio

En esta sección se detallará la mayor parte de la experimentación realizada en el marco de la presente tesis. La misma se organizará de la siguiente manera: en la sección 5.3.1, se continuará con el estudio de la propiedad log P_{liver} presentado en el capítulo anterior. En esta instancia se realizará la selección de descriptores haciendo uso de VIDEAN, obteniendo un modelo de regresión y otro de clasificación. La sección 5.3.2 está avocada el campo del diseño de materiales y al estudio de una propiedad mecánica específica, presentando también en este caso diferentes modelos de regresión y clasificación. Luego, en 5.3.3, se analizan distintos modelos de regresión y clasificación obtenidos para dos propiedades fundamentales para el diseño racional de fármacos.

En estas secciones se presentarán distintos casos de estudios que ilustran diferentes maneras cómo utilizar la herramienta. Además, para algunas propiedades específicas se realizará una comparación de modelos obtenidos a partir de enfoques de selección de características versus técnicas de aprendizaje de descriptores. Adicionalmente, para estos mismos casos, se presentará un enfoque híbrido que combina estas dos variantes.

Para todos los casos de estudio, los descriptores moleculares fueron calculados usando DRAGON (DRAGON, 2007). Se utilizó como herramienta automatizada de selección de descriptores a DELPHOS (Soto, Cecchini, et al., 2009) y para algunos casos WEKA (M. Hall et al., 2009), con el soporte de VIDEAN (Martínez et al., 2015) para agregar el conocimiento del experto al proceso. En otros casos, para el cálculo y aprendizaje de descriptores se utilizó la combinación de herramientas CODES-TSAR (Dorronsoro et al., 2004; TSAR, 2000). Para generar modelos y evaluar performance fueron utilizadas las rutinas de WEKA. Para los casos en los que se utilice otra herramienta, se hará explícita mención.

5.3.1 Medioambiente: análisis de la propiedad *log* P_{liver}

Retomando el caso de estudio presentado en el capítulo 4, aquí se presentará una extensión de ese trabajo, pero ahora utilizando VIDEAN para el proceso de selección de descriptores. Se ilustrará la utilización de la herramienta siguiendo un hilo conductor que nos llevará a elegir un conjunto de descriptores y generar con él un modelo de regresión. Por otro lado, se presentará un análisis similar, pero en este caso para armar un modelo de clasificación para la misma propiedad, pero orientado a cuantificar la potencial toxicidad de un VOC.

5.3.1.1 Elección del mejor subconjunto de descriptores para regresión

Aquí la motivación estuvo centrada en analizar un caso de estudio conocido, con la idea de evaluar la utilización de VIDEAN en la selección de descriptores. La metodología para llevar adelante la experimentación se ilustra en la figura 5.2. En este caso, se utilizó la herramienta para elegir un subconjunto de descriptores a partir de cuatro subconjuntos candidatos obtenidos automáticamente (M1, M2, M3 y M4) a partir del conjunto de datos reportado en el trabajo de Palomba et al. (2012). Este caso de estudio se encuentra disponible en la página web de VIDEAN[5].

Con cada subconjunto de descriptores se generaron distintos modelos de regresión utilizando redes neuronales, regresiones lineales y árboles de decisión. Se aplicó una estrategia de validación cruzada de 4-folds para la partición del conjunto de entrenamiento y validación. En la tabla 5.1 se reporta para cada modelo las siguientes métricas: la mejor calidad predictiva obtenida por los métodos de regresión, su cardinalidad, el número de descriptores frecuentes (es decir, aquellos que aparecen en más de un modelo) y el número de descriptores que comparte con otros modelos.

[5] http://lidecc.cs.uns.edu.ar/VIDEAN/, accediendo a la pestaña "Uses Cases" y luego cargando "Example 1".

Modelo	Calidad predictiva	Cardinalidad	# de descriptores frecuentes	# de descriptores compartidos con otros modelos
M1(ALOGP, Mor29u, AMW, Se, Pol)	r^2 = 0.81 MAE = 0.15 RMSE = 0.20	5	2	2
M2 (ALOGP, SP15, RDF015v, RDF020e, H6v)	r^2 = 0.76 MAE = 0.17 RMSE = 0.23	5	1	1
M3 (ALOGP, Mor29u, X4Av, ESpm15, Mor31e, Ui)	r^2 = 0.79 MAE = 0.16 RMSE = 0.21	6	3	3
M4 (ALOGP, Mor29u, X4Av, DP06, QZZv, Mor02v, F01[C–C])	r^2 = 0.79 MAE = 0.16 RMSE = 0.21	7	3	3

Tabla 5.1 Modelos candidatos obtenidos para log P_{liver}.

Cuando se comparan subconjuntos de descriptores con capacidades predictivas similares, la elección de un subconjunto generalmente se centra en aquel con cardinalidad inferior. Esto se debe a que un subconjunto con menor número de descriptores es generalmente más fácil de interpretar y tiene más probabilidades de ser generalizable (basado en el principio de la navaja de Ockham, (Audi, 1999)). A pesar de esto, y dado que las diferencias entre las cardinalidades de los subconjuntos candidatos son mínimas, otros criterios son tomados en cuenta, como la interpretabilidad o la calidad predictiva.

De esta manera, una primera estrategia a tener en cuenta puede ser analizar los subconjuntos que contienen una mayor proporción de "descriptores frecuentes" (aquellos que están presentes en más de un subconjunto candidato). La razón que nos motivó para iniciar el análisis con la exploración de estos descriptores, es que su aparición en más de un subconjunto indicaría una mayor probabilidad de aportar información relevante para el modelado de la propiedad.

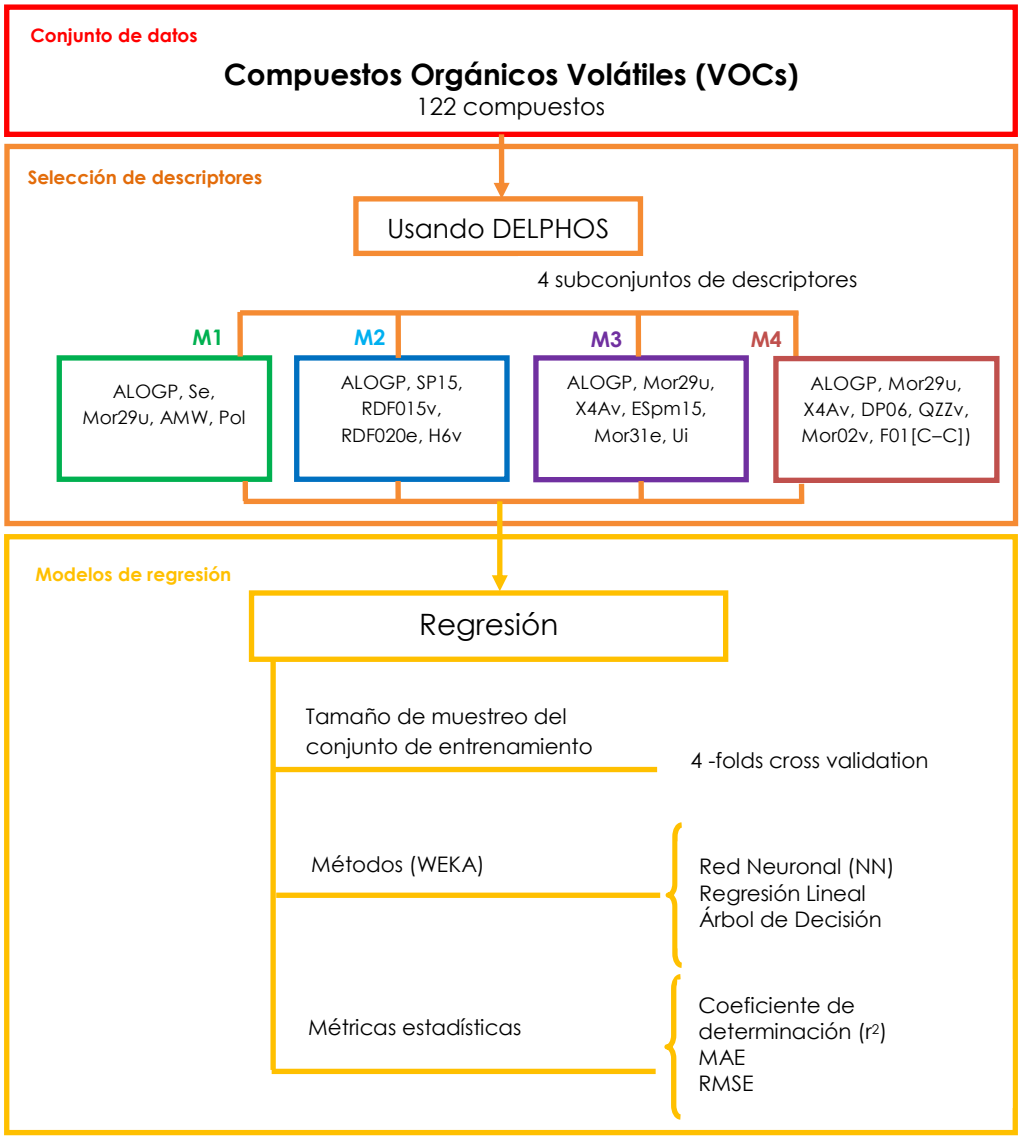
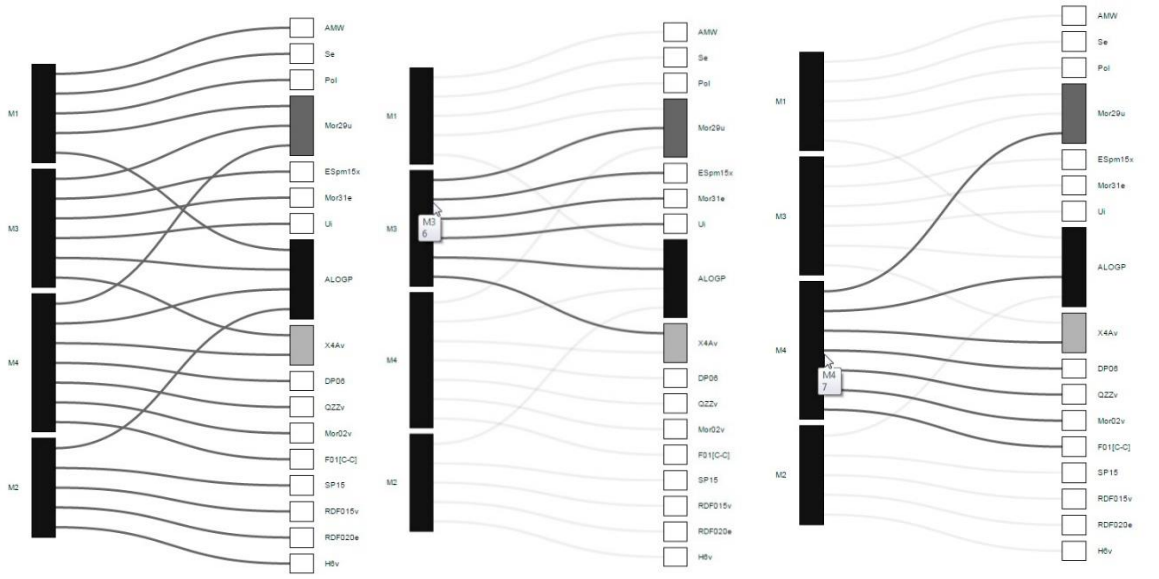


Figura 5.2 Metodología aplicada para la selección de descriptores y la generación de modelos de regresión para log P_{liver}.

Esta información puede ser fácilmente visualizada en el grafo bipartito de modelos y descriptores. La Figura 5.3 muestra que M3 y M4 son los modelos que contienen descriptores más frecuentes (tres), pero también mayor cardinalidad. De este modo, si consideramos cardinalidad y número de descriptores frecuentes, se puede concluir que M2 sería la opción menos promisoria.



a) **b)** **c)**

Figura 5.3 a) Relaciones entre modelos y descriptores. Los descriptores frecuentes corresponden a los nodos que están coloreados de un color gris más oscuro a negro. **b)** Visualización al pasar el puntero sobre M3. **c)** Visualización al pasar el puntero sobre M4.

Descartado M2, un paso siguiente puede consistir en analizar la información mutua entre los descriptores que están presentes en los modelos restantes. El objetivo en este caso es determinar qué modelos tienen mayor proporción de descriptores que aportan información no redundante para la predicción de la propiedad. La Figura 5.4 muestra esta información para M1, M3 y M4. Se puede observar claramente que M4 tiene mayor información mutua entre sus pares de descriptores, y por lo tanto será considerado como el modelo menos interesante debido a la alta redundancia entre sus descriptores.

Después de que M4 es descartado, el siguiente análisis se centra en decidir entre M1 y M3 usando las restantes visualizaciones. Un siguiente paso consiste en evaluar la co-ocurrencia de descriptores en estos modelos.

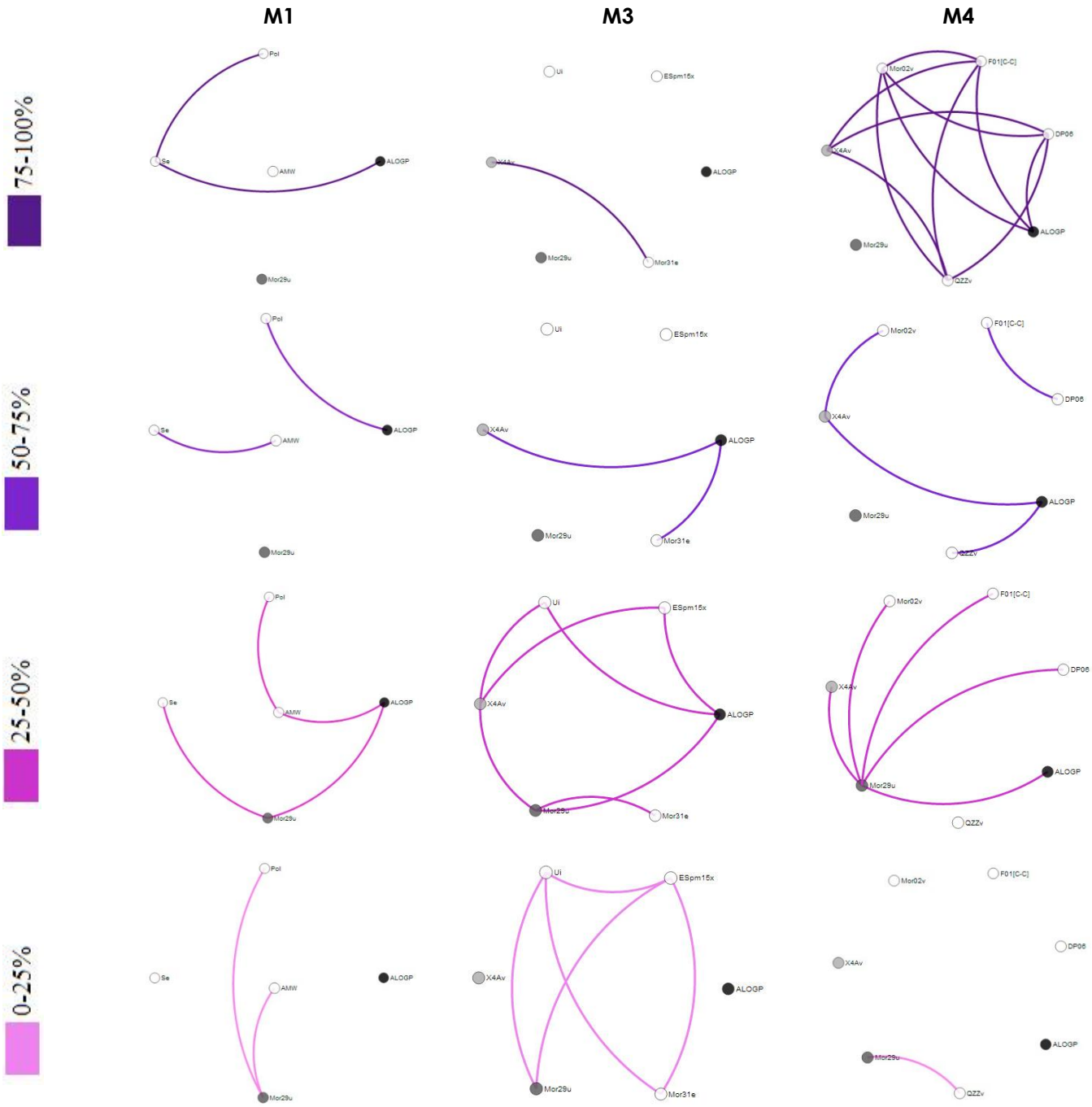


Figura 5.4 Enlaces con los cuatro niveles de información mutua (columnas) entre los descriptores de M1, M3 y M4 (filas). Este filtro puede ser obtenido siguiendo estos pasos: (1°) seleccionar cada modelo haciendo clic en el nodo correspondiente del grafo bipartito, (2°) teniendo seleccionado el modo basado en entropía, mover el umbral de enlaces del grafo G_p hacia la derecha hasta que se detenga, (3°) filtrar los enlaces haciendo doble clic sobre el rango de colores que se encuentra arriba del grafo.

Esto es con el fin de identificar las similitudes y diferencias entre los descriptores de los diferentes modelos. Esta información se muestra en la Figura 5.5, donde se puede observar una proporción similar de descriptores con un grado de co-ocurrencia medio y alto.

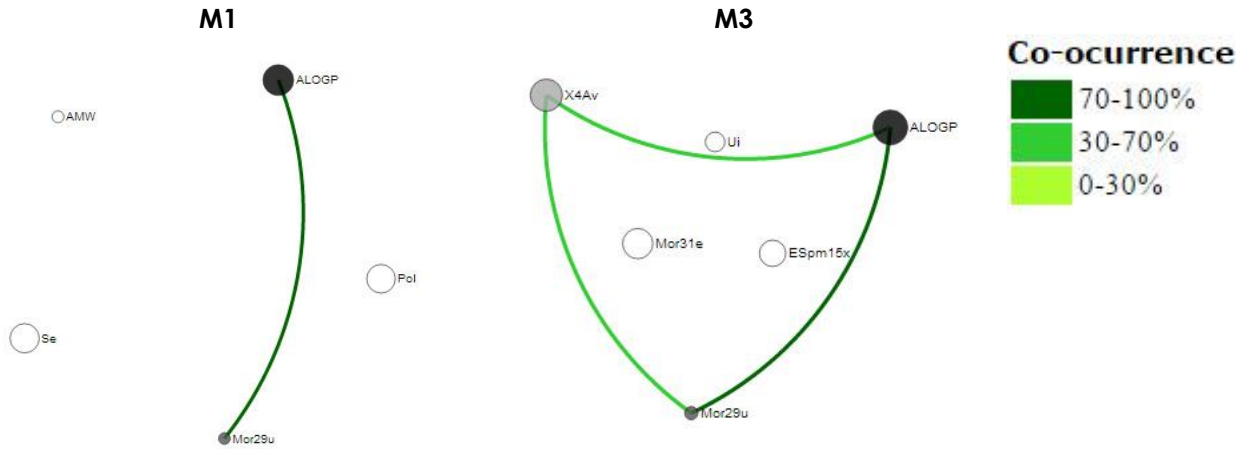


Figura 5.5 Grado de co-ocurrencia mediano y alto entre los descriptores de M1 y M3. Este filtrado se puede obtener mediante la selección de cada modelo haciendo clic en el nodo correspondiente en el grafo bipartito, y luego modificando el umbral de las aristas en el grafo G_s.

En este punto, una opción es utilizar los histogramas y gráficos de dispersión para analizar cuál de los dos modelos es más interpretable desde un punto de vista físico-químico. M1 consta de cinco descriptores: AMW (peso molecular dividido por el número de átomos), Se (suma de electro-negatividades atómicas de Sanderson), Pol (número de polaridad), Mor29u (representación molecular 3-D de estructuras basadas en difracción de electrones - no ponderado/señal 29) y ALOGP (coeficiente de partición octanol-agua de Ghose-Crippen). Todos ellos aportan, en un grado mayor o menor, información importante acerca de las propiedades moleculares relacionadas con la capacidad de la molécula para distribuirse entre los dos medios en estudio: el tejido del hígado y la sangre. En este sentido, un análisis por parte de los expertos sobre los descriptores de M1 (presentado en el capítulo 4), permitiría determinar que estos descriptores contemplan más aspectos físico-químicos relacionados con la propiedad bajo

estudio que M3. Por otra parte, M1 también tiene una cardinalidad menor que M3. Así, concluimos que la opción más acertada es seleccionar a M1 como el mejor modelo de los cuatro modelos candidatos, ya que reúne las características deseables de baja cardinalidad, buena calidad predictiva e interpretabilidad de los descriptores en términos físicos-químicos.

De esta manera, se ha presentado un posible flujo de análisis para elegir entre modelos alternativos, utilizando nuestra herramienta de analítica visual. Sin embargo, es importante destacar que esta estrategia no es el único enfoque a seguir para el análisis. Por ejemplo, un usuario podría iniciar el análisis evaluando las interacciones entre los descriptores de los modelos, a través de la inspección de los grafos no dirigidos. Si bien los distintos flujos de análisis pueden dar lugar a diferentes subconjuntos de descriptores, un aspecto importante de este tipo de selección es que aumenta la confianza del experto sobre el modelo.

5.3.1.2 Elección del mejor subconjunto de descriptores para clasificación

La motivación aquí fue tratar de determinar la potencial nocividad de un VOC en el organismo. Para esto, a través de distintos modelos QSPR evaluamos la afinidad del compuesto con un determinado medio (acuoso o grasoso). Más específicamente, lo que se busca es tener un primer modelo para la toxicidad de un VOC identificando cuáles de ellos son más afines a acumularse en el hígado. De esta manera, se pretende contribuir con el desarrollo de modelos PBPK (*Physiologically Based Pharmacokinetic*) para temas de salud pública, mediante el diseño de un modelo QSPR que permita clasificar si un VOC es más afín a un medio acuoso (sangre) o a un medio grasoso (hígado) (Cravero, Martínez, Díaz, y Ponzoni, 2017). La metodología y el diseño experimental llevados a cabo se ilustra en la Figura 5.6. La selección automática de los descriptores se realizó de la misma manera que para el caso de regresión.

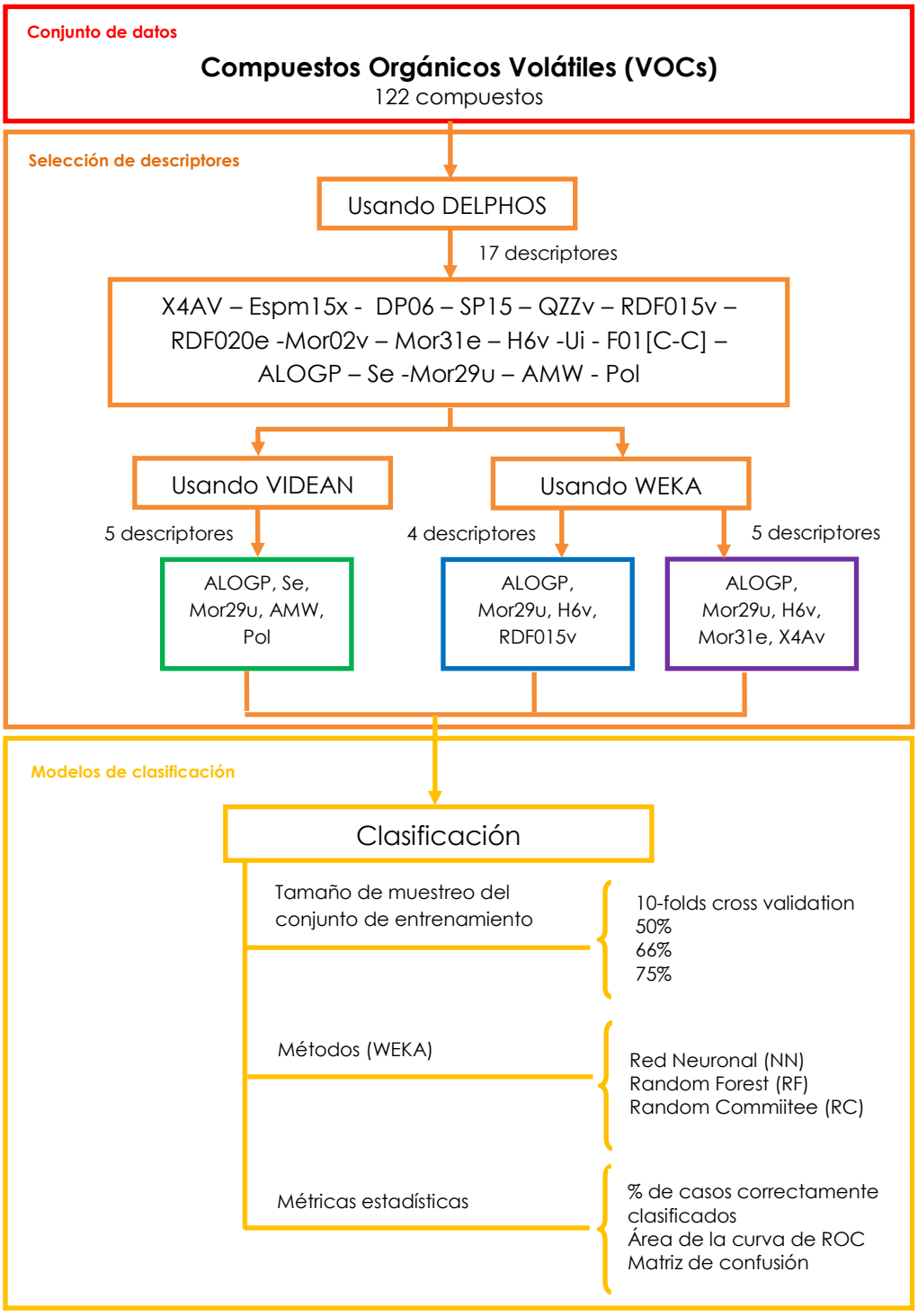


Figura 5.6 Metodología y diseño experimental llevado a cabo para obtener el modelo de clasificación para log P_{liver}

Luego a partir del conjunto de 17 descriptores resultantes, se realizó una segunda selección siguiendo dos enfoques: (i) utilizando VIDEAN, mediante la cual se obtuvo un subconjunto de 5 descriptores (Sub-A) (ii) usando WEKA, de la cual resultaron dos subconjuntos de 4 (Sub-B) y 5 (Sub-C) descriptores. Para obtener Sub-B fue utilizado un método wrapper y para obtener Sub-C se utilizó el método CfsSubset (Hall y Smith, 1998) como evaluadores de atributos. CfsSubset es un método que evalúa el valor de un subconjunto de características teniendo en cuenta la capacidad predictiva individual de cada característica junto con el grado de redundancia entre ellos. Siempre se prefiere subconjuntos de características que estén altamente correlacionados con la clase y que tengan baja intercorrelación (Hall et al., 2009). Luego, para la etapa de modelado, se aplicaron para cada uno de los tres subconjuntos, diferentes métodos: redes neuronales, bosque aleatorio (*random forest*) y comité aleatorio (*random committee*) (Frank et al., 2005), con distintos tamaños para el conjunto de entrenamiento. El modelo con más alto rendimiento que obtuvimos, clasifica correctamente el 72.13% de los VOCs y tiene un área de la Curva de Característica Operativa del Receptor (curva de ROC) de 0.83.

Definición de umbrales para la discretización de log P_{liver}

Proponemos clasificar el conjunto de datos en tres clases de VOCs: los que tienen afinidad con la sangre, los que tienen afinidad con el hígado y los que no presentan una preferencia por ninguno de los dos medios. Recordemos que la propiedad log P_{liver} es un coeficiente de partición, esto quiere decir que representa la relación entre la concentración del compuesto en cada medio, luego de alcanzar un equilibrio. Más específicamente, altos valores de log P_{liver} indican la afinidad del compuesto por un medio grasoso (hígado) mientras que valores bajos de la propiedad representan la afinidad por un medio acuoso (sangre).

$$\log P_{liver} = log \frac{[VOC\ en\ el\ hígado]}{[VOC\ en\ la\ sangre]} = \log \frac{afinidad\ con\ la\ sangre\ (medio\ grasoso)}{afindad\ con\ el\ hígado\ (medio\ acuoso)}$$

Una relación de concentraciones igual a 1 indica la misma afinidad por los medios. Luego consideramos el rango: -0.04 a +0.04 para log P_{liver} como "zona gris". Estos valores provienen de las siguientes ecuaciones:

$$\log P_{liver} = \log \frac{[1]}{[1.1]} = -0.04 \qquad \log P_{liver} = \log \frac{[1.1]}{[1]} = +0.04$$

A partir de esta zona, definimos los valores más bajos (<-0.04) como la clase "afinidad con la sangre" y los valores más altos (> +0.04) como la clase "afinidad con el hígado". En la Tabla 5.2, se puede observar la distribución del conjunto de datos para esta clasificación. Los resultados son coherentes con la naturaleza de los compuestos (VOCs) ya que se presentan valores de log P_{liver} altos para los VOCs no-polares, y viceversa para los polares. Se debe tener en cuenta que el 75% del conjunto de datos tiene una afinidad con el hígado; este es un resultado esperado ya que la mayoría de los VOCs tienen baja solubilidad en agua.

Clase	Umbral	Número de moléculas	% moléculas
Afinidad con la sangre	$(-\infty; 0.04]$	15	12.30%
Zona gris	$(-0.04; +0.04)$	15	12.30%
Afinidad con el hígado	$[+0.04; +\infty)$	92	75.40%

Tabla 5.2 Umbrales para la discretización de la propiedad objetivo y porcentaje de muestras para cada clase. Se definen tres clases de VOCs en términos de la afinidad al medio.

El modelo obtenido permite predecir el medio de afinidad (sangre o hígado) de un compuesto orgánico volátil (VOC) y puede aplicarse en el desarrollo de modelos fármaco-cinéticos basados en la fisiología (PBPK)(Sager et al., 2015). Hasta donde sabemos, es el primer modelo de clasificación QSAR derivado de log P_{liver} propuesto en la literatura. Además, se debe resaltar el aporte realizado en cuanto a la definición de las clases de afinidad a un determinado medio, que pueden ser utilizadas para otros estudios QSAR relacionados con la propiedad.

5.3.1.3 Conclusiones

Se ha estudiado la propiedad log P_{liver} para los compuestos orgánicos volátiles (VOCs). En primera instancia se ha presentado un modelo de regresión cuyos descriptores han sido obtenidos mediante una combinación de selección automática e intervención del experto con el soporte de VIDEAN. Se ha detallado un posible flujo de análisis utilizando nuestra herramienta logrando concluir con un modelo que presenta buen rendimiento y es interpretable en términos físico-químicos.

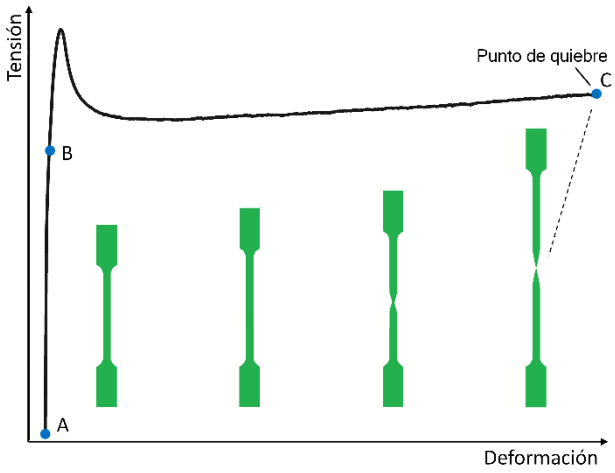
Por otro lado, se presentó un modelo de clasificación con el fin de evaluar la potencial toxicidad de un VOC en el organismo. En este sentido el modelo clasifica si un compuesto es más afín al hígado o a la sangre. A partir de esto podemos analizar aquellos que son más propensos a acumularse en el hígado y por ende ser más perjudiciales para la salud. Estos modelos contribuyen a brindar un panorama de cómo se distribuyen estos tipos de compuestos en el organismo y pueden emplearse como herramienta para evaluar riesgos y tomar decisiones en materia de salud pública.

5.3.2 Diseño de Materiales

En esta sección se presentan las experimentaciones realizadas con tres propiedades mecánicas de los polímeros derivadas del ensayo de tensión: elongación a la rotura, resistencia a la rotura y módulo elástico, que brindan información acerca de la ductilidad, resistencia y rigidez de un material, respectivamente. Los resultados de esta caracterización son comúnmente utilizados para seleccionar un material para una aplicación específica o control de calidad (Van Krevelen, 2009).

A continuación, se explican brevemente los conceptos básicos del test. En la Figura 5.7 se puede ver el equipo utilizado (Maquina universal de ensayos mecánicos marca INSTRON©) y la curva de tensión-deformación típica de un polímero sintético (no elastómero). En el comienzo de la curva (tramo A-B), el comportamiento del material se considera elástico y en esta zona se calcula el *módulo elástico*. Se trata del cociente entre la tensión aplicada y la deformación y es un indicador de la rigidez de un material. Luego, el comportamiento se torna plástico y se pueden calcular varias propiedades de interés. Específicamente, en esta tesis se ha estudiado el punto de rotura (punto C de la Figura 5.7), de donde se desprenden los valores de resistencia a la rotura y elongación a la rotura. En el ensayo de tensión, a medida que se incrementa la velocidad transversal, el módulo y la resistencia a la rotura aumentan mientras que la elongación a la rotura generalmente disminuye (excepto en elastómeros).

a)



b)

Figura 5.7 a) Máquina universal de ensayos mecánicos INSTRON© y b) Curva derivada del ensayo de tensión de un polímero sintético no elastómero.

A velocidades de deformación lentas las moléculas tienen la posibilidad de sufrir cambios de conformación u orientarse en la dirección del esfuerzo aplicado, de modo que presentan mayor ductilidad y elongación a la rotura que a velocidades altas. Cuando los ensayos de los distintos materiales son realizados a diferentes velocidades, se hace imprescindible para el modelado de estas propiedades incorporar este parámetro (CHS, del inglés crosshead speed) como descriptor, ya que los resultados son altamente dependientes del mismo. A su vez, cuando se construyen las bases de datos, se pretende que los valores sean derivados de testeos bajo normas (por ejemplo, ASTM), y que estén llevados a cabo en el mismo rango de temperatura, que es otra variable que afecta el comportamiento de un polímero.

El estudio de las propiedades mecánicas de materiales poliméricos es de gran importancia, debido a que dependiendo de su aplicación se intentará determinar un perfil estructural que va a estar sujeto a las características de estas propiedades. En relación a esto, se presentarán en las siguientes secciones diferentes modelos que permiten relacionar de manera cuantitativa diferentes descriptores de una entidad química (compuesto) con una determinada propiedad.

5.3.2.1 Análisis de la propiedad *elongación a la rotura*

En la industria de los materiales poliméricos es fundamental definir el perfil de aplicación, para cual se realiza una completa caracterización fisicoquímica. Hay numerosas propiedades que pueden describir este perfil, y entre ellas, las mecánicas ocupan un lugar de preponderancia. Por ejemplo, la capacidad de resistir la ruptura bajo una tensión de tracción es una de las propiedades de los materiales más importantes y ampliamente utilizada en aplicaciones estructurales (Ward y Sweeney, 2012). En esta sección se explora una propiedad obtenida a partir de un ensayo de tensión: elongación a la rotura (*elongation at break*), que es una medida de la ductilidad del material. Como se mencionó antes, es

importante tener en cuenta que no sólo la temperatura ambiente afecta el valor final de la elongación a la rotura, sino que la velocidad transversal (CHS) del ensayo es un parámetro que modifica en gran medida los resultados (Callister y Rethwisch, 2011).

En esta sección detallaremos la experimentación llevada a cabo para obtener modelos tanto de regresión como de clasificación para la elongación a la rotura. Para tal fin, se utilizó un conjunto de datos de los polímeros de alto peso molecular extraídos de Palomba et al. (2014), que incluye 655 descriptores moleculares procedentes de 77 polímeros del tipo: resinas puras, amorfos, lineales, no reticulados, y no elastómeros. Para ambos tipos de modelo, se ilustra la situación en el que el analista quiere hacer una intervención en la selección automática de descriptores con el fin de incorporar un parámetro experimental de relevancia para el modelo: la velocidad transversal de ensayo (CHS). Una observación a tener en cuenta, es que este parámetro aporta información clave a los modelos ya que no todos los ensayos de tensión realizados sobre los polímeros de la base de datos se hicieron a la misma velocidad. De este modo, demostramos cómo el experto puede utilizar nuestra herramienta para analizar la información procedente de subconjuntos de descriptores existentes e incluso incorporar nuevos descriptores de acuerdo a su propio conocimiento de la propiedad de destino.

5.3.2.1.1 Diseño de un modelo de regresión incorporando CHS

En este caso particular, el objetivo es predecir valores de la *elongación a la rotura* por medio de un modelo de regresión que tenga un alto rendimiento, buena interpretabilidad y que cumpla el requisito especial de incluir CHS (velocidad de crosshead) como un descriptor experimental debido a su alta influencia en el

valor de la propiedad (Martínez et al., 2015). Este caso de estudio se encuentra disponible en la página web de VIDEAN[6].

La experimentación se esquematiza en la Figura 5.8. El análisis se inició a partir de los mejores diez subconjuntos obtenidos automáticamente por DELPHOS (Tabla 5.3), a los cuáles se les aplicó tres métodos de aprendizaje automática: regresiones lineales, árboles de decisión y redes neuronales efectuando una validación cruzada de 4-folds, con los parámetros por defecto que ofrece WEKA.

Modelo	Calidad predictive	Cardinalidad
M1 (Mn/MW, Sp, RHyDp, ETA_EtaP_F_L)	r^2 = 0.26 MAE = 4.62 RMSE = 8.14	4
M2 (Mn/MW, MDEO-11, D/Dr09, SMTIV)	r^2 = 0.32 MAE = 5.94 RMSE = 8.31	4
M3 (Mn/MW, nHBint4, nHBint10, ETA_dEpsilon_B)	r^2 = 0.56 MAE = 4.03 RMSE = 6.22	4
M4 (Mn/MW, nsCH3, nF6Ring, ALOGP2, RDCHI)	r^2 = 0.41 MAE = 3.94 RMSE = 6.75	5
M5 (Mn/MW, nROH, n6Ring, nHCsatu, ALOGP2)	r^2 = 0.68 MAE = 3.28 RMSE = 5.78	5
M6 (Mn/MW,nP, minHBa, T(O..P), ETA_Epsilon_3)	r^2 = 0.25 MAE = 4.48 RMSE = 7.20	5
M7 (Mn/MW, ETA_dEpsilon_B, C-005, SHaaCH, nHBint9, nCt)	r^2 = 0.31 MAE = 4.19 RMSE = 7.20	6
M8 (Mn/MW, ndssC, minHBint9, MSD, C-004, Mw/Mn (PDI), crosshead speed(CHS))	r^2 = 0.39 MAE = 3.92 RMSE = 6.86	7
M9 (Mn/MW, Pol, Wap, maxHAvin, nHAvin, MWC04)	r^2 = 0.15 MAE = 4.92 RMSE = 7.88	6
M10 (Mn/MW, maxHBint6, ETA_dEpsilon_A, TIC2, ndO, nHdCH2)	r^2 = 0.48 MAE = 4.02 RMSE = 7.09	6

Tabla 5.3 Calidad predictiva y cardinalidad de los diez mejores modelos obtenidos por DELPHOS.

[6] http://lidecc.cs.uns.edu.ar/VIDEAN/, accediendo a la pestaña "Uses Cases" y luego cargando "Example 2".

Como una primera estrategia, y utilizando VIDEAN, se exploraron los "descriptores frecuentes" (aquellos que en el grafo bipartito están coloreados en una escala de grises más oscura e indican que han sido elegidos por más de un modelo). De este análisis se puede observar que el descriptor Mn/MW fue elegido por todos los modelos, y ETA_dEpsilon_B al igual que ALOGP2 fueron seleccionados en dos modelos cada uno. Entonces, el modelador se enfrenta con el análisis de si estos descriptores proporcionan información valiosa para describir la propiedad objetivo desde un punto de vista físico-químico. Esto merece una breve discusión. Mn/MW es un descriptor que relaciona el peso molecular promedio en número (Mn) con el peso molecular del monómero del polímero, dando como resultado el número de unidades repetitivas promedio en una cadena de polímero. En otras palabras, Mn/MW proporciona información macro del material. Por otro lado, ETA_dEpsilon_B es un descriptor de átomo de topoquímico extendido, lo que representa una medida de la contribución de insaturación y que se calcula en el monómero. ALOGP2 es el cuadrado de ALogP (coeficiente de partición de Ghose-Crippen). Estos coeficientes de partición representan una medida de la hidrofobicidad del monómero. ETA_dEpsilon_B y ALOGP2 proporcionan información a nivel micro porque se calcularon en un monómero y no en una molécula polimérica promedio (imposible de calcular). Por lo tanto, los "descriptores frecuentes" están representando los aspectos macro y micro de las moléculas del conjunto de datos y todos ellos afectan a la propiedad en estudio: *elongación a la rotura*.

Una posible pregunta para el modelador podría ser cuán independiente entre sí es la información proporcionada por cada descriptor. Para esto se analizó la información mutua con el objetivo de construir un modelo cuyos descriptores son independientes en términos de la información numérica que proporcionan, evitando así la redundancia.

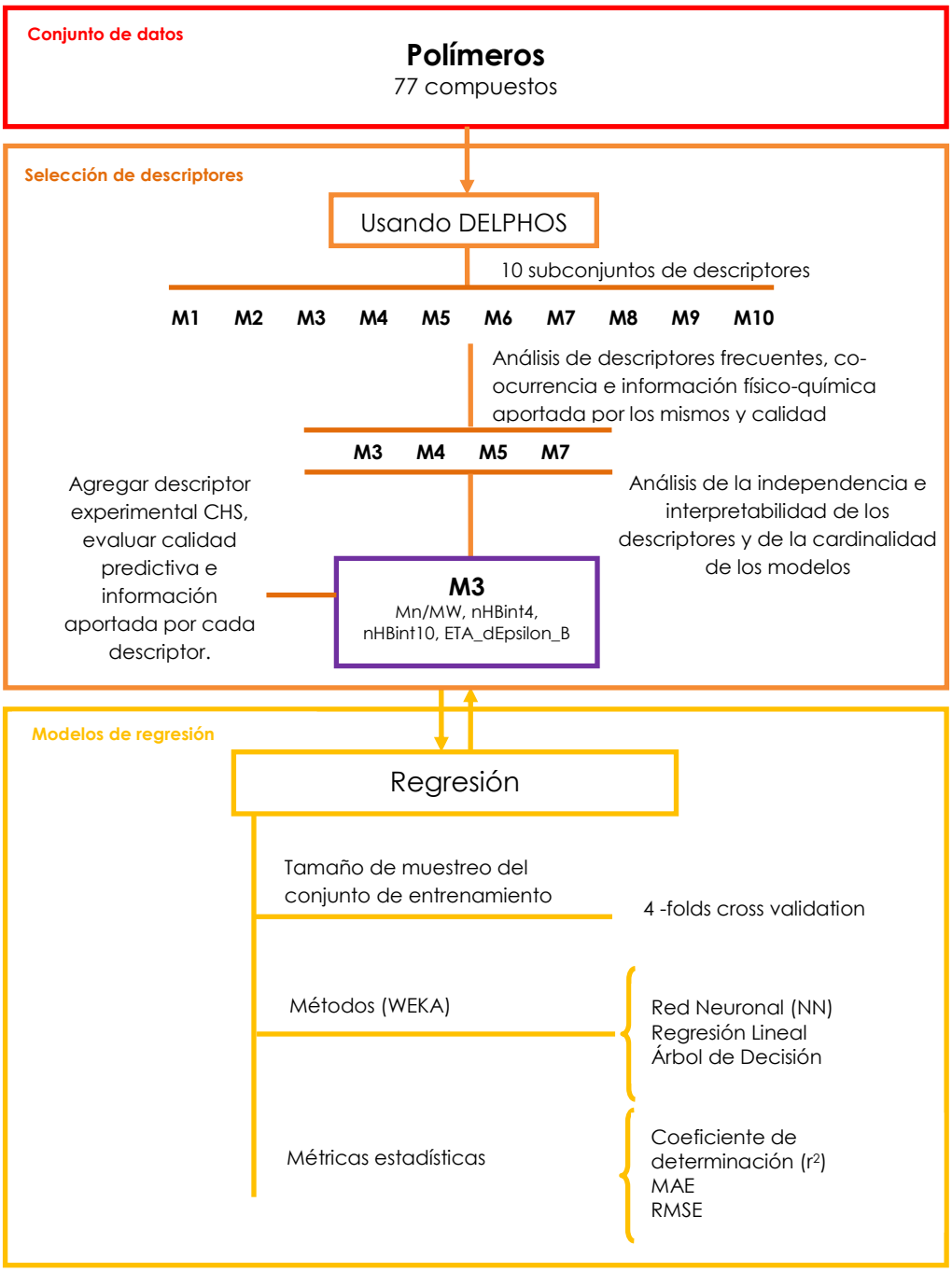


Figura 5.8 Esquematización de la experimentación realizada utilizando VIDEAN para arribar a un modelo de regresión para la *elongación a la rotura*.

La co-ocurrencia de estos descriptores "frecuentes" también se analizó para encontrar alguna información complementaria entre ellos. De este análisis se pudo observar que dos pares de descriptores aparecen dos veces: Mn/MW - ETA_dEpsilon_B y Mn/MW - ALOGP2, pero ETA_ dEpsilon_B y ALOGP2 no aparecen juntos en ningún modelo.

Después de esta primera etapa de análisis, se eligieron los cuatro modelos que contienen "descriptores frecuentes": M3, M4, M5 y M7 (Figura 5.5), ya que proporcionan información valiosa para la descripción de la propiedad. La calidad de los cuatro subconjuntos seleccionados se evaluó teniendo en cuenta la independencia, cardinalidad e interpretación.

En la Figura 5.9, se puede observar que el subconjunto que mejor cumple con todos estos factores es M3. Este subconjunto se compone de: Mn/MW, ETA_dEpsilon_B, nHBint4 y nHBint10. Los dos últimos son descriptores electro-topológicos que representan aspectos estructurales y químicos de los monómeros.

Teniendo en cuenta el análisis desde un punto de vista de predicción, es importante destacar que la mejor precisión de predicción pertenece a M3 (r^2 = 0.56) y M5 (r^2 = 0.68). Sin embargo, M5 tiene mayor cardinalidad (5) que M3 (4) y dos de los descriptores de M5 (nROH y nHCsatu) casi no tienen variación con respecto a la propiedad objetivo. Esta falta de varianza no es deseable, ya que significa que la información numérica proporcionada es muy limitada.

Posteriormente, el modelador puede querer modificar M3 agregando descriptores adicionales. En este caso, se considera esencial el descriptor experimental CHS para describir la propiedad ya que, para el conjunto de datos, los valores de *elongación a la rotura* se midieron a diferentes velocidades de crosshead. El subconjunto resultante contiene cinco descriptores (M3 + CHS) y el siguiente paso es analizar qué tan bueno es estadísticamente, con el fin de comprobar si la hipótesis del modelador mejora la calidad de la predicción.

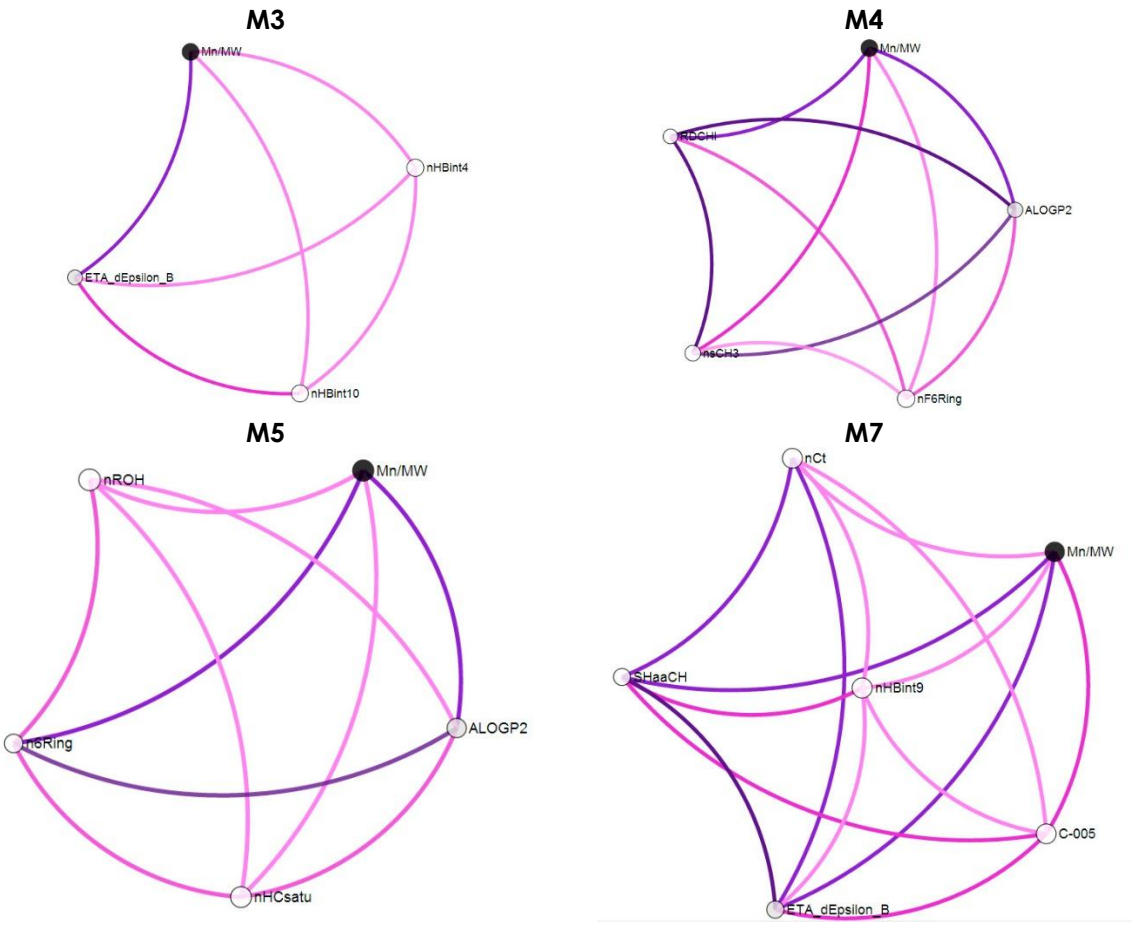


Figura 5.9 Información mutua (alta, media y baja) entre los descriptores de modelos 3, 4, 5 y 7.

En la Tabla 5.4, se puede observar que los valores estadísticos confirman la sugerencia del experto, y por lo tanto el rendimiento del nuevo modelo (r^2: 0,62) supera a la original (r^2: 0,56). Con el fin de mejorar aún más el estudio se analizaron los diagramas de dispersión provistos en VIDEAN, que muestran la variación de los valores del descriptor con respecto a la *elongación a la rotura*. De este análisis, rápidamente se pudo detectar la poca variación del descriptor nHBint4 (Figura 5.10). Este comportamiento motivó la eliminación de nHBint4 del modelo.

Modelo	Calidad predictiva
M3 (Mn/MW, nHBint4, nHBint10, ETA_dEpsilon_B)	r^2 = 0.56 MAE = 4.03 RMSE = 6.22
M3 + CHS (Mn/MW, nHBint4, nHBint10, ETA_dEpsilon_B,CHS)	r^2 = 0.62 MAE = 3.43 RMSE = 5.89
M3 + CHS - nHBint4 (Mn/MW, nHBint10, ETA_dEpsilon_B,CHS)	r^2 = 0.69 MAE = 3.24 RMSE = 5.68

Tabla 5.4 Capacidad predictiva de los modelos M3, (M3 + CHS) y (M3 + CHS − nHBint4). La segunda columna muestra la calidad predictiva del "mejor" modelo después de aplicar validación cruzada de 4-folds con tres métodos diferentes (regresión lineal, árboles de decisión y redes neuronales). En este caso, la mejor exactitud de predicción para los tres modelos se obtuvo mediante el uso de un árbol de decisión (M5P).

En la Tabla 5.4, se puede ver la precisión de predicción del nuevo modelo mejorado (r^2: 0,69). Además, analizando el resto de los diagramas de dispersión, se ve que todos los descriptores varían de manera diferente con la propiedad, lo que es otra indicación de que probablemente proporcionen información independiente.

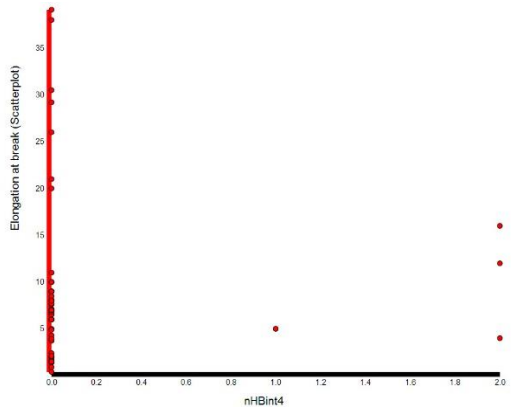


Figura 5.10 Diagrama de dispersión entre el descriptor nHBint4 y la propiedad. Puede observarse la poca variación.

En resumen, este ejemplo ilustra un posible flujo de análisis para generar un modelo sólido para la predicción de la *elongación a la rotura* de polímeros mediante el uso de VIDEAN.

El modelo final (M3 + CHS - nHBint4) contiene cuatro descriptores que proporcionan información significativa e independiente: Mn/MW, ETA_dEpsilon_B, nHBint10 y CHS. Mn/MW representa el número de unidades repetitivas de moléculas poliméricas promedio (macro), nHBint10 y ETA_dEpsilon_B representan las propiedades estructurales a nivel micro y finalmente CHS proporciona información experimental de los test de medición de la propiedad, la cual afectó la capacidad predictiva del modelo notablemente. Por lo tanto, el modelo presentado aquí, considera flexibilidad molecular, interacciones intermoleculares y la tasa de testo, lo que hace que sea un subconjunto confiable e interpretable.

5.3.2.1.2 Diseño de un modelo de clasificación incorporando CHS

En esta instancia, resulta de interés plantear modelos de clasificación en términos de la ductilidad de un material en fases tempranas de diseño previo a la síntesis (Martínez et al., 2017). Esto representa una valiosa herramienta, debido a la demanda de materiales poliméricos con requisitos específicos y dado que, como hemos mencionado en otras oportunidades, el estudio del perfil mecánico de un polímero ayuda a definir su campo de aplicación. De esta manera, presentaremos modelos de clasificación QSPR para la caracterización de ductilidad de materiales poliméricos, utilizando la información proporcionada por experimentos de ensayos de tensión.

La Figura 5.11 esquematiza la metodología llevada a cabo. De la misma manera que en las experimentaciones anteriores, una vez realizada la selección automática de descriptores, se prosiguió a realizar una segunda selección de la siguiente manera: (i) con WEKA, obteniendo un subconjunto de 11 descriptores

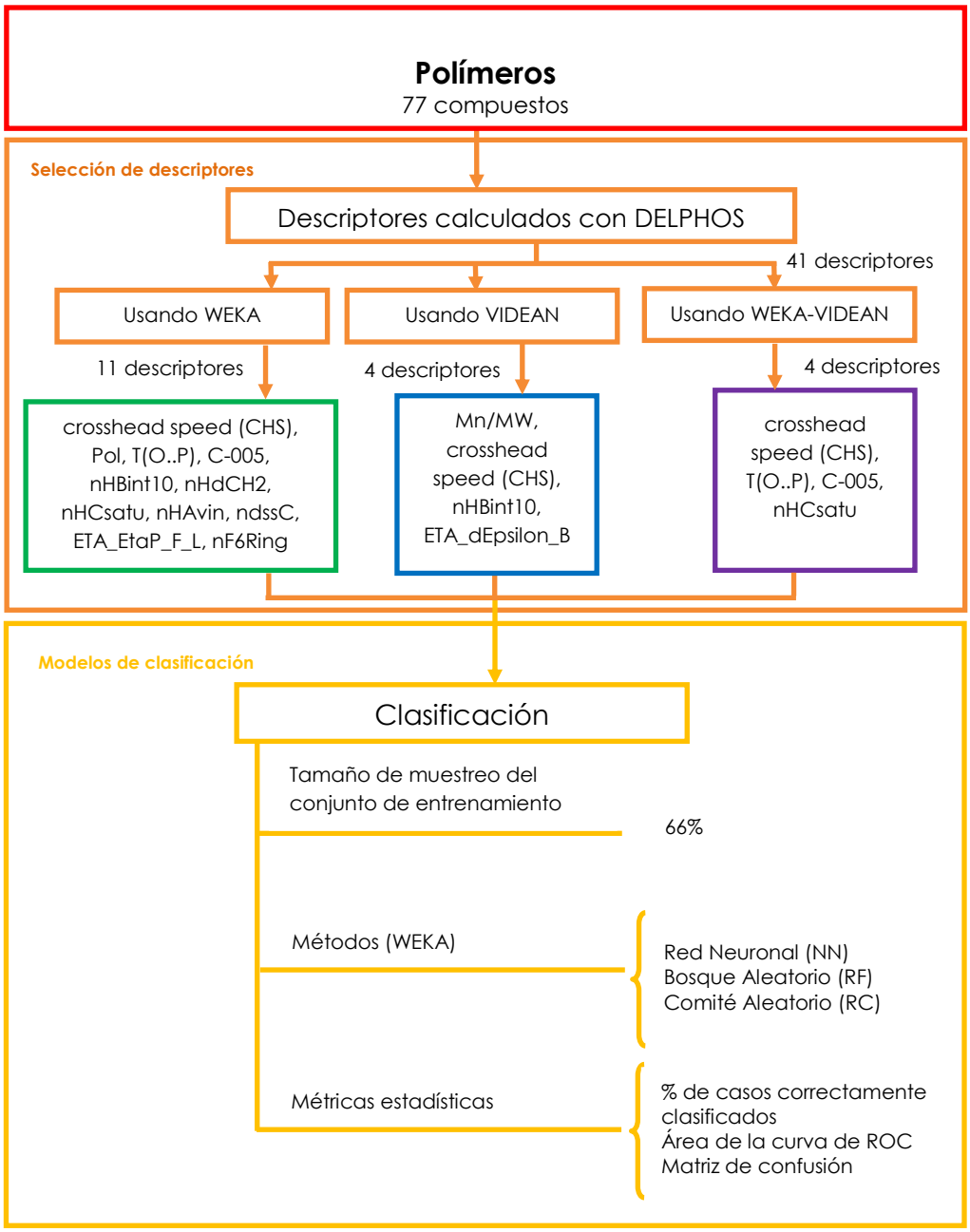


Figura 5.11 Esquema del proceso experimental para obtener el modelo de clasificación para *elongación a la rotura*.

(Sub-1); (ii) con VIDEAN, de la cual se obtuvo un conjunto de 4 descriptores y (iii) con una combinación de WEKA y VIDEAN, teniendo como resultado otro subconjunto de 4 descriptores. En todos estos subconjuntos se encuentra la propiedad CHS, ampliamente analizada en la sección anterior, ya que tiene un papel fundamental en la predicción de la *elongación a la rotura*.

Se infirieron distintos modelos QSPR, utilizando tres técnicas, Red Neuronal, Bosque Aleatorio y Comité Aleatorio, con distintas particiones para el conjunto de entrenamiento. El modelo con el más alto rendimiento clasifica correctamente el 88.46% de los polímeros y tiene un área de la curva de ROC de 0.97. Este modelo obtenido puede predecir con un alto nivel de confianza, si un material será dúctil o no en fases tempranas de diseño del polímero, previo a realizarse su síntesis.

5.3.2.2 Análisis de la propiedad *resistencia a la rotura* con VIDEAN

En esta sección se estudia la *resistencia a la rotura* utilizando la combinación de técnicas CODES-TSAR para extracción de características. Si bien esta estrategia ha sido aplicada en el estudio de compuestos químicos de interés para el diseño de drogas con un buen desempeño (Guerra et al., 2008), su efectividad nunca había sido evaluada en el campo de los materiales poliméricos. El objetivo aquí fue explorar las fortalezas y limitaciones de la metodología propuesta en la inferencia de propiedades mecánicas. En particular, se buscó realizar una comparación de modelos QSPR obtenidos por CODES-TSAR con un modelo QSPR formado por descriptores elegidos a través de técnicas de selección de características (Cravero et al., 2016). Por otro lado, teniendo en cuenta que CODES-TSAR sólo captura la información estructural 2D de los compuestos, la combinación de estas variables aprendidas con los descriptores moleculares 3D obtenidos a través de métodos de selección de característica podría llevar a obtener modelos QSPR con alto rendimiento predictivo. Es por esto que se decidió

evaluar el impacto de aplicar la hibridación de los dos tipos de estrategias (selección y aprendizaje de características).

El conjunto de datos utilizado tiene 66 polímeros (Palomba et al., 2014). Para el análisis se han tomado como base cinco subconjuntos de descriptores con los que se han experimentado en estudios y reportes previos (Cravero et al., 2015) y que fueron obtenidos aplicando los dos tipos de estrategias mencionadas anteriormente. El esquema de la metodología aplicada se ilustra en la Figura 5.12. De los cinco subconjuntos de descriptores, uno es representativo del método de selección de características (MR) y otros dos son representativos de la técnica de extracción de características (CT-N2 y CT-N3). Los últimos dos nacen de la combinación de MR con CT-N2 y CT-N3. El rendimiento de los cinco modelos puede verse en la Tabla 5.5. La evaluación estadística fue realizada con WEKA, utilizando árboles de decisión y realizando una validación cruzada de 10 folds.

Modelo	Cardinalidad	r^2
CT-N2	2	0.5788
CT-N3	3	0.4385
MR: Mn + CHS + Eta_dEpsilon_D + M_{MC}/M_{SC}	4	0.8172
C-N2: CT-N2 + MR	6	0.8488
C-N3: CT-N3 + MR	7	0.8514

Tabla 5.5 Para cada modelo se reporta su cardinalidad y el coeficiente de determinación para la validación.

Los descriptores del modelo de referencia (MR) fueron seleccionados utilizando DELPHOS en conjunto con un análisis ad hoc por parte del experto. En el siguiente apartado se detallará el proceso realizado para obtener los subconjuntos CT-N2 y CT-N3. A su vez, los compararemos en rendimiento con MR.

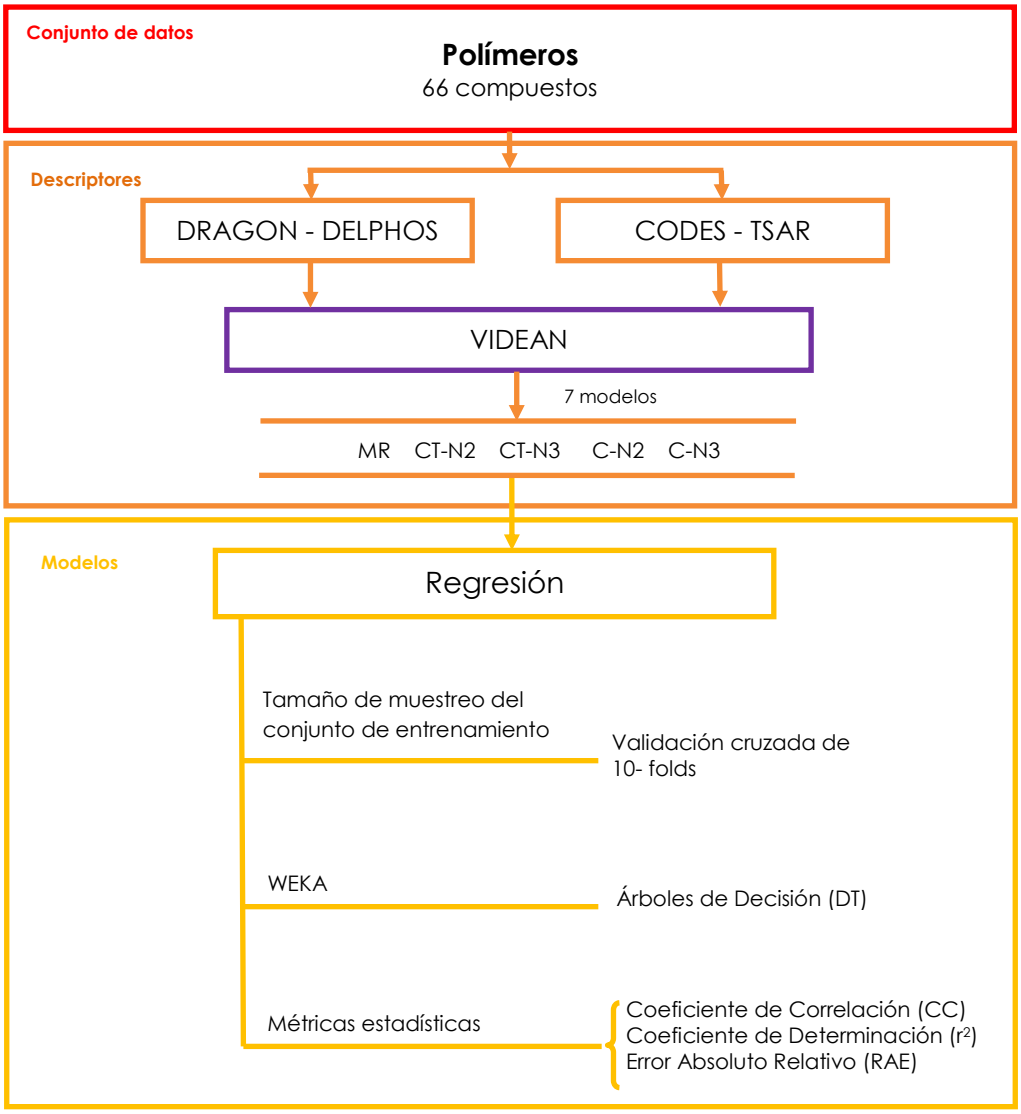


Figura 5.12 Esquema de la metodología aplicada para el estudio de la *resistencia a la rotura*.

5.3.2.2.1 Contrastación de enfoques de selección y aprendizaje de descriptores

Utilizando la herramienta CODES se procesó cada estructura de molécula contenida en el conjunto de datos, obteniendo una matriz dinámica. Esta matriz es la entrada que utiliza el software TSAR, que finalmente calcula los descriptores moleculares para cada compuesto. Los valores generados por CODES-TSAR tomados en conjunto, representan un descriptor determinado de la estructura 2D.

En la tabla 5.5 se puede observar que el rendimiento de los modelos CT-N2 y CT-N3 que contienen sólo descriptores creados por CODES-TSAR, está bastante por debajo del rendimiento de MR.

Los modelos fueron analizados en VIDEAN. En primera instancia utilizando los gráficos de dispersión e histogramas de los valores del descriptor y la propiedad. De este análisis se pudo observar que CT-N3 mostraba el mismo tipo de comportamiento para los tres descriptores respecto de la propiedad (Figura 5.13), puede deducirse que todos los descriptores del modelo están aportando información en la misma zona y debido a esto es el bajo rendimiento del modelo (0.4385). Un comportamiento similar se concluyó para CT-N2. Para el caso de MR, pudo observarse (Figura 5.13) que los cuatro descriptores del modelo presentaban comportamientos diferentes respecto de la propiedad, pero con una pérdida de información en la zona más hacia la derecha de cada histograma. A continuación, se analizarán los modelos que surgieron de la combinación de MR con CT-N2 y CT-N3.

5.3.2.2.2 Enfoque híbrido de selección y aprendizaje de descriptores

Aquí analizaremos los modelos combinados (C-N2 y CN-3), que surgieron a partir de la hibridación de los descriptores obtenidos mediante ambas técnicas: selección y aprendizaje de características.

Regresando a la Tabla 5.5, se puede observar que ambos modelos combinados aumentan el rendimiento con respecto a cada modelo por separado. Del análisis realizado anteriormente de la Figura 5.13, podemos deducir que el buen rendimiento de los modelos combinados se debe a que todos los descriptores juntos (MR + CT) están contribuyendo con información de diferentes zonas, completando así todo el espacio de la propiedad.

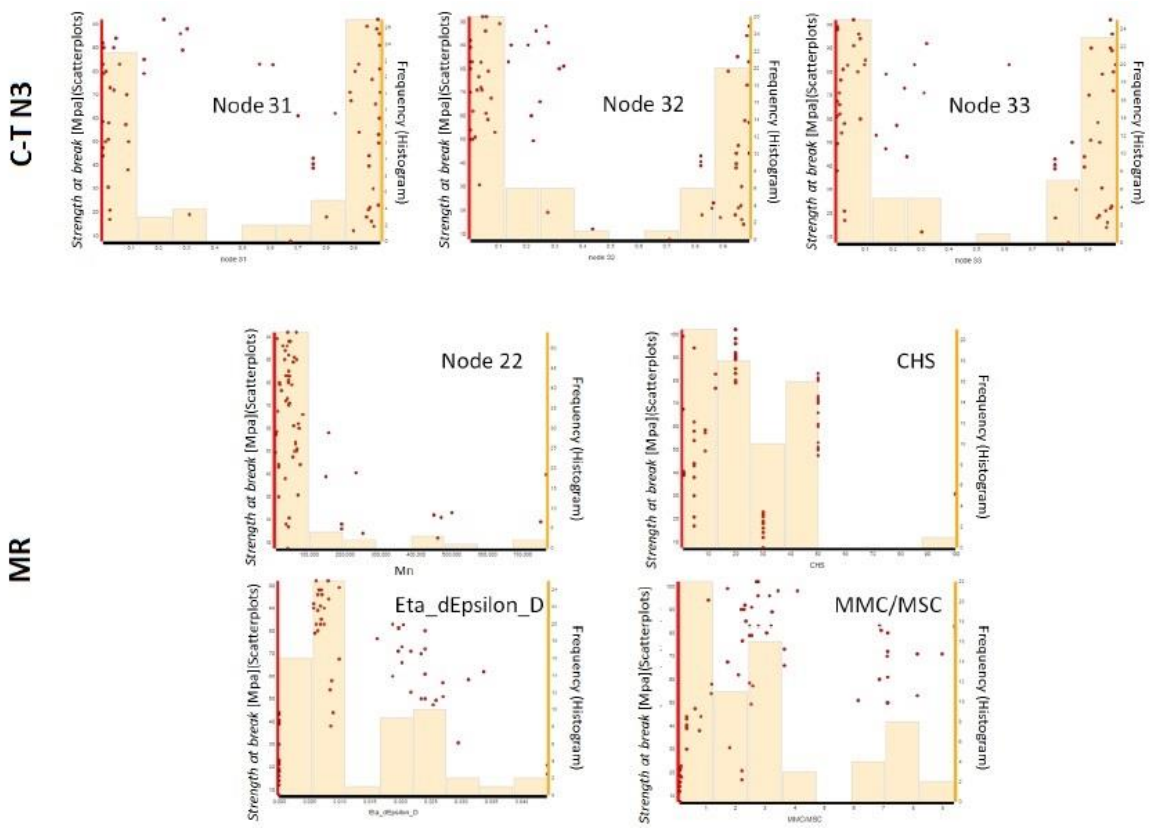


Figura 5.13 Gráficos de dispersión e histogramas para CT-N3 y RM que muestran los valores de los descriptores versus los valores de la propiedad.

Adicionalmente, se analizaron los grafos de correlación entre los descriptores en términos de la información mutua (IM). En la Figura 5.14 se puede observar las aristas en una escala de tonos de rosa a violeta representando los grados de IM, rosa claro cuando los descriptores tienen poca IM y violeta en caso contrario. En este sentido, es deseable que los descriptores tengan poca IM entre ellos para que no aporten información redundante al modelo. En la figura se puede observar que casi todos los descriptores tienen un alto grado de IM excepto Mn (todas sus aristas son de color rosa claro). Esto es coherente, debido a que Mn es un descriptor muy significativo para los polímeros, ya que estos materiales tienen la característica de tener una distribución de peso molecular en lugar de tener un peso molecular único. Mn representa el peso molecular promedio en número y

proporciona información macromolecular al modelo. Cuando Mn fue removido del modelo, el rendimiento en la predicción se deterioró.

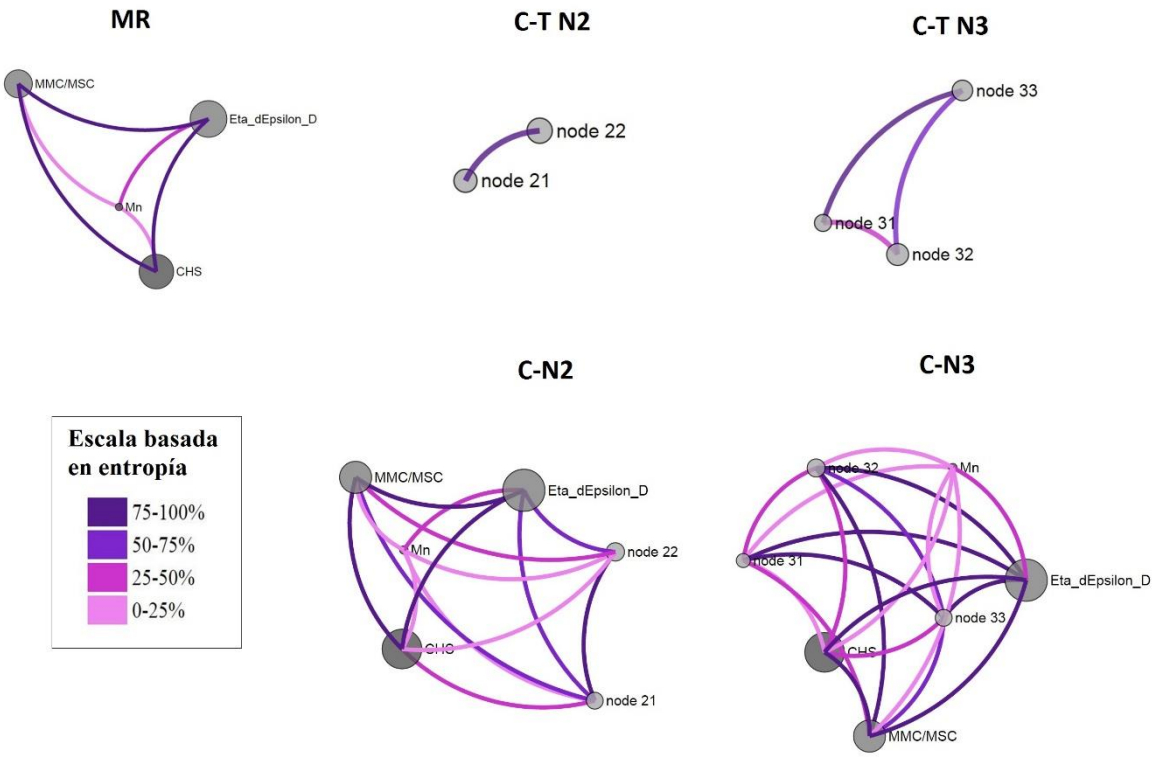


Figura 5.14 Información mutua entre los descriptores de cada modelo

Finalmente, podemos concluir que el uso exclusivo del método CODES-TSAR sobre esta base de datos, no ha sido suficiente para modelar y predecir la *resistencia a la rotura*. A pesar de esto, los descriptores aprendidos mediante esta técnica de extracción de características pueden contribuir con información a los modelos QSPR que se han generado usando técnicas de selección de características. De este modo, se han reportado modelos QSAR para predecir esta propiedad mecánica, observando que los conjuntos de descriptores obtenidos por ambas técnicas proporcionan información complementaria y relevante para la inferencia de la propiedad objetivo.

5.3.2.3 Análisis de la propiedad *módulo elástico*

Siguiendo en mente con nuestro objetivo de desarrollar modelos computacionales que permitan obtener rápidamente, y sin necesidad de sintetizar al compuesto valores estimados de la propiedad, aquí presentaremos un modelo de regresión para la propiedad *módulo elástico* obtenido a través de VIDEAN (Cravero et al., 2015). En primera instancia, partiendo de la base de datos extraída de Palomba et al. (2014), se calcularon descriptores utilizando HyperChem y DRAGON. Luego, se realizó una selección de descriptores a través de DELPHOS. De este procedimiento, se obtuvieron 10 subconjuntos de descriptores que posteriormente fueron analizados con VIDEAN. En esta instancia, se procedió a evaluar distintas combinaciones de descriptores interviniendo distintos modelos, para luego evaluarlos y llegar a obtener el modelo final.

En la Figura 5.15 puede observarse la conformación de los 10 modelos generados con DELPHOS (M1-M10) más un modelo M11 tomado como referencia de un trabajo previo (Palomba et al., 2014) en el que se estudió el *módulo elástico*. En ese trabajo, el modelo fue obtenido utilizando herramientas computacionales para cálculo y selección de descriptores y realizando el análisis de forma manual. Aquí proponemos mejorar el rendimiento de ese modelo, sin perder interpretabilidad, mediante el soporte de VIDEAN. Al igual que para las otras dos propiedades mecánicas estudiadas en las secciones anteriores, el parámetro experimental CHS debe incorporarse al modelo final.

Comenzando con el estudio, como primer paso, se analizó M11 utilizando las visualizaciones provistas por VIDEAN. De este análisis, más precisamente utilizando los gráficos de dispersión e histogramas de cada descriptor versus el *módulo elástico*, se pudo observar que ninguno de ellos presentaba valores de descriptores para valores altos de la propiedad. Es decir, que había una zona no descripta por estos descriptores. Por otro lado, en cuánto al rendimiento, M11 presentaba un coeficiente de determinación $r^2 = 0.616$.

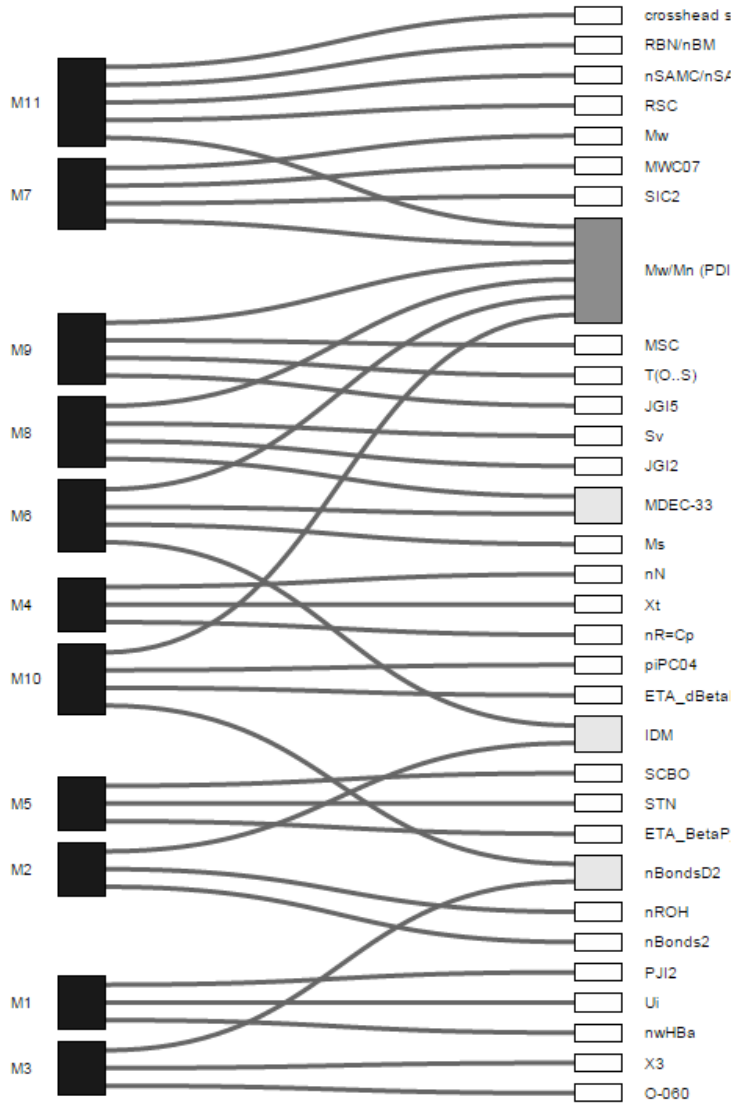


Figura 5.15 En el grafo bipartito se muestran los 10 modelos obtenidos por DELPHOS más el modelo M11 tomado como referencia de Palomba et al. (2014) que contiene el parámetro experimental CHS.

Luego, los modelos generados por DELPHOS (M1-M10) fueron analizados de la misma manera: mirando la independencia entre los descriptores y la dispersión de los valores versus la propiedad, además de evaluar su rendimiento en términos estadísticos. De ese análisis se pudo concluir que estos 10 modelos, al igual que M11, fallaban en describir la zona de valores altos de descriptor y propiedad. Por

otro lado, de estos 10 modelos los de más alta calidad predictiva fueron: M9 (r^2=0.493) y M10 (r^2 = 0.533).

El próximo paso, consistió en intervenir los modelos que mejor rendimiento habían tenido (M9 y M11). Con este fin, se realizó un análisis detallado de cada descriptor, detectando que xt y JGI5 eran buenos candidatos para incluir en intervenciones futuras, debido a que presentaban valores altos para valores altos de la propiedad (Figura 5.16). Esto permitiría completar el espacio de información no cubierto hasta este momento.

Una vez identificados estos descriptores candidatos, se procedió a intervenir los modelos M9 y M11. Al modelo M11 se le adicionó de manera separada los dos descriptores. Ambos modelos resultantes mejoraron su rendimiento (M11 + xt, r^2 = 0.630; M11 + JGI5, r^2 = 0.620). Por otro lado, se intervino el modelo M9, que ya contenía al descriptor JGI5. Este modelo incluía al descriptor T(O..S) que no aportaba información independiente respecto de los otros descriptores (analizado con el grafo de correlaciones de VIDEAN), por lo tanto, se decidió eliminarlo. Al sacar este descriptor, se pudo observar que el rendimiento del nuevo modelo M9 - T(O..S) se mantenía (r^2 = 0.496), otro indicador de que este descriptor no estaba aportando información relevante al modelo. Luego, a este modelo se le agregó el parámetro experimental CHS. Por lo tanto, se evaluó el rendimiento del nuevo modelo (M9 - T(O..S) + CHS) mostrando una notable mejoría (r^2 = 0.660).

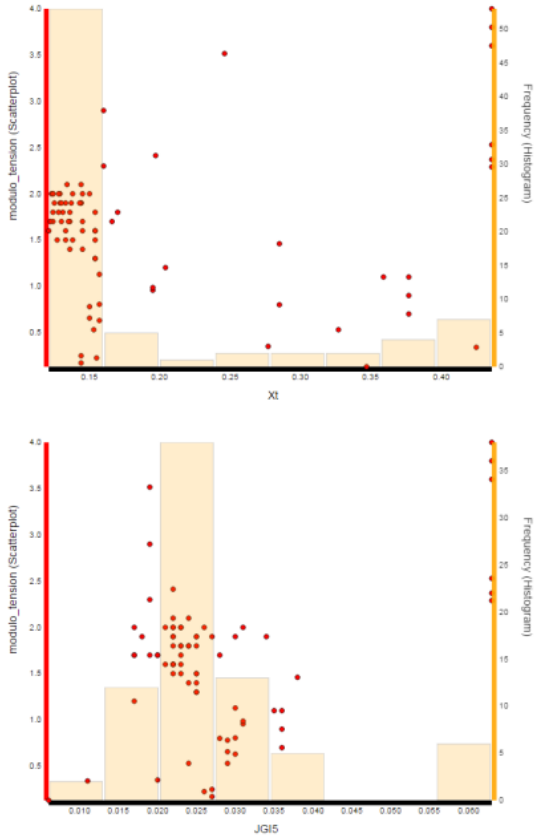


Figura 5.16 Gráficos de dispersión para los descriptores candidatos (xt y JGI5), donde se puede observar que aportan información para valores altos de la propiedad.

Además, se decidió agregar el otro descriptor candidato xt, mejorando aún más en su desempeño (r^2 = 0.68). Por último, y teniendo en cuenta que los modelos que arrojaban una buena capacidad predictiva tenían entre sus descriptores al ETA_bBetaP, se decidió incorporar este descriptor en lugar del descriptor xt. Desde un punto de vista estadístico, este último modelo (M9 - T(O..S) + CHS + ETA_bBetaP) fue el mejor obtenido con un coeficiente de determinación (r^2) de 0.693.

Comparación entre M11 y el nuevo modelo intervenido por el experto

En términos estadísticos, el nuevo modelo (M9 - T(O..S) + CHS + ETA_bBetaP) presenta un mejor rendimiento (r^2 = 0.693) que M11 (r^2 = 0.616). Además, se mantiene la misma cardinalidad (5 descriptores). Ahora se debe analizar si los descriptores tienen una interpretación físico-química en términos de la propiedad, al igual que pasaba con M11. Si bien se intenta explicar el aporte de cada descriptor por separado, no se debe perder de vista que los descriptores en el modelo actúan en conjunto y el aporte es global. Para esto debemos repasar un poco la propiedad bajo estudio.

En la porción de la curva del ensayo de tensión donde se mide el *módulo elástico,* las deformaciones moleculares son relativamente pequeñas y se asocian con la extensión de los enlaces existentes entre los átomos de las moléculas del material. Por lo tanto, la información relevante para describir la propiedad, entre otros aspectos, debería estar relacionada con los enlaces atómicos, y el tipo de átomos de los materiales de la base de datos. Por otro lado, otra característica a tener en cuenta es la polidispersión (PDI), ya que un material con distribución extensa de pesos, presentará una fracción de bajos pesos importante que influirá sobre la propiedad en estudio. Como hablamos anteriormente en el texto, el parámetro experimental CHS es muy importante cuando los ensayos de los distintos materiales se han realizado a diferentes velocidades.

El modelo (M9 - T(O..S) + CHS + ETA_bBetaP) quedó formado por los siguientes descriptores:

- Mw/Mn (PDI): es la polidispersión del material polimérico que contiene la información de la distribución de pesos.
- MSC: es la masa de la cadena lateral del monómero y brinda información relacionada al tamaño de la cadena lateral.
- JGI5: es un índice topológico que considera las transferencias de carga entre pares de átomos y la global en la molécula.
- CHS: es la velocidad de ensayo que influye firmemente en el valor de la propiedad.

- ETA_bBetaP: es un índice atómico topoquímico extendido y contiene información sobre tipo de átomo y enlace.

Finalmente, como conclusión general, pudimos observar que con el soporte de las herramientas computacionales adecuadas y utilizando el criterio químico del experto, se pueden lograr combinaciones de descriptores que modelen considerablemente bien a la propiedad bajo estudio y que sean confiables. Particularmente el modelo obtenido aquí presenta mejor capacidad predictiva que el modelo de referencia obtenido sin el soporte de una herramienta de analítica visual, manteniendo la misma cardinalidad y sin perder interpretabilidad en términos físico-químicos.

5.3.2.4 Conclusiones

Se han estudiado tres propiedades mecánicas de los polímeros. Para el caso de *elongación a la rotura* se generaron modelos de regresión y clasificación utilizando DELPHOS como técnica de selección de descriptores, y con el soporte de VIDEAN y del conocimiento del experto.

Luego para la *resistencia a la rotura*, se han estudiado diferentes modelos con el fin de analizar y comparar la aplicación de una técnica de selección de características (DELPHOS) con un método de extracción de características (CODES-TSAR). Comparando ambos métodos, podemos ver que CODES considera que la propiedad bajo estudio depende de la estructura química de la molécula, y no de una contribución de diferentes variables independientes. Por lo tanto, con esta herramienta no es necesario realizar una selección de características como la que hace DELPHOS. Por otra parte, cada característica calculada por DRAGON tiene su propia interpretación físico-química y puede utilizarse en un modelo QSAR como una pieza individual de información. Por lo tanto, se hace posible la interpretación de los modelos QSAR en términos de la contribución individual de los descriptores moleculares, ayudando a obtener modelos más comprensibles.

En nuestro caso particular, pudimos observar que los conjuntos de descriptores generados por CODES-TSAR no lograban modelar con precisión a la propiedad, mientras que la combinación (hibridación) de ambos conjuntos de descriptores (DELPHOS + CODES-TAR) lograba aumentar el rendimiento de cada conjunto por separado. Por esto, concluimos que ambos conjuntos de descriptores aportan al modelo la información necesaria para cubrir todo el espacio de la propiedad.

Por último, para la propiedad *módulo elástico*, se generaron una variedad de modelos de regresión, que fueron analizados visualmente e intervenidos según el criterio del experto para llegar a un modelo superador.

5.3.3 Diseño de Fármacos: Análisis de HIA y BBB

Aquí retomaremos el estudio presentado en el capítulo 4, en el que presentamos dos modelos generados a partir de descriptores obtenidos por la técnica de aprendizaje CODES-TSAR. Analizaremos distintos modelos de regresión y clasificación obtenidos para las propiedades *HIA (Human Intestinal Absorption, HIA)* y *BBB (Blood-Brain Barrier)*. Estos resultados se encuentran publicados en Ponzoni et al. (2017). Comenzaremos presentando modelos diseñados con descriptores obtenidos a partir de técnicas de selección, para luego analizar la hibridación (combinando técnicas de selección y aprendizaje) de los descriptores de estos últimos modelos con los presentados en el capítulo 4.

El procedimiento completo de experimentación realizado se esquematiza en la Figura 5.17. Utilizando DRAGON se calcularon descriptores para cada conjunto de datos, excluyendo el cálculo de los descriptores 3D. Esto es debido a que más adelante contrastaremos y combinaremos estos descriptores con descriptores obtenidos utilizando una herramienta de aprendizaje de descriptores que no captura las características 3D de la molécula. Luego de obtener este conjunto de descriptores, se utilizó DELPHOS para seleccionar los más relevantes.

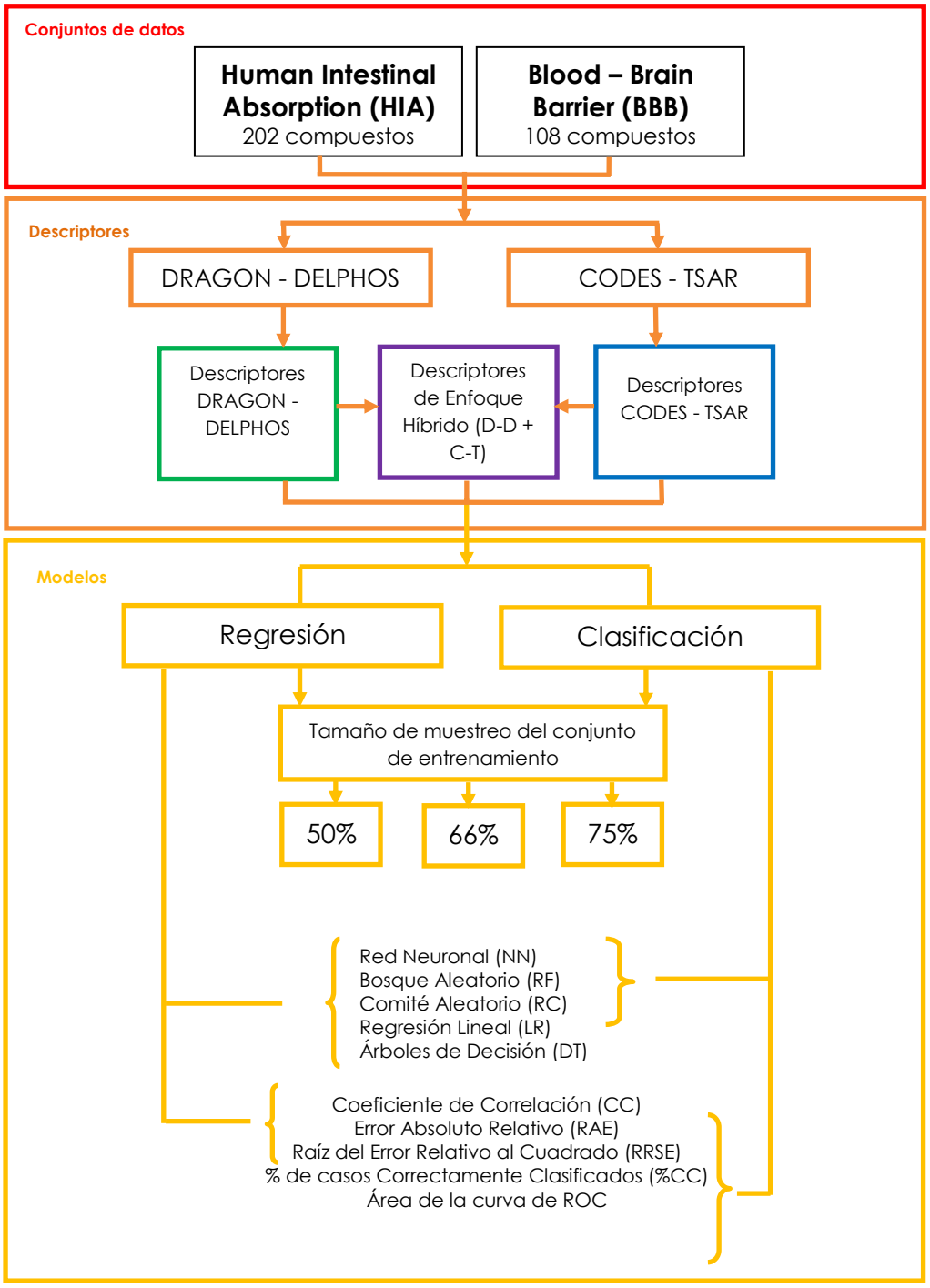


Figura 5.15 Esquema de la metodología llevada a cabo para *HIA* y *BBB*

Se han calculados diferentes modelos utilizando WEKA: Regresión Lineal (LR), Árboles de Decisión (DT), Redes Neuronales (NN), Bosque Aleatorio (BA, *random forest*) y Comité Aleatorio (CA, *random committee*). Para cada uno de estos, se utilizó la configuración de parámetros por defecto. Para evaluar el rendimiento se computaron distintas métricas utilizando WEKA. Para el caso de regresión se reportan el coeficiente de correlación (CC), el error absoluto relativo (RAE) y la raíz del error cuadrático relativo (RRSE). Para clasificación se reportaron el porcentaje de casos correctamente clasificados (%CC), el área de la curva de ROC junto con la matriz de confusión y la raíz del error cuadrático relativo (RRSE). Para la deducción de los modelos, se probaron con distintas configuraciones para la división del conjunto de datos en entrenamiento y testeo (50-50, 66-34, 75-25).

En todos los casos se utilizó el muestreo estratificado provisto por WEKA. En las Tablas 5.6 y 5.7 se muestra un resumen de los mejores modelos (regresión y clasificación) obtenidos para cada conjunto de datos, luego de aplicar todas las combinaciones de experimentos.

	Mejores modelos QSAR de regresión			
Datos	**CC**	**%datos entrenamiento**	**Método**	**Conjunto de descriptores**
HIA	0.75	75%	Redes Neuronales	Híbrido
BBB	0.76	66%	Comité Aleatorio	DRAGON-DELPHOS

Tabla 5.6 Resumen de los mejores modelos de regresión obtenidos para HIA y BBB

	Mejores modelos QSAR de clasificación				
Datos	**%CC**	**ROC**	**% datos entrenamiento**	**Método**	**Conjunto de descriptores**
HIA	86.96%	0.865	66%	Random Forest	DRAGON-DELPHOS
BBB	86.49%	0.720	66%	Redes Neuronales	DRAGON-DELPHOS

Tabla 5.7 Resumen de los mejores modelos de clasificación obtenidos para HIA y BBB.

Modelado QSAR de la *absorción intestinal humana (HIA)*

Se utilizó un conjunto de datos de 202 compuestos con valores de HIA (Guerra et al., 2010). Sobre este conjunto de datos se ejecutó DELPHOS. En este caso, se obtuvieron 25 conjuntos, de los cuáles decidimos quedarnos con aquellos dos con el menor error absoluto relativo (RAE): $M5_{HIA}$ y $M9_{HIA}$. Utilizando estos subconjuntos de descriptores se derivaron una gran variedad de modelos QSAR utilizando diferentes enfoques de aprendizaje automático.

En la tabla 5.8 se muestra un resumen de los mejores modelos obtenidos para estos dos conjuntos de descriptores. El mejor modelo QSAR de clasificación se obtuvo utilizando el subconjunto $M9_{HIA}$. Este modelo alcanza la mayor precisión (86.96%) y un área de ROC promedio de 0.865. Para los experimentos de clasificación de *HIA*, se definieron dos clases: moléculas no absorbidas y moléculas absorbidas. Cuando el valor de *HIA* está por debajo de 0.7 se considera que no absorbe, mientras que para valores iguales o por encima de 0.7 se considera que absorbe. A partir de la matriz de confusión, podemos observar que este clasificador tiene una alta precisión para los compuestos absorbidos (92.30%) y una precisión más moderada para los compuestos no absorbidos (70%). El decaimiento en el rendimiento para la segunda clase puede estar relacionada con el desequilibrio de clase en el conjunto de pruebas porque sólo el 25% de las muestras corresponden a moléculas que no se absorben.

	Mejores modelos QSAR de regresión			Mejores modelos QSAR de clasificación			
Conjunto de descriptores	**CC**	**% de datos en el conjunto de entrenamiento**	**Método**	**%CC**	**ROC**	**% de datos en el conjunto de entrenamiento**	**Método**
$M5_{HIA}$ (AMW, MATS7m, ESpm01d, TPSA(NO))	0.68	66%	Regresión Lineal	81.16%	0.803	66%	Bosque Aleatorio
$M9_{HIA}$ (AMW, GATS6v, JGI4, VRp2, TPSA(NO))	0.68	66%	Regresión Lineal	**86.96%**	**0.865**	**66%**	**Bosque Aleatorio**

Tabla 5.8 Métricas estadísticas para los modelos de regresión y clasificación obtenidos con los descriptores de $M5_{HIA}$ y $M9_{HIA}$ para *HIA*. En negrita se resalta el mejor modelo obtenido para clasificación.

Volviendo al conjunto $M9_{HIA}$, en la Figura 5.18 podemos observar la correlación de Spearman entre cada par de descriptores. Los tonos más claros indican poco nivel de correlación, es decir que cada descriptor está aportando información única al modelo.

Por otro lado, analizaremos la relevancia físico-química de los descriptores. $M9_{HIA}$ presenta un total de cinco descriptores moleculares, AMW y TPSA, reforzando así la importante correlación de estas propiedades moleculares con los valores de HIA. Otro descriptor encontrado en el modelo es GATS6v (autocorrelación de Geary de retardo 6 ponderado por el volumen de van der Waals), un índice general de autocorrelación espacial sobre el volumen de van der Waals en este caso. Además de GATS6v, hay otro descriptor en el modelo perteneciente a la familia de autocorrelaciones 2D, el JGI4 (índice de carga topológica media de orden 4). Este descriptor es capaz de evaluar la transferencia de carga entre pares de átomos y, por lo tanto, la transferencia de carga global en la molécula. Finalmente, el último descriptor de este modelo es el VRp2 (índice medio basado en vectores de tipo Randic a partir de la matriz de distancia ponderada

polarizada), un índice basado en valores propios. Estos descriptores se analizaron en términos de correlación de Spearman mediante el uso de VIDEAN. A través de este análisis, se pudo observar el resultado deseable que cada descriptor proporciona información única al modelo.

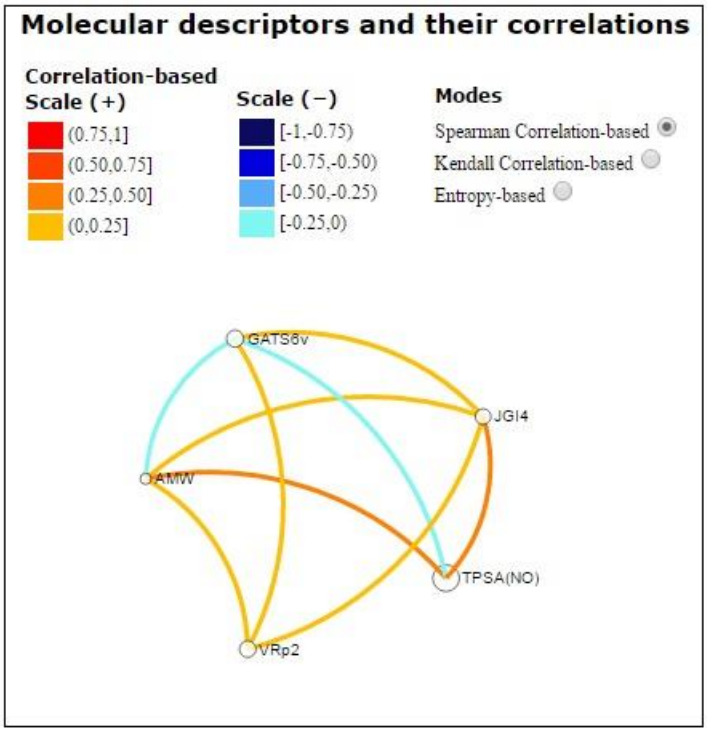


Figura 5.18 Análisis de correlación entre los descriptores de $M9_{HIA}$.

En resumen, se puede afirmar que para el conjunto de datos HIA el modelo QSAR de clasificación basado en los descriptores obtenidos por DELPHOS ($M9_{HIA}$), presenta una capacidad predictiva razonable, así como también los descriptores tienen una clara relación con la propiedad aportando cada uno información relevante. Para el caso de regresión, ambos subconjuntos presentan un bajo y parecida calidad predictiva. Más adelante presentaremos un modelo superador, aplicando un enfoque híbrido para la elección de los descriptores (selección y aprendizaje).

Modelado QSAR de la *barrera hematoencefálica (BBB)*

El conjunto de datos utilizado tiene 108 compuestos con valores conocidos de *logBB* publicados en Guerra et al. (2008). Respecto a los subconjuntos de descriptores moleculares, al igual que para el caso anterior, decidimos tomar los dos subconjuntos con menor error absoluto relativo (RAE) reportados por DELPHOS, que corresponden a los subconjuntos $M2_{BBB}$ y $M13_{BBB}$. A partir de estos dos subconjuntos, se generaron una variedad de modelos de regresión y clasificación. Para los experimentos de clasificación BBB, se definieron tres clases: moléculas que cruzan la barrera hematoencefálica (BBB <= 0.7), moléculas que no atraviesan la barrera hematoencefálica (BBB > -0.3) y una zona gris que representa incertidumbre (BBB <= -0.3 y BBB > 0.7). En la tabla 5.9 se muestra para cada subconjunto los mejores modelos obtenidos tanto para regresión como para clasificación. De esta tabla podemos analizar que los mejores modelos tanto para regresión como para clasificación fueron obtenidos para el subconjunto $M13_{BBB}$. Para el caso de clasificación, a partir de la matriz de confusión, podemos observar que este modelo QSAR tiene una alta precisión para compuestos que atraviesan la barrera hematoencefálica (85.71%) y compuestos que no atraviesan la barrera hematoencefálica (100%). Sin embargo, la precisión del clasificador disminuye para los compuestos en la zona gris. La caída de rendimiento para esta clase intermedia puede estar relacionada con el desequilibrio de clase en el conjunto de pruebas, ya que sólo el 10,81% de las muestras corresponde a moléculas en la zona gris. Este hecho también puede explicar el valor moderado del área ROC promedio (0.72).

	Mejores modelos QSAR de regresión			Mejores modelos QSAR de clasificación			
Conjunto de descriptores	**CC**	**% de datos en el conjunto de entrenamiento**	**Método**	**%CC**	**ROC**	**% de datos en el conjunto de entrenamiento**	**Método**
$M2_{BBB}$ (nR06, SIC1, CIC5)	0.63	50%	Bosque Aleatorio	75.95%	0.793	50%	Bosque Aleatorio
$M13_{BBB}$ (AMW, RBN, MATS5e, MATS4p, EEig12d, JGI7, Hy)	**0.76**	**66%**	**Comité Aleatorio**	**86.49%**	**0.720**	**66%**	**Redes Neuronales**

Tabla 5.9 Métricas estadísticas de los mejores modelos obtenidos con los descriptores $M2_{BBB}$ y $M13_{BBB}$ para BBB. En negrita se resaltan los mejores modelos obtenidos para regresión y clasificación.

Como en el caso anterior, para asegurarnos que los descriptores de $M13_{BBB}$ estaban aportando información no redundante entre ellos, analizamos los descriptores con VIDEAN. Más precisamente observamos la correlación de Spearman entre pares de descriptores pudiendo corroborar la baja redundancia de datos en el modelo (Figura 5.19).

Por otro lado, se realizó un análisis dese el punto de vista físico-químico de los descriptores de $M13_{BBB}$. Algunos de ellos están relacionados con los índices constitucionales (0D), tales como el peso molecular promedio (AMW) o el número de enlaces giratorios (RBN). Otra familia importante de descriptores encontrados fue la autocorrelación 2D. Por otra parte, dos de cada tres descriptores en esta familia están en relación con el coeficiente de Moran (MATS4P, MATS5e) con respecto a polarizabilidad y a la electronegatividad de Sanderson, respectivamente. El otro descriptor de esta familia involucrada en el modelo es JGI7, y está relacionado con la carga topológica. Por último, dos descriptores más están presentes en el modelo, EEig12d y Hy, que son parte de los índices de adyacencia de borde y las familias de propiedades moleculares,

respectivamente. La primera está en relación con los momentos dipolares, mientras que la segunda tiene una relación directa con la hidrofobicidad de las moléculas.

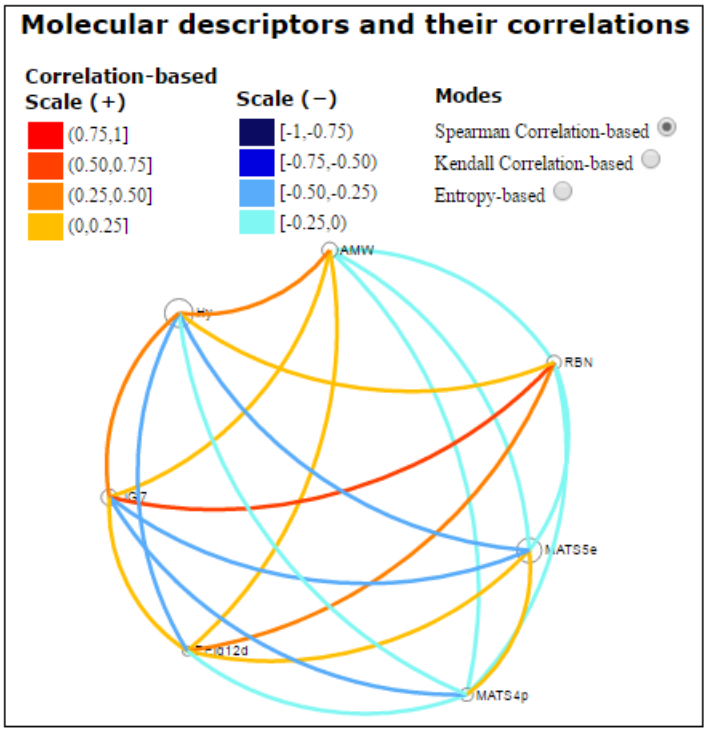


Figura 5.19 Análisis de información aportada entre los descriptores de $M13_{BBB}$.

En conclusión, algunos de estos descriptores están en relación directa con propiedades bien conocidas de moléculas que permiten o impiden que los compuestos pasen por la barrera hematoencefálica, tales como el peso molecular o logP. Se ha probado extensamente que los compuestos con valores de logP más altos son más propensos a pasar la BBB, mientras que los compuestos con valores logP bajos tienen dificultades para cruzar la barrera. De manera similar, los compuestos que muestran un peso molecular muy alto normalmente no son propensos a cruzar la barrera hematoencefálica. Además, parámetros como polarizabilidad, electronegatividad de Sanderson, momentos dipolares y carga

topológica están en relación con la hidrofobicidad y la distribución de carga de las moléculas, afectando así su capacidad para cruzar la BBB.

En resumen, podemos afirmar que para el conjunto de datos BBB los modelos QSAR basado en los descriptores obtenidos por DELPHOS ($M13_{BBB}$), logran una precisión de predicción razonable. Con respecto a esto, los descriptores moleculares elegidos por el método de selección de características están claramente relacionados con el logBBB en términos fisicoquímicos y contienen bajos niveles de redundancia.

5.3.3.1 Enfoque híbrido de selección y aprendizaje de descriptores

El propósito aquí es presentar modelos QSAR híbridos, en el sentido que una parte de los descriptores de los modelos se obtuvieron a partir de técnicas de selección de descriptores y otra parte con técnicas de aprendizaje de descriptores.

Luego de probar la hibridización en el campo del diseño de materiales, otra idea que resultaba interesante explorar era la evaluación de esta estrategia híbrida en el contexto de diseño de drogas.

Modelado QSAR para la absorción intestinal humana (HIA)

En este caso combinamos los descriptores de los subconjuntos seleccionados a través de DELPHOS: $M5_{HIA}$ y $M9_{HIA}$, con el conjunto obtenido mediante CODES-TSAR: CT_{HIA}, resultando de esta manera dos subconjuntos más: $M5_{HIA} \cup CT_{HIA}$ y $M9_{HIA} \cup CT_{HIA}$. En la tabla 5.12 se resumen las métricas para los mejores modelos.

Para poder evaluar la contribución real de los descriptores de CT_{HIA} al rendimiento del modelo QSAR de regresión combinado, decidimos ejecutar un experimento adicional. La idea era evaluar la significancia estadística de la contribución de los

descriptores de CT_{HIA} al modelo combinado, en contraste con una selección aleatoria de descriptores moleculares.

	Mejores modelos QSAR de regresión			Mejores modelos QSAR de clasificación			
Conjunto de descriptores	**CC**	**% de datos en el conjunto de entrenamiento**	**Método**	**%CC**	**ROC**	**% de datos en el conjunto de entrenamiento**	**Método**
$M5_{HIA}$ U CT_{HIA} (AMW, MATS7m, ESpm01d, TPSA(NO), CODES-T1, CODES-T2, CODES-T3)	0.76	75%	Redes Neuronales	80.00%	0.764	75%	Bosque Aleatorio
$M9_{HIA}$ U CT_{HIA} (AMW, GATS6v, JGI4, VRp2, TPSA(NO), CODES-T1, CODES-T2, CODES-T3)	0.68	66%	Regresión Lineal	81.16%	0.856	66%	Bosque Aleatorio

Tabla 5.12 Métricas para los mejores modelos de descriptores combinados obtenidos para *HIA*.

En este caso para regresión, el mejor modelo obtenido para $M5_{HIA}$ U CT_{HIA} logró superar en rendimiento a los modelos obtenidos anteriormente sobre cada subconjunto por separado ($M5_{HIA}$: 0.68 y C_{THIA}: 0.41). Al igual que para los casos anteriores, los descriptores fueron analizados utilizando VIDEAN, con el objetivo de analizar las relaciones entre los descriptores. En este análisis se pudo observar que los descriptores del conjunto CT_{HIA} poseen alta información mutua entre ellos, pero baja entre cada uno de ellos con respecto a los descriptores de $M5_{HIA}$. Por lo que podemos observar el aporte de información complementaria entre los dos subconjuntos (Figura 5.20).

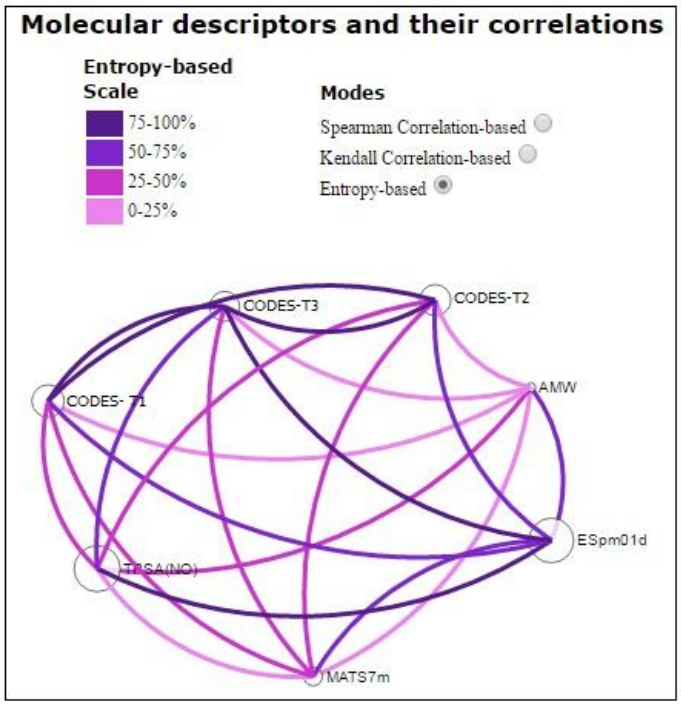


Figura 5.20 Información mutua entre los descriptores del modelo combinado $M5_{HIA}$ ∪ CT_{HIA}.

Para este análisis, se realizaron cien repeticiones de un experimento aleatorio. En cada repetición, tres descriptores moleculares incluidos en CT_{HIA} son sustituidos por tres descriptores moleculares ($RAND_{HIA}$) tomados aleatoriamente de todo el conjunto de descriptores moleculares calculados por DRAGON (excluyendo los descriptores moleculares de $M5_{HIA}$). Después de este reemplazo, se vuelve a calcular un modelo QSAR de regresión usando el nuevo subconjunto combinado $M5_{HIA}$ + $RAND_{HIA}$, aplicando las mismas condiciones experimentales reportadas para el mejor modelo QSAR de regresión para *HIA* (ver Tabla 5.6). De esta manera, se obtiene una distribución de frecuencia de los coeficientes de correlación (CC) a partir de cien modelos QSAR deducidos con estos subconjuntos aleatorios (Figura 5.21). Analizando estos resultados, podemos observar que el valor medio de precisión de estos modelos QSAR es muy baja (0.397). Además, ningún modelo de regresión generado a partir de los subconjuntos aleatorios logró el coeficiente de correlación del mejor modelo de

regresión para *HIA* (0.76). Por lo tanto, podemos concluir que la contribución del subconjunto CT_{HIA} al modelo combinado es claramente relevante en términos estadísticos.

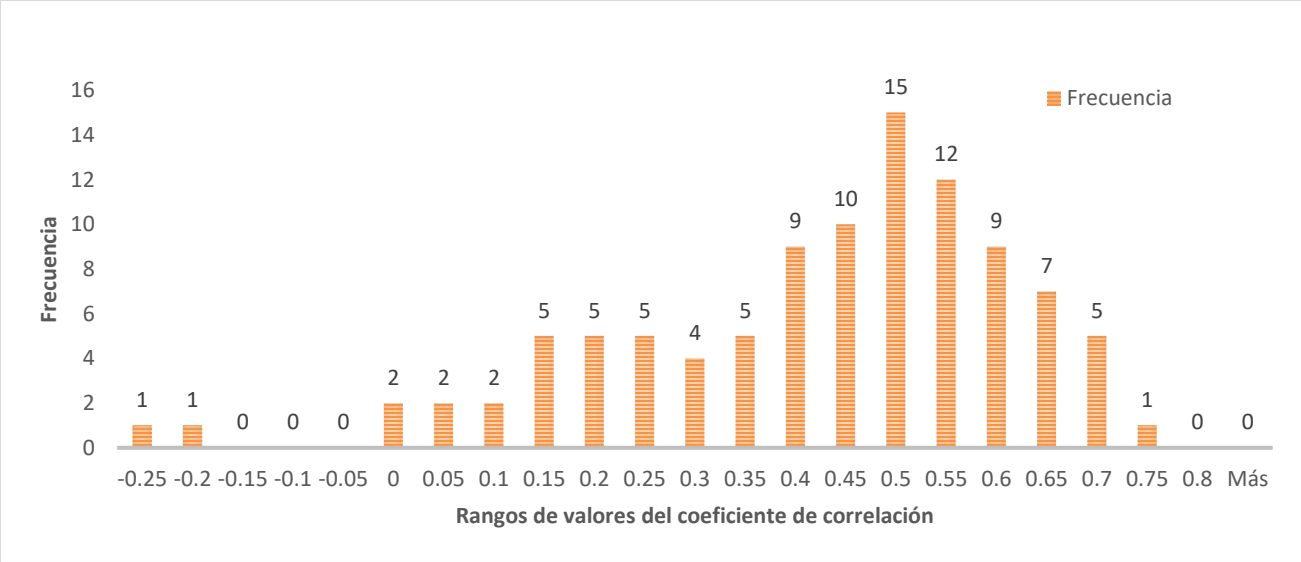


Figura 5.21 Valores de frecuencia de precisión para los modelos de regresión obtenidos tomando subconjuntos aleatorios de descriptores.

En resumen, podemos afirmar que el mejor modelo QSAR obtenido para el conjunto de datos *HIA* mediante la combinación de DELPHOS con CODES-TSAR logra una precisión razonable en las predicciones. Los descriptores moleculares elegidos por el método de selección de características están claramente relacionados con *HIA* en términos físico-químicos, y también presentan bajos niveles de redundancia. Para el caso particular del modelo de regresión combinado, podemos observar que los descriptores moleculares proporcionados por CODES-TSAR mejoraron el rendimiento del modelo QSAR generado por DRAGON/DELPHOS con significancia estadística.

Modelado QSAR para la barrera hematoencefálica (BBB)

Siguiendo la misma idea que para *HIA,* aquí combinamos los subconjuntos $M2_{BBB}$ y $M13_{BBB}$ con CT_{BBB}, realizando siempre el mismo conjunto de experimentos. En la Tabla 5.13 se puede observar que los mejores modelos combinados inferidos no superan en rendimiento a los mejores presentados anteriormente (regresión con $M13_{BBB}$: 0.76 y clasificación con $M13_{BBB}$: 86.5%).

	Mejores modelos QSAR de regresión			**Mejores modelos QSAR de clasificación**			
Conjunto de descriptores	**CC**	**% de datos en el conjunto de entrenamiento**	**Método**	**%CC**	**ROC**	**% de datos en el conjunto de entrenamiento**	**Método**
$M2_{BBB}$ U CT_{BBB} (nR06, SIC1, CIC5, CODES-T1, CODES-T2, CODES-T3)	0.59	50%	Bosque Aleatorio	77.78%	0.852	75%	Bosque Aleatorio
$M13_{BBB}$ U CT_{BBB} (AMW, RBN, MATS5e, MATS4p, EEig12d, JGI7, Hy, CODES-T1, CODES-T2, CODES-T3)	0.70	75%	Bosque Aleatorio	85.19%	0.743	50%	Redes Neuronales

Tabla 5.13 Métricas para los mejores modelos de descriptores combinados obtenidos para *BBB*.

5.3.3.3. Conclusiones

En esta sección hemos comparado dos metodologías de aprendizaje automático, DRAGON-DELPHOS y CODES-TSAR, ya aplicadas con anterioridad para el estudio de la *resistencia a la rotura*. La primera como representante de las técnicas de selección de descriptores y la segunda en representación de las metodologías de aprendizajes de descriptores. Los experimentos se llevaron a

cabo con bases de datos pertenecientes a dos propiedades: *absorción intestinal humana (HIA)* y *barrera hematoencefálica(BBB)*.

En todos los casos, los resultados de los modelos QSAR se contrastaron en varias condiciones experimentales, variando los parámetros de muestreo y las técnicas utilizadas para inferir los modelos de clasificación y regresión. También se analizaron los subconjuntos de descriptores moleculares obtenidos por las estrategias DRAGON-DELPHOS y CODES-TSAR en términos de información mutua y correlación, a fin de evaluar las asociaciones entre pares de descriptores relevantes y sus relaciones con las propiedades estudiadas. A partir de los resultados, se observó que ninguno de los métodos supera al otro en todos los escenarios, ya que la exactitud de la predicción depende de las características de la base de datos y las condiciones experimentales. Sin embargo, con respecto a los métodos de entrenamiento utilizados para la inferencia del modelo QSAR, las técnicas basadas en conjuntos, Bosque Aleatorio y Comité Aleatorio, superaron a los algoritmos más tradicionales en los dos tipos de modelos QSAR (regresión y clasificación). Por esta razón, se recomienda aplicar métodos combinados de aprendizaje para la construcción de un modelo.

Otro punto a tener en cuenta, al igual que cuando se aplicó en el campo de diseño de materiales, está asociado con las características intrínsecas de cada metodología. Por esta razón, cada modelador puede elegir una metodología dependiendo en qué aspecto desea centrarse: los esfuerzos computacionales o la interpretabilidad del modelo.

Aquí también decidimos evaluar el impacto de combinar estas dos técnicas. Estos experimentos de hibridación sobre los conjuntos de datos revelaron que la precisión de los modelos QSAR se puede mejorar uniendo subconjuntos de descriptores moleculares obtenidos por ambas metodologías si estos subconjuntos contienen información complementaria para los modelos, como ocurrió con el mejor modelo de regresión de *HIA*. Por esta razón, como conclusión general,

recomendamos considerar este enfoque híbrido como una estrategia adicional para tener en cuenta en los experimentos.

5.4 Sumario

Como venimos hablando a lo largo de los distintos capítulos, una estrategia de selección de descriptores en la que se incorpora al experto en el dominio, es sumamente requerida, con el fin de mejorar la confianza del modelador respecto del conjunto final de descriptores y de esta manera lograr diseñar modelos más generalizables e interpretables en términos físico-químicos.

En el presente capítulo, con el fin de alcanzar este desafío, presentamos el desarrollo de VIDEAN, una herramienta de software que combina métodos estadísticos con visualizaciones interactivas. La idea principal fue aprovechar la experiencia de los químicos para realizar una exploración visual interactiva de diferentes porciones de datos. Se diseñaron varias representaciones visuales coordinadas con el objetivo de captar diferentes aspectos y relaciones entre los descriptores y la propiedad objetivo. A través del uso de estas visualizaciones interactivas se pueden obtener piezas de información valiosa relacionada con: descriptores redundantes, descriptores que proporcionan información discriminativa, descriptores relevantes por consenso entre los modelos alternativos, y descriptores cuyo conocimiento ayuda a reducir la incertidumbre sobre el valor de la propiedad objetivo.

Las capacidades de VIDEAN fueron evaluadas a través de múltiples escenarios, en los cuales se pudo observar el soporte sólido que brinda al modelador en el proceso de la selección de descriptores, requiriendo de un esfuerzo cognitivo menor comparado con el análisis manual de cada uno de ellos. Además de las ventajas mencionadas anteriormente de incorporar el analista en el proceso, también es importante para establecer las limitaciones y el alcance de VIDEAN. Aquí nos centramos en la etapa de selección de descriptores. Con el fin de abordar el problema más general de la predicción de la actividad/propiedad,

sería de mucha utilidad incluir aspectos tales como el estudio de los compuestos de manera individual o en familias, así como la incorporación de visualizaciones relacionadas con el de dominio de aplicación de los modelos QSAR.

Capítulo 6: Identificación del Dominio de Aplicación de Modelos QSAR/QSPR

En los capítulos anteriores nos hemos enfocado principalmente en la generación de modelos QSAR/QSPR y más específicamente en una etapa muy importante de este proceso que es la selección de los descriptores más relevantes para este modelo. Aquí nos centraremos en otra etapa importante dentro del modelado QSAR/QSPR relacionada con la confiablidad del modelo obtenido, que ha sido introducida en el capítulo 3. Una vez entrenado el modelo, se debería poder determinar si para un compuesto nuevo, la predicción va resultar confiable o no. En otras palabras, lo que queremos encontrar es el dominio de aplicación (DA) del modelo QSAR/QSPR.

Si bien este es un requerimiento importante, muchos modelos presentados en la literatura científica no tienen disponible su correspondiente DA. Uno de los motivos principales reside en la dificultad de su determinación y la manera en que debe ser medido. Aunque una buena definición del DA ayuda al usuario del modelo a estimar la fiabilidad de sus predicciones, no debe asumirse que todas las predicciones dentro del DA definido son necesariamente confiables (OECD, 2007). En la práctica, una predicción podría ser poco confiable aunque el compuesto se encuentre dentro de los dominios estructurales y físico-químicos establecidos para el modelo. Esto podría ocurrir, por ejemplo, en los casos en que el compuesto nuevo actúa por un mecanismo de acción diferente que no ha sido capturado por el el algoritmo de aprendizaje.

En este capítulo se revisarán algunos métodos propuestos en la literatura para determinar el DA de un modelo y se presentará una metodología para aplicar sobre modelos de clasificación basada en el método presentado en Soto et al. (2009).

6.1 Formas de definir el dominio de aplicación

Clasificaremos las técnicas para definir el domino de aplicación en dos grandes grupos, siguiendo la propuesta presentada en Mathea et al. (2016) en conjunto con la discutida en Sahigara et al. (2012). De esta manera, podemos determinar el DA: (i) en términos del espacio de los descriptores moleculares, y (ii) en términos de la estimación de la confianza de las predicciones. Si bien las presentamos por separado por una cuestión de taxonomía, creemos que es indispensable utilizar estas técnicas en conjunto. De esta manera, no sólo se tiene en cuenta el espacio de los descriptores, sino también el rendimiento del modelo. En lo sucesivo, siempre que hablemos de un "compuesto nuevo" (CN) estaremos haciendo referencia al compuesto sobre el que se quiere determinar el dominio de aplicación.

6.1.1 AD basado en el espacio de los descriptores moleculares

Estos métodos se basan en la comparación del CN con los compuestos del conjunto de entrenamiento. Las características que se comparan en este caso son los descriptores moleculares. A continuación, se presentarán los métodos más comúnmente utilizados para este fin.

Métodos basados en distancias: este tipo de métodos permite determinar la distancia entre un CN y un compuesto del conjunto de entrenamiento o un conjunto de ellos (vecinos más cercanos). Para este último caso, es posible calcular las distancias a los *k* vecinos más cercanos y usar la distancia media o la distancia máxima al k-ésimo vecino como criterio de decisión. Siempre debe definirse un umbral para poder discriminar entre compuestos similares o no. Otra variante es requerir que al menos *k* objetos de conjunto de entrenamiento estén dentro de un umbral de distancia particular (es decir, este método garantiza una densidad local predefinida). Hay diferentes tipos de distancias que se pueden

utilizar. La más comúnmente empleada es la Euclídea. Sin embargo, la correlación entre las variables puede distorsionar esta medida. Por lo tanto, para que la distancia euclídea sea significativa, las variables deben ser no correlacionadas y ser de escalas comparables (Krzanowski, 2000). Generalmente se suele utilizar la distancia Euclídea para datos continuos, la distancia de Mahalanobis (Mahalanobis, 1936) para datos continuos en caso de multicolinearidad, la distancia de Manhattan (Krause, 2012) para enteros en casos de datos discretos y la similitud de Tanimoto simple (Rogers y Tanimoto, 1960) para datos binarios.

Métodos geométricos y basados en rangos: en algunos casos, en lugar de calcular distancias, el método simplemente determina que el CN se encuentre dentro del rango de las variables cubiertas por el conjunto de entrenamiento, ya sea uno por uno o incluyendo a todos los compuestos (Jaworska et al., 2005). Entre los métodos más conocidos se encuentran:

1. Cuadro delimitador (*bounding box*), que considera el rango de los descriptores individuales utilizados para construir el modelo. Suponiendo una distribución uniforme, el dominio de aplicación se puede imaginar como una caja de límites que es un hiper-rectángulo *p*-dimensional definido sobre la base de valores máximos y mínimos de cada descriptor utilizado para construir el modelo. Los lados de este hiper-rectángulo son paralelos con respecto a los ejes de coordenadas. Sin embargo, pueden presentarse varios inconvenientes asociados a este enfoque, ya que sólo se tienen en cuenta los rangos de los descriptores, sin tener en cuenta por ejemplo, la correlación entre los descriptores y sin detectar posibles regiones vacías dentro del espacio de interpolación (Jaworska, et al., 2005; Netzeva et al., 2005).

2. Casco convexo (*Convex Hull*), donde el espacio de interpolación se define por el área convexa más pequeña que contiene el conjunto de entrenamiento completo. La implementación de un casco convexo puede

ser un desafío con el aumento de la dimensionalidad de los datos. En el caso de datos de dos o tres dimensiones, existen varios algoritmos propuestos, aunque el aumento de las dimensiones contribuye al orden de la complejidad (Preparata y Shamos, 1991). Además, los límites de los conjuntos se analizan sin considerar la distribución real de los datos. De manera similar al enfoque basado en rangos, este enfoque no puede identificar las posibles regiones vacías internas dentro del espacio de interpolación (Jaworska et al., 2005; Netzeva et al., 2005).

Métodos basados en la distribución de la densidad de probabilidad: estos métodos se basan en estimar la *función de densidad de probabilidad* para los datos dados. Una característica principal de estos enfoques es su capacidad para identificar las regiones vacías internas. Además, si es necesario, la distribución real de datos puede reflejarse generando regiones cóncavas alrededor de los bordes espaciales de interpolación (Jaworska et al., 2005; Netzeva et al., 2005). Generalmente, estos enfoques se implementan estimando la densidad de probabilidad del conjunto de datos seguido por la identificación de la región de mayor densidad que consiste en una fracción conocida (dada por el usuario) de la masa de probabilidad total (Netzeva et al., 2005). De esa manera, se crea el potencial para cada instancia en el conjunto de entrenamiento de tal forma que sea más alto para esa instancia y disminuya con la distancia. Una vez que el potencial es calculado para todo el conjunto de datos, el potencial global se obtiene mediante la adición de los potenciales individuales, lo que define la densidad de probabilidad (Forina et al., 1991; Jouan-Rimbaud et al., 1999).

6.1.2 AD basado en la estimación de la confianza

En la sección anterior se exploraron los métodos más comúnmente utilizados para determinar el DA basándose en el espacio de descriptores. Aquí, hablaremos de aquellos que se basan en estimar la confianza de las predicciones analizando la

calidad predictiva del modelo en la proximidad del CN. Para esto pueden utilizarse varios enfoques. Uno de ellos consiste en emplear diferentes conceptos de la estadística multivariada para realizar la estimación, utilizando por ejemplo la varianza de las predicciones de los compuestos similares al CN dentro del conjunto de entrenamiento (Soto et al., 2009; Briesemeister et al., 2012; Keefer et al., 2013). También pueden utilizarse enfoques que utilicen las probabilidades posteriores para evaluar la confiabilidad en las predicciones (Mathea et al., 2016; Soto et al., 2011). Por último, un enfoque que se está utilizando en este último tiempo es la predicción de conformidad (*conformal prediction*), que proporciona un marco para producir medidas válidas de confianza en las predicciones, suponiendo que los datos son independientes e idénticamente distribuidos (Mathea et al., 2016; Norinder et al., 2016).

6.2 Metodología propuesta

La metodología que proponemos busca identificar si la predicción estimada por un modelo QSAR de clasificación sobre un compuesto nuevo será confiable o no. Para desarrollar esta técnica tomamos como punto de partida el método presentado en Soto et al. (2009), en el que se propone una metodología híbrida para determinar el DA de un modelo. A grandes rasgos, esa metodología utiliza mapas auto-organizativos (SOMs) como enfoque no supervisado para determinar la similaridad entre los compuestos en conjunto con estadística multivariada para evaluar las predicciones del modelo de regresión. De esta manera permite determinar si la predicción de un compuesto nuevo será *confiable*, *no confiable* o *no categorizada* (Figura 6.1).

La propuesta ahora es determinar si la predicción de un compuesto nuevo será confiable (o no) cuando tenemos un modelo de clasificación. Para esto, decidimos experimentar con la técnica de agrupamiento difuso (Bezdek et al.,

1984) para determinar el conjunto de compuestos vecinos. En el agrupamiento difuso (*soft clustering*) se le asigna a cada objeto un grado de pertenencia a cada cluster, cuanto más alto es este valor mayor es la similitud entre el objeto y el resto del cluster. La elección de utilizar este enfoque difuso estuvo basada en la flexibilidad que supone que un compuesto tenga distintos grados de pertenencia a clústeres diferentes, lo que permite definir distintos puntos de cortes para la selección de los compuestos más similares.

Para aplicar la metodología comenzamos con un conjunto de datos, el cual será dividido en un set de entrenamiento(E) y otro set de testeo(T). Los compuestos de este último serán utilizados para evaluar la metodología de DA. Con el conjunto E se armarán los clusters y a partir de ellos se determinarán los clusters vecinos (los más similares) para el compuesto nuevo (CN). Es decir, que lo que queremos lograr es poder representar a ese CN con compuestos de E y así poder estimar el nivel de confianza en la predicción. Cada CN pertenece al conjunto de datos T. Además, se supone que tenemos un clasificador que ha sido entrenado con los compuestos de E. En este sentido parte del rendimiento del método de AD tiene que ver con la buena generalidad del modelo. Los valores de los descriptores del conjunto E son escalados entre 0 y 1, para luego aplicar sobre ellos el algoritmo de agrupamiento difuso.

Dado un CN al que se le quiere determinar su DA, los pasos que sigue el algoritmo son:

1. Calcular su valor de pertenencia a cada uno de los clusters definidos sobre el conjunto E.
2. Seleccionar el cluster donde el CN tiene mayor valor de pertenencia.
3. De ese cluster, se definirán como vecinos del CN, aquellos compuestos que cumplan las siguientes condiciones:
    - su valor de pertenencia sea mayor o igual al valor de pertenencia del CN

- su distancia al centroide sea menor o igual a la distancia entre el centroide y el CN

4. Evaluar condiciones para determinar la categoría del CN (*confiable*, *no confiable* o *no categorizado*)

Las condiciones del punto 4 pretenden cubrir dos aspectos. El primer aspecto tiene como objetivo evaluar que la cantidad de compuestos que integran el conjunto de vecinos sea significativa, para de este modo asegurar que el CN está siendo debidamente representado. El segundo aspecto pretende determinar si la lógica del modelo ha sido capturada correctamente en el entorno definido por los vecinos de CN. Para esto se evalúa dentro del conjunto de vecinos, las probabilidades de las clases y el porcentaje de compuestos correctamente clasificados.

Más específicamente, dado un modelo de clasificación con un conjunto de etiquetas de clases $C = \{C_1, C_2, \ldots C_n\}$, la probabilidad P_i definida como $\max\{p_i\}_{i=1..n}$ donde p_i representa la probabilidad de cada clase C_i dentro del conjunto de vecinos, la cantidad de compuestos vecinos (CV), el porcentaje de compuestos que han sido correctamente clasificados (%ccv) y la definición de los correspondientes umbrales, se procede a evaluar las condiciones para categorizar a un CN de la manera que se ilustra en la Figura 6.1.

6.2.1 Diseño de experimentos

Con el fin de evaluar el funcionamiento del método, se utilizaron dos conjuntos de datos presentados en capítulos anteriores. El primer conjunto de datos está formado por 5 descriptores, 202 compuestos y corresponde al estudio de la *absorción intestinal humana* (HIA). En este caso, un compuesto puede ser clasificado como *absorbido* o *no absorbido* por el intestino. El segundo conjunto tiene 122 compuestos, 5 descriptores y la variable experimental es *log* P_{liver}, la cual

permite discriminar si un compuesto tiene *afinidad por el hígado*, *afinidad por la sangre* o la *misma afinidad por ambos medios*, con el fin de determinar la potencial toxicidad del mismo.

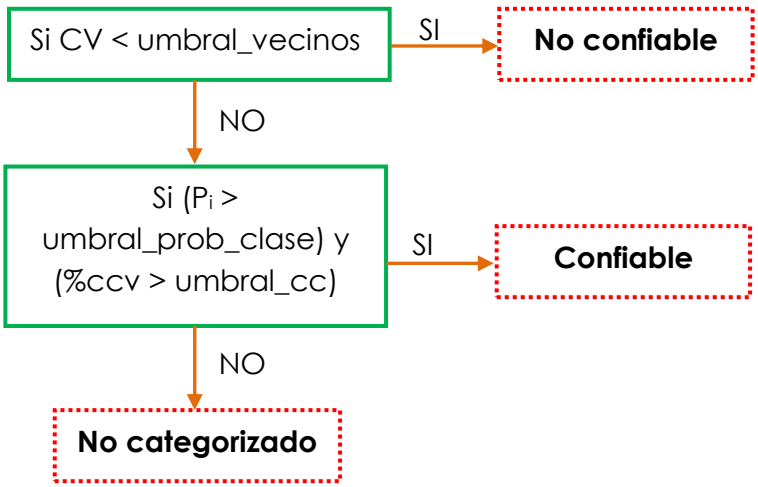


Figura 6.1 Evaluación de condiciones para categorizar el dominio de aplicación de un compuesto nuevo.

El método utilizado para entrenar los modelos fue bosque aleatorio (*random forest*). Previamente a la aplicación de este método de modelado, los compuestos fueron divididos de manera que un 75% de los datos se utilizaron para el entrenamiento y el 25% restante para testeo. Esta separación fue realizada de manera aleatoria y estratificada, con el fin de asegurar que ambos conjuntos queden formados por compuestos que representen todos los posibles valores de la variable experimental. La metodología ha sido aplicada 10 veces a cada uno de los dos conjuntos de datos, realizando en cada réplica una nueva separación de los compuestos en entrenamiento y testeo.

Por último, el número de clusters utilizados para *HIA* fue 4 y para *log* P_{liver}, 3. Si bien no existe un criterio objetivo para elegir el número indicado de clusters, existen varios métodos que pueden utilizarse para tener una aproximación. Uno que

hemos explorado fue el método de Elbow (Kodinariya y Makwana, 2013), que busca el punto en el que se produce un cambio abrupto en la curva representada entre el número de clusters y la suma de las distancias al cuadrado de cada objeto del cluster a su centroide. La ubicación de ese cambio abrupto en la curva sugiere un número adecuado de clusters. Es importante aclarar, que además de esta aproximación, se ha tenido en cuenta el tamaño del conjunto de datos y el rendimiento del método. El umbral para evaluar la cantidad de vecinos de un CN fue definido como el 5% de la cantidad total de compuestos del conjunto de entrenamiento y los umbrales para la probabilidad de clase y los compuestos correctamente clasificados se fijaron en 0.6 y 0.7 respectivamente. En la tabla 6.1 se resumen los aspectos más importantes tanto de los conjuntos de datos como de los dos experimentos llevados a cabo.

	HIA	***log P_{liver}***
Cantidad de compuestos	202	122
Cantidad de etiquetas de clase par la variable experimental	2	3
Cantidad de descriptores	5	5
Cantidad de compuestos de entrenamiento (75%)	102	92
Cantidad de compuestos de testeo (25%)	50	29
Método de modelado	Bosque Aleatorio	Bosque Aleatorio
Cantidad de clústeres	4	3
Umbral para la cantidad de vecinos requerida (umbral_vecinos)	8	5
Umbral para la probabilidad de clase P_i	0.5	0.5
Umbral para compuestos correctamente clasificados (%cc)	0.7	0.7

Tabla 6.1 Resumen de las características de las bases de datos y de los parámetros utilizados para aplicar el método de DA.

6.2.2 Resultados

Las Figuras 6.2 y 6.3 muestran para HIA y *log* P_{liver} respectivamente, los porcentajes de compuestos del conjunto de testeo en cada categoría. En la figura 6.2 se puede observar que el 55% de los datos fue categorizado como *confiable* y de este porcentaje, un 64% de los compuestos coincidió en la etiqueta predicha por el modelo con la etiqueta asignada por el método de DA.

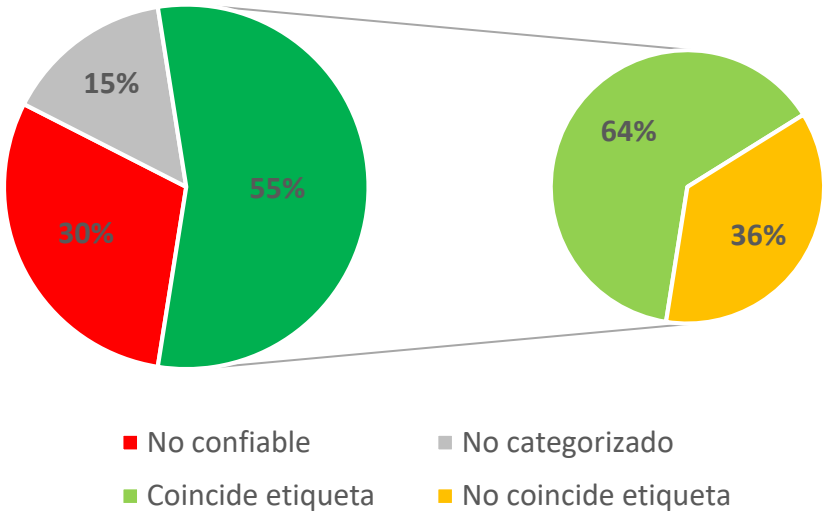


Figura 6.2 Porcentaje de compuestos clasificados en cada clase para HIA, luego de aplicar el método de DA.

Para log P_{liver} (Figura 6.3) el porcentaje de compuestos categorizados como *confiables* fue del 26%, con un 85% de compuestos para los que coincidió la etiqueta de clase predicha con la etiqueta de clase asignada por el método. En este caso los porcentajes de compuestos clasificados como *no confiables* y *no categorizados* son muy similares (37% y 38% respectivamente). Estos resultados no son alentadores en principio para la valoración del modelo QSAR, no siendo de la misma manera en relación al método de DA, ya que dentro de los compuestos

categorizados como *confiables*, un porcentaje alto (85%) fue clasificado de manera consistente con la predicción del modelo.

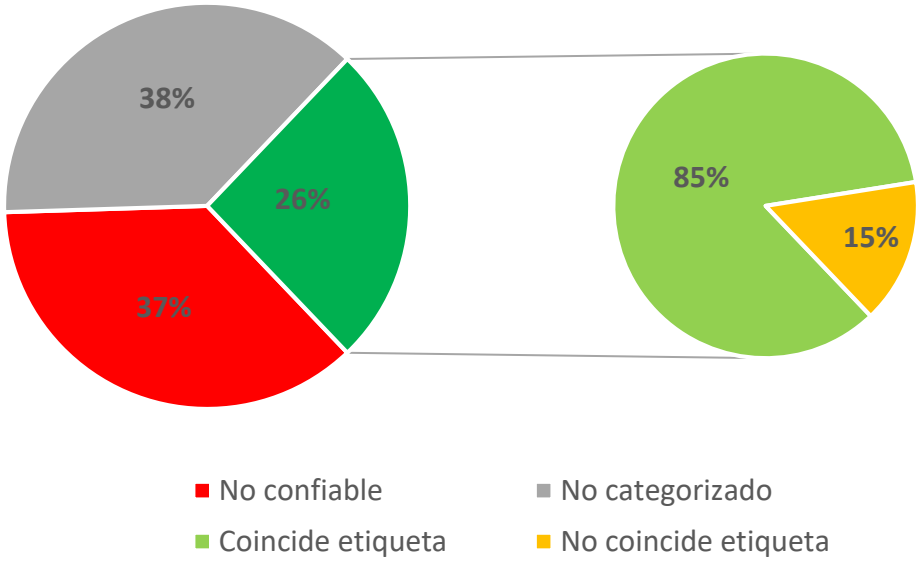


Figura 6.3 Porcentaje de compuestos clasificados en cada clase para log P_{liver}, luego de aplicar el método de DA.

6.2.3 Conclusiones

Hemos presentado un método que permite categorizar la predicción de un compuesto nuevo como *confiable*, *no confiable* o *no categorizada* para el caso de clasificación. Presentamos dos escenarios para evaluar la metodología: en el primero, utilizamos una base de datos de 202 compuestos para la cual *HIA* podía tomar dos posibles valores. En el segundo, la base de datos presentada tenía 122 compuestos con tres clases definidas para log P_{liver}. Los resultados resultaron más favorables para *HIA* que para log P_{liver}. Hay que tener en cuenta que la base de datos para este último caso es reducida en cantidad de compuestos, una de las

posibles razones por las que el porcentaje de compuestos con pocos vecinos ha sido más alto (37%). Si bien los resultados obtenidos son aceptables, creemos que puede mejorarse el rendimiento, en primera instancia, modificando la forma en que se calculan los compuestos vecinos y realizando pruebas con bases de datos que tengan un número más considerable de compuestos. Por otro lado, la técnica presentada es sensible al nivel de variabilidad en términos del espacio de descriptores que posee el conjunto de datos utilizado, lo cual influye en la determinación de los compuestos vecinos de un CN. Por ende, los criterios para determinar los umbrales que utiliza el método propuesto es otro tema central a seguir investigando.

En adición a esto, es importante mencionar que la utilización de métodos de DA nunca proporcionará una certeza absoluta, ya que un compuesto nuevo puede parecer estar dentro del DA definido y, sin embargo, la predicción podría ser poco fiable. Por el contrario, el compuesto puede parecer que está fuera del DA y, sin embargo la predicción podría ser confiable (OECD, 2007). Por lo tanto, el usuario del modelo debe ser consciente de que los métodos de DA, al igual que otras técnicas estadísticas, tales como la selección de descriptores, proporcionan un medio útil para apoyar decisiones basadas en el conocimiento del experto.

6.3 Sumario

En este capítulo se han presentado conceptos relacionados con el dominio de aplicación de un modelo QSAR/QSPR. Se resumieron diferentes técnicas utilizadas para su definición, en términos del espacio de los descriptores y en términos de la estimación de confianza de las predicciones. En este sentido, un aporte presentado aquí fue una metodología para determinar si un compuesto nuevo podrá predecirse de manera confiable para el caso de clasificación. Esta propuesta ha sido evaluada utilizando modelos predictivos para dos propiedades específicas.

Capítulo 7: Conclusiones

El objetivo general de esta tesis consistió en la propuesta e implementación de distintas técnicas computacionales que asistan al experto en etapas importantes del modelado QSAR/QSPR. Más específicamente, se buscó desarrollar metodologías computacionales que asistan al experto en el proceso de selección de los descriptores más relevantes para una propiedad o actividad biológica, y en la determinación del dominio de aplicación de modelos QSAR/QSPR de clasificación. Estas metodologías han sido evaluadas y validadas con el desarrollo de una variedad de modelos para distintas propiedades de interés en las áreas de diseño racional de fármacos, diseño de materiales poliméricos y medioambiente.

La selección de descriptores constituye una etapa muy importante del modelado, ya que implica aplicar diferentes estrategias para obtener un conjunto reducido de descriptores a partir de un universo muy grande de ellos. Existen muchas técnicas automatizadas para realizar esta tarea, discutidas en el capítulo 4. Esas metodologías no evalúan ciertos aspectos deseables de los descriptores, como por ejemplo la interpretabilidad de los mismos en términos físicos químicos. Como principal aporte en este sentido, se desarrolló VIDEAN, una herramienta de analítica visual para dar soporte al experto, permitiendo realizar un análisis exploratorio e interactivo de los datos.

Por otro lado, cuando tenemos modelos QSAR/QSPR es deseable conocer el dominio para el cuál ese modelo realizará las predicciones de una manera confiable. Esto no constituye una tarea sencilla, ya que se deben identificar los límites de ese modelo. Muchos enfoques han sido implementados para determinar el dominio de aplicación y aplicados en modelos de regresión, como se detalla en el capítulo 6. Sin embargo, existen escasos enfoques presentados para el caso de la clasificación, por lo que en el marco de esta tesis presentamos

la idea de una metodología para la determinación del dominio de aplicación para modelos de clasificación.

Por otro lado, como una manera de validar las estrategias implementadas se han desarrollado múltiples modelos QSAR/QSPR de regresión y clasificación, los cuales representan también un aporte en cada una de las tres áreas de aplicación exploradas en esta tesis.

A partir de los resultados obtenidos, se puede concluir que las ideas indagadas en esta tesis representan una valiosa iniciativa en el área del modelado QAR/QSPR, brindado soporte al experto en el proceso de desarrollo de un modelo. En el resto del capítulo se expondrán, a manera de resumen, las principales contribuciones de esta tesis, junto con sugerencias de extensiones y mejoras del trabajo realizado, así como también ideas para realizar futuras investigaciones.

7.1 Contribuciones

Las principales contribuciones de esta tesis se dan en el contexto del modelado QSAR/QSPR, tanto en la implementación de técnicas computacionales que brindan apoyo al modelador, como en el aporte de modelos para predecir distintas propiedades de interés que serán detalladas más adelante.

Selección de descriptores

La selección de un subconjunto de descriptores moleculares durante el diseño de un modelo QSAR/QSPR es una tarea difícil, donde se deben cumplir con varios objetivos simultáneos. Muchos métodos de selección utilizados para hacer frente a este problema se centran en las relaciones estadísticas entre los descriptores y la propiedad objetivo, dejando de lado los aspectos relacionados con el conocimiento químico. Por lo tanto, la interpretabilidad y la generalidad de los

modelos obtenidos por estos métodos se ven afectados drásticamente. Es decir que se requiere de una estrategia que incorpore el conocimiento experto en el proceso de selección con el fin de mejorar la confianza del usuario en el conjunto final de descriptores. A partir de esto, es que surge la propuesta y posterior implementación de VIDEAN, ampliamente detallada en el capítulo 5. Esta herramienta permite cargar diferentes subconjuntos de descriptores para que sean analizados de manera exploratoria e interactiva, permitiendo analizar cantidades considerables de datos con un mínimo esfuerzo cognitivo del usuario. A través de diferentes casos de estudio se demostró su potencial para dar soporte al experto en el desarrollo de modelos QSAR/QSPR, permitiendo mejorar la confiablidad de éstos, debido a la intervención del modelador en la etapa de selección de descriptores.

Dominio de Aplicación

En el modelado QSAR/QSPR, la definición del dominio de aplicación es una cuestión de mucha importancia, ya que a través de su definición se incrementa la confianza en las predicciones y se permite hacer un uso más práctico de los modelos. Es decir que, si tenemos un modelo con su dominio de aplicación definido, podremos saber para un compuesto nuevo si las predicciones arrojadas por el modelo serán confiables (el compuesto está dentro del dominio) o no (caso contrario). En este sentido, se realizó un planteo y posterior implementación de una primera idea para determinar el dominio de aplicación en modelos de clasificación. El enfoque utilizado se basa en evaluar la similaridad estructural entre un compuesto nuevo y los compuestos del conjunto de entrenamiento del modelo, con el fin de determinar el conjunto de moléculas más representativas (vecinos) del nuevo compuesto. Luego se analizan las etiquetas de este conjunto de vecinos y se clasifica el nuevo compuesto como *confiable*, *no confiable* o *no categorizado*.

Desarrollo de modelos QSAR/QSPR

Se han desarrollado numerosos modelos de regresión y clasificación para propiedades/actividades biológicas de interés en tres campos de aplicación, utilizando distintas herramientas computacionales: DRAGON (DRAGON, 2007) y CODES (Dorronsoro et al., 2004) para al cálculo de los descriptores; DELPHOS (Soto et al., 2009), TSAR (TSAR, 2000) y VIDEAN (Martínez et al., 2015) para la selección y aprendizaje de descriptores; y WEKA (M. Hall et al., 2009) para la generación y evaluación de modelos. En el proceso de desarrollo de todos los modelos siempre se priorizaron cuestiones fundamentales como un buen rendimiento estadístico, baja cardinalidad, interpretabilidad de sus descriptores en relación a la propiedad, y por ende la generalidad del modelo. Los mismos se encuentran detallados en el capítulo 5 y serán resumidos a continuación:

Diseño racional de fármacos: se estudiaron dos propiedades de mucha relevancia en el desarrollo de fármacos: *HIA (*Human Intestinal *Absorption)* y *BBB (*Blood–Brain Barrier). El conjunto de datos para *HIA* fue extraído de Guerra et al. (2010) y el de *BBB* de Guerra et al. (2008). Para ambas propiedades se desarrollaron modelos de regresión y clasificación. Para el caso de clasificación el modelo obtenido para *HIA* permite determinar si una molécula ha sido absorbida o no, mientras que para *BBB* se definieron tres clases: moléculas que cruzan la barrera hematoencefálica, moléculas que no atraviesan la barrera hematoencefálica y una zona gris que representa incertidumbre.

A partir de los modelos mencionados anteriormente, que fueron obtenidos utilizando técnicas de selección de descriptores, se ha realizado una contrastación entre estos modelos versus aquellos obtenidos utilizando técnicas de aprendizaje de descriptores. De una manera similar y dando un paso más hacia adelante, se han reportado nuevos modelos que surgen de una selección híbrida de descriptores, es decir que contemplan descriptores seleccionados por las dos técnicas mencionadas anteriormente.

Diseño de materiales poliméricos (plásticos): en este campo se han estudiado propiedades mecánicas de los polímeros derivadas del ensayo de tensión, que brindan información acerca de la ductilidad, resistencia y rigidez de un material polimérico; y que junto con otras propiedades permiten definir su perfil de aplicación estructural. La base de datos utilizada fue extraída de Palomba et al. (2014). Una de las propiedades estudiadas fue la e*longación a la rotura,* una propiedad mecánica que brinda información referente a la ductilidad de un material polimérico. En torno a ella se han desarrollado modelos de regresión y de clasificación. Para el caso de clasificación, el modelo permite predecir si un material será dúctil o no, previo a su síntesis. Otra de las propiedades que abordamos fue la *resistencia a la rotura* que brinda información acerca de la resistencia de un material. Para este caso se han desarrollado una gran variedad de modelos de regresión.

Para estas dos propiedades, del mismo modo que para las propiedades *HIA* y *BBB*, se ha realizado una contrastación entre estos modelos versus aquellos obtenidos utilizando técnicas de aprendizaje de descriptores. Así como también se han desarrollado modelos utilizando una estrategia híbrida para la selección de descriptores, es decir que están formados por descriptores obtenidos por las dos técnicas mencionadas anteriormente.

Finalmente, la propiedad *módulo elástico*, que aporta información relacionada a la rigidez de un material, ha sido abordada de manera inicial. Presentando diversos subconjuntos de descriptores posibles y llegando a obtener un modelo de regresión con buen rendimiento e interpretabilidad.

Medioambiente: en esta área se estudió una propiedad llamada log P_{liver}, que es el coeficiente de partición sangre-hígado para los compuestos orgánicos volátiles (VOCs). La base de datos de VOCs fue extraída de Abraham et al. (2007). También para log Pliver se desarrollaron modelos tanto de regresión como de clasificación. Para el caso de clasificación, el modelo permite identificar si un compuesto tiene afinidad con la sangre, afinidad con el hígado o la misma

afinidad por ambos medios, como paso preliminar a la evaluación de su potencialidad tóxica.

7.2 Trabajos Futuros

Como una extensión natural al trabajo realizado en esta tesis, se pretende seguir mejorando las metodologías propuestas. Para la selección de descriptores, más precisamente en relación a la herramienta VIDEAN, quedaron algunos aspectos a considerar que resultaría de utilidad para el proceso general de desarrollo de un modelo, tales como permitir que se puedan analizar también los compuestos (no sólo los descriptores) de manera individual o por familias. También desde el punto de vista de la evaluación de los descriptores en un determinado modelo, se prevé incorporar nuevas técnicas para los casos de clasificación ya que al momento sólo se encuentran disponibles técnicas para regresión. Por otra parte, para mejorar la detección del dominio de aplicación quedan realizar muchas pruebas aún, e incorporar estrategias adicionales para mejorar el cálculo de los límites del modelo. En este sentido, se prevé explorar en profundidad la teoría de predicción conformacional (*conformal prediction*) que permite determinar niveles precisos de confianzas en nuevas predicciones, independientemente del método con el que haya sido entrenado el modelo (Mathea et al., 2016; Norinder et al., 2016). Finalmente, pensamos seguir profundizado en el análisis de las propiedades presentadas en esta tesis, e impulsar nuevas experimentaciones y el estudio de otras propiedades/actividades.

Referencias

Abraham, M. H., Ibrahim, A., y Acree, W. E. (2007). Air to liver partition coefficients for volatile organic compounds and blood to liver partition coefficients for volatile organic compounds and drugs. *European Journal of Medicinal Chemistry*, *42*(6), 743-751.

Abraham, M. H., y Weathersby, P. K. (1994). Hydrogen bonding. 30. Solubility of gases and vapors in biological liquids and tissues. *Journal of Pharmaceutical Sciences*, *83*(10), 1450-1456.

Afantitis, A., Melagraki, G., Makridima, K., Alexandridis, A., Sarimveis, H., y Iglessi-markopoulou, O. (2005). Prediction of high weight polymers glass transition temperature using RBF neural networks. *Journal of molecular*, *716*, 193-198. https://doi.org/10.1016/j.theochem.2004.11.021

Alexander, D., Tropsha, A., y Winkler, D. A. (2015). Beware of R2: simple, unambiguous assessment of the prediction accuracy of QSAR and QSPR models. *Journal of Chemical Information and Modeling*, *55*(7), 1316-1322.

Ali, H. M., y Ali, I. H. (2015). QSAR and mechanisms of radical scavenging activity of phenolic and anilinic compounds using structural, electronic, kinetic, and thermodynamic parameters. *Medicinal Chemistry Research*, *24*(3), 987-998.

Alyass, A., Turcotte, M., y Meyre, D. (2015). From big data analysis to personalized medicine for all: challenges and opportunities. *Bmc Medical Genomics*, *8*. https://doi.org/ARTN 33\r10.1186/s12920-015-0108-y

Andrienko, N., y Andrienko, G. (2013). A visual analytics framework for spatio-temporal analysis and modelling. *Data Mining and Knowledge Discovery*, *27*(1), 55-83.

Attwood, T., Pettifer, S., y Thorne, D. (2016). Bioinformatics Challenges at the Interface of Biology and Computer Science: Mind the Gap.

Audi, R. (1999). *Ockham's razor, the Cambridge dictionary of philosophy* (2nd ed.). Cambridge: Cambridge University Press.

Balaz, S., y Lukacova, V. (1999). A Model-based Dependence of the Human Tissue/Blood Partition Coefficients of Chemicals on Lipophilicity and Tissue Composition. *Molecular Informatics*, *18*(4), 361-368.

Basak, S. C., Vracko, M., y Bhattacharjee, A. K. (2015). Big Data and New Drug Discovery: Tackling «Big Data» for Virtual Screening of Large Compound Databases. *Current Computer - Aided Drug Design*, *11*(3), 197-201.

Basant, N., Gupta, S., y Singh, K. P. (2016). Predicting the acute neurotoxicity of diverse organic solvents using probabilistic neural networks based QSTR modeling approaches. *NeuroToxicology*, *53*, 45-52.

Baumann, K., Ecker, G. F., Mestres, J., y Schneider, G. (2014). Molecular Informatics Going Fully Online. *Molecular Informatics*, *33*(1), 2-2. https://doi.org/10.1002/minf.201480131

Bezdek, J. C., Ehrlich, R., y Full, W. (1984). FCM: The fuzzy c-means clustering algorithm. *Computers & Geosciences*, *10*(2-3), 191-203.

Bishop, C. (2006). *Pattern Recognition and Machine Learning*.

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5-32. https://doi.org/10.1023/A:1010933404324

Breiman, L., Friedman, J., Stone, C. J., y Olshen, R. A. (1984). *Classification and regression trees*. CRC press.

Briesemeister, S., Rahnenführer, J., y Kohlbacher, O. (2012). No longer confidential: estimating the confidence of individual regression predictions. *PloS one*, *7*(11), e48723.

Buist, H. E., Lianne, W.-B., Bouwman, T., y Vaes, W. H. J. (2012). Predicting blood: air partition coefficients using basic physicochemical properties. *Regulatory Toxiology and Pharmacology*, *62*(1), 23-28.

Callister, W. D., y Rethwisch, D. G. (2011). *Materials science and engineering: an introduction* (Vol. 5). NY: John Wiley & Sons.

Cao, D.-S., Xu, Q.-S., Liang, Y.-Z., Chen, X., y Li, H.-D. (2010). Automatic feature subset selection for decision tree-based ensemble methods in the prediction of bioactivity. *Chemometrics and Intelligent Laboratory Systems*, *103*(2), 129-136.

Carpendale, S. (2008). Evaluating information visualizations. En *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (Vol. 4950 LNCS, pp. 19-45). Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-70956-5_2

Castaño, T., Encinas, A., Pérez, C., Castro, A., Campillo, N., y Gil, C. (2008). Design, synthesis, and evaluation of potential inhibitors of nitric oxide synthase. *Bioorganic & Medicinal Chemistry*, *16*(1), 6193-6206.

Cherkasov, A., Muratov, E. N., Fourches, D., Varnek, A., Baskin, I. I., Cronin, M., … Tropsha, A. (2014). QSAR modeling: where have you been? Where are you going to? *Journal of medicinal chemistry*, *57*(12), 4977-5010. https://doi.org/10.1021/jm4004285

CODESSA. (2005). CODESSA PRO. Recuperado a partir de http://www.codessa-pro.com/

Cover, T. M., y Thomas, J. A. (1991). Entropy, relative entropy and mutual information. *Elements of information theory*, *2*, 1-55.

Cover, T. M., y Thomas, J. A. (2012). *Elements of information theory*. John Wiley & Sons.

Cravero, F., Martínez, M. J., Díaz, M. F., y Ponzoni, I. (2017). QSAR Classification

Models for Predicting Affinity to Blood or Liver of Volatile Organic Compounds in e-Health. *Lecture Notes in Computer Sciences.*, 1-10.

Cravero, F., Martínez, M. J., Vazquez, G. E., Díaz, M. F., y Ponzoni, I. (2016). Feature Learning applied to the Estimation of Tensile Strength at Break in Polymeric Material Design. *Journal of Integrative Bioinformatics*, *13*, 286-301.

Cravero, F., Vazquez, G. E., Díaz, M. F., y Ponzoni, I. (2015). Modelado QSPR de Propiedades Mecánicas de Materiales Poliméricos Empleando Técnicas de Reducción de Variables basadas en algoritmos de Aprendizaje Automático. En *In Proceeding of the Conference of Chemical Engineering*.

Curtis, R. E., Kinnaird, P., y Xing, E. P. (2011). GenAMap: Visualization strategies for structured association mapping. En *IEEE Symposium on Biological Data Visualization 2011, BioVis 2011 - Proceedings* (pp. 87-94). https://doi.org/10.1109/BioVis.2011.6094052

Danishuddin, y Khan, A. U. (2016). Descriptors and their selection methods in QSAR analysis: paradigm for drug design. *Drug Discovery Today*, *21*(8), 1291-1302.

Dashtbozorgi, Z., y Golmohammadi, H. (2010). Prediction of air to liver partition coefficient for volatile organic compounds using QSAR approaches. *European Journal of Medicinal Chemistry*, *45*(6), 2182-2190.

Dharma Rahadi, P. S. A., Shobirin, K. A., y Ariyani, S. (2016). Big Data Managemen. *International Journal of Engineering and Emerging Technology*, *1*(1).

Dietzsch, J., Heinrich, J., Nieselt, K., y Bartz, D. (2009). SpRay: A visual analytics approach for gene expression data. *2009 IEEE Symposium on Visual Analytics Science and Technology*, 179-186. https://doi.org/10.1109/VAST.2009.5333911

Dorronsoro, I., Chana, A., Abasolo, M. I., Castro, A., Gil, C., Martínez, A., … Martínez, A. (2004). CODES/Neural Network Model: a Useful Tool for in Silico Prediction of Oral Absorption and Blood-Brain Barrier Permeability of Structurally Diverse Drugs. *Molecular Informatics*, *23*(2-3), 89-98.

DRAGON. (2007). DRAGON 5.5. Milan, Italy: Talete srl.

Dragos, H., Gilles, M., Alexandre, V., Marcou, G., y Varnek, A. (2009). Predicting the predictability: A unified approach to the applicability domain problem of QSAR models. *Journal of Chemical Information and Modeling*, *49*(7), 1762-1776.

Duch, W. (2006). Filter methods. En *Feature Extraction* (pp. 89-117).

Forina, M., Armanino, C., Leardi, R., y Drava, G. (1991). A class-modelling technique based on potential functions. *Journal of Chemometrics*, *5*, 435-453.

Frank, E., Hall, M., Holmes, G., Kirkby, R., Pfahringer, B., Witten, I. H., y Trigg, L. (2005). Weka. *Weka. Data Mining and Knowledge Discovery Handbook*, 1305-1314.

Freese, N. H., Norris, D. C., y Loraine, A. E. (2016). Integrated genome browser: Visual analytics platform for genomics. *Bioinformatics*, *32*(14), 2089-2095. https://doi.org/10.1093/bioinformatics/btw069

Gajewicz, A., Rasulev, B., Dinadayalane, T. C., Urbaszek, P., Puzyn, T., Leszczynska, D., y Leszczynski, J. (2012). Advancing risk assessment of engineered nanomaterials: Application of computational approaches. *Advanced drug delivery reviews*, *64*(15), 1663-1693.

Ganguly, M., Brown, N., Schuffenhauer, A., Ertl, P., Gillet, V. J., y Greenidge, P. A. (2006). Introducing the consensus modeling concept in genetic algorithms: application to interpretable discriminant analysis. *Journal of Chemical Information and Modeling*, *46*(5), 2110-2124.

Gasteiger, J. Chemoinformatics: Achievements and challenges, a personal view, 21 Molecules § (2016). https://doi.org/10.3390/molecules21020151

Gerasch, A., Faber, D., Küntzer, J., Niermann, P., Kohlbacher, O., Lenhof, H.-P., y Kaufmann, M. (2014). BiNA : a visual analytics tool for biological network data. *PloS one*, *9*(2), e87397.

Ghose, A. K., Viswanandhan, V. N., y Wendoloski, J. J. (1998). Prediction of Hydrophobic (Lipophilic) Properties of Small Organic Molecules Using Fragmental Methods: An Analysis of ALOGP and CLOGP Methods. *The Journal of Physical Chemistry*, *102*(21), 3762-3772.

Goodarzi, M., Dejaegher, B., y Heyden, Y. Vander. (2012). Feature selection methods in QSAR studies. *Journal of AOAC International*, *95*(3), 636-651.

Gramatica, P. (2006). Principles of QSAR models validation: internal and external. *QSAR & Combinatorial Science*, *26*(5), 694-701.

Gramatica, P. (2010). Chemometric Methods and Theoretical Molecular Descriptors in Predictive QSAR Modeling of the Environmental Behavior of Organic Pollutants. *Recent Advances in QSAR Studies*, 327-366.

Gramatica, P., Chirico, N., Papa, E., Cassani, S., y Kovarich, S. (2013). Software News and Updates: {QSARINS}: a new software for the development, analysis, and validation of {QSAR MLR} models. *Journal of Computational Chemistry*, *34*(24), 2121-2132. https://doi.org/http://dx.doi.org/10.1002/jcc.23361

Gramatica, P., Pilutti, P., y Papa, E. (2003). Predicting the NO3 radical tropospheric degradability of organic pollutants by theoretical molecular descriptors. *Atmospheric Environment*, *37*(22), 3115-3124.

Green, T. M., Ribarsky, W., y Fisher, B. (2009). Building and applying a human cognition model for visual analytics. *Information visualization*, *8*(1), 1-13.

Guerra, A., Campillo, N., y Páez, J. A. (2010). Neural computational prediction of oral drug absorption based on CODES 2D descriptors. *European Journal of Medicinal Chemistry*, *45*(3), 930-940.

Guerra, A., Páez, J. A., y Campillo, N. (2008). Artificial Neural Networks in ADMET Modeling: Prediction of Blood – Brain Barrier Permeation. *Molecular Informatics*, *27*(5), 586-594.

Guo, Z., Ward, M. O., y A., R. E. (2009). Model space visualization for multivariate

linear trend discovery. En *Proceedings of the IEEE symposium on visual analytics science and technology VAST'09* (pp. 75-82).

Gütlein, M., Karwath, A., y Kramer, S. (2014). CheS-Mapper 2 . 0 for visual validation of (Q) SAR models. *Journal of cheminformatics*, 6(41), 1-18. https://doi.org/10.1186/s13321-014-0041-7

Guyon, I., y Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of machine learning research*, *3*, 1157-1182.

Hall, M. A., y Smith, L. A. (1998). *Practical feature subset selection for machine learning*.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., y Witten, I. H. (2009). The WEKA Data Mining Software: An Update. *ACM SIGKDD explorations newsletter*, *11*(1), 10-18.

Hanley, J. A. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, *143*(1), 29-36.

Hansch, C. (1969). A quantitative approach to biochemical structure-activity relationships. *Accounts of Chemical Research*, *2*(8), 232-239.

Hao, M. C., Janetzko, H., Mittelsadt, S., Hill, W., U., D., y Keim, D. A. (2011). A visual analytics approach for peak-preserving prediction of large seasonal time series. *Comput Graph Forum*, *30*(3), 691-700.

Hecht-Nielsen, R. (1988). Theory of the backpropagation neural network. *Neural Networks*, *1*(1), 445-448.

Hegarty, M. (2011). The cognitive science of visual-spatial displays: Implications for design. *Topics in cognitive science*, *3*(3), 446-474.

Hewitt, M., Ellison, C. M., Enoch, S. J., Madden, J. C., y Cronin, M. T. D. (2010). Integrating (Q)SAR models, expert systems and read-across approaches for the prediction of developmental toxicity. *Reproductive Toxicology*, *30*(1), 147-160.

Hinton, G. E., y Zemel, R. S. (1994). Autoencoders, minimum description length and Helmholtz free energy. En *Advances in neural information processing systems* (pp. 3-10).

HyperChem. (2006). HyperChem. Gainesville, FL, USA: Hypercube, Inc.

Jalali-Heravi, M., y Parastar, F. (2000). Use of artificial neural networks in a QSAR study of anti-HIV activity for a large group of HEPT derivatives. *Journal of chemical information and computer sciences*, *40*(1), 147-154.

Jaworska, J., Aldenberg, T., y Nikolova, N. (2005). Review of methods for assessing the applicability domains of SARs and QSARs. *Alternatives to Laboratory Animals*, *33*, 445-459.

Jaworska, J., Nikolova-Jeliazkova, N., y Aldenberg, T. (2005). QSAR applicability domain estimation by projection of the training set descriptor space: a review. *ATLA-NOTTINGHAM-*, *33*(5), 445.

Jouan-Rimbaud, D., Bouveresse, E., Massart, D. L., y de Noord, O. E. (1999). Detection of prediction outliers and inliers in multivariate calibration. *Analytica Chimica Acta*, *388*, 283-301.

Katritzky, A. R., Kuanar, M., Fara, D. C., Karelson, M., Acree Jr., W. E., Solov, V. P., y Varnek, A. (2005). QSAR modeling of blood:air and tissue:air partition coefficients using theoretical descriptors. *Bioorganic & Medicinal Chemistry*, *13*(23), 6450-6463.

Katritzky, A. R., Sild, S., Lobaniv, V., y Karelson, M. (1998). Quantitative structure property relationship (QSPR) correlation of glass transition temperatures of high molecular weight polymers. *Journal of chemical information and computer sciences*, *38*(2), 300-304.

Kaufman, L., y Rousseeuw, P. J. (2009). *Finding groups in data: an introduction to cluster analysis*. John Wiley & Sons.

Keefer, C. E., Kauffman, G. W., y Gupta, R. R. (2013). Interpretable, probability-based confidence metric for continuous quantitative structure–activity relationship models. *Journal of Chemical Information and Modeling*, *53*(2), 368-383.

Keim, D. A. (2001). Visual exploration of large data sets. *Communications of the ACM*, *44*(8), 38-44. https://doi.org/10.1145/381641.381656

Keim, D. A., Kohlhammer, J., Ellis, G., y Mansmann, F. (2010). *Mastering the Information Age Solving Problems with Visual Analytics*.

Keim, D. A., Mansmann, F., Schneidewind, J., y Ziegler, H. (2006). Challenges in visual data analysis. En *Information Visualization* (pp. 9-16).

Kodinariya, T. M., y Makwana, P. R. (2013). Review on determining number of Cluster in K-Means Clustering. *International Journal*, *1*(6), 90-95.

Kohavi, R., y John, G. H. (1997). Wrappers for feature subset selection. *Artificial Intelligence*, *97*(1-2), 273-324.

Kohonen, T. (1990). The self-organizing map. *Proceedings of the IEEE*, *78*(9), 1464-1480.

Kojadinovic, I. (2005). On the use of mutual information in data analysis: an overview. En *Proc Int Symp Appl Stochastic Models Data Anal* (pp. 738-747).

Krause, E. F. (2012). *Taxicab geometry: An adventure in non-Euclidean geometry*. Courier Corporation.

Krzanowski, W. (2000). *Principles of multivariate analysis*. OUP Oxford.

Lam, H., Bertini, E., Isenberg, P., Plaisant, C., y Carpendale, S. (2012). Empirical Studies in Information Visualization: Seven Scenarios. *Visualization and Computer Graphics, IEEE Transactions on*, *18*(9), 1520-1536. https://doi.org/10.1109/tvcg.2011.279

Levitin, D. J. (2014). *The organized mind: Thinking straight in the age of information overload*. Penguin.

Liu, H. X., Yao, X. J., Zhang, R. S., Liu, M. C., Hu, Z. D., y Fan, B. T. (2005). Prediction of the tissue/blood partition coefficients of organic compounds based on the molecular structure using least-squares support vector machines. *Journal of Computer-Aided Molecular Design*, *19*(7), 499-508.

Livingstone, D. J., Hesketh, G., y Clayworth, D. (1991). Novel method for the display of multivariate data using neural networks. *Journal of Molecular Graphics*, *9*(2), 115-118.

Loh, W.-Y. (2011). Classification and regression trees. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *1*(1), 14-23.

Mahalanobis, P. C. (1936). On the generalized distance in statistics. En *Proceedings of the National Institute of Sciences* (pp. 49-55).

Martin-Biosca, Y., TorreS-Cartas, S., Villanueva Camañas, R. M., Sagrado, S., y Medina Hernández, M. J. (2009). Biopartitioning micellar chromatography to predict blood to lung, blood to liver, blood to fat and blood to skin partition coefficients of drugs. *Analytica Chimica Acta*, *632*(2), 296-303.

Martínez, M. J., Cravero, F., Díaz, M. F., y Ponzoni, I. (2017). QSPR Modeling Applied to High Molecular Weight Polymers: Ductility Characterization from Elongation at Break. En *VIII International Symposium on Materials. MATERIAIS 2017*.

Martínez, M. J., Ponzoni, I., Díaz, M. F., Vazquez, G. E., y Soto, A. J. (2015). Visual analytics in cheminformatics: user-supervised descriptor selection for QSAR methods. *Journal of Cheminformatics*, *7*(39). https://doi.org/10.1186/s13321-015-0092-4

Mathea, M., Klingspohn, W., y Baumann, K. (2016). Chemoinformatic Classification Methods and their Applicability Domain. *Molecular Informatics*, *35*(5), 160-180.

Mitchell, T., y McGraw, H. (1997). *Machine Learning*.

Mosier, P. D., y Jurs, P. C. (2002). QSAR/QSPR studies using probabilistic neural networks and generalized regression neural networks. *Journal of Chemical Information and Computer Sciences*, *42*(6), 1460-1470.

Munzner, T. (2009). A Nested Process Model for Visualization Design and Validation. *IEEE Transactions on Visualization and Computer Graphics*, *15*(6), 921-928. https://doi.org/10.1109/TVCG.2009.111

Netzeva, T. I., Worth, A. P., Aldenberg, T., Benigni, R., Cronin, T. D. M., Gramatica, P., … Yang, C. (2005). Current Status of Methods for Defining the Applicability Domain of (Quantitative) Structure–Activity Relationships. *ATLA*, *33*, 155-173.

Niculescu, S. P. (2003). Artificial Neural Networks and Genetic Algorithms in QSAR. *Journal of Molecular Structure*, *622*(1), 71-83.

Nikolova, N., y Jaworska, J. S. (2003). Approaches to measure chemical similarity: a review. *Molecular Informatics*, *22*(9-10), 1006-1026.

Norinder, U., Rybacka, A., y Andersson, P. L. (2016). Conformal prediction to define

applicability domain – A case study on predicting ER and AR binding. *SAR and QSAR in Environmental Research*, *27*(4), 303-316.

OECD. (2007). Guidance document on the validation of (quantitative) structure-activity relationships [(Q) SAR] models. *OECD Series on Testing and Assessment No. 69. ENV/JM/MONO (2007) 2*, *154*.

Palczewska, A., Neagu, D., y Ridley, M. (2013). Using Pareto points for model identification in predictive toxicology. *Journal of Cheminformatics*, *5*(1), 16.

Palomba, D., Martínez, M. J., Ponzoni, I., Díaz, M. F., Vazquez, G. E., y Soto, A. J. (2012). QSPR models for predicting log pliver values for volatile organic compounds combining statistical methods and domain knowledge. *Molecules2le*, *17*(12), 14937-14953.

Palomba, D., Martínez, M. J., Ponzoni, I., Vazquez, G. E., y Soto, A. J. (2012). QSPR Models for Predicting Log Pliver Values for Volatile Organic Compounds Combining Statistical Methods and Domain Knowledge. *Molecules*, 14937-14953.

Palomba, D., Vazquez, G. E., y Díaz, M. F. (2014). Prediction of Elongation at Break for Linear Polymers. *Chemometrics and Intelligent Laboratory Systems*, *139*, 121-131. https://doi.org/10.1016/j.chemolab.2014.09.009

Partl, C., Lex, A., Streit, M., Strobelt, H., Wassermann, A. M., Pfister, H., y Schmalstieg, D. (2014). ConTour: Data-driven exploration of multi-relational datasets for drug discovery. *IEEE Transactions on Visualization and Computer Graphics*, *20*(12), 1883-1892. https://doi.org/10.1109/TVCG.2014.2346752

Picard, R. R., y Cook, R. D. (1984). Cross-Validation of Regression Models. *journal of the American Statistical Association*, *79*(387), 575-583.

Platt, J. R. (1947). Influence of Neighbor Bonds on Additive Bond Properties in Paraffins. *The Journal of Chemical Physics*, *15*(6), 419-420.

Ponzoni, I., Sebastian, V., Requena, C., Roca, C., Martínez, M. J., Cravero, F., … Campillo, N. E. (2017). Hybridizing Feature Selection and Feature Learning Approaches in QSAR Modeling for Drug Discovery. *Scientific Reports*, *7*, 2404.

Poulin, P., y Theil, F. P. (2000). A priori prediction of tissue:plasma partition coefficients of drugs to facilitate the use of physiologically-based pharmacokinetic models in drug discovery. *Journal of Pharmaceutical Sciences*, *89*(1), 16-35.

Preparata, F. P., y Shamos, M. I. (1991). Convex Hulls: Basic Algorithms. En F. P. Preparata y M. I. Shamos (Eds.), *Computational Geometry: An Introduction* (pp. 95–148). New York, NY, USA: Springer-Verlag.

Puzyn, T., Leszczynski, J., y Cronin, M. T. (2010). *Recent advances in QSAR studies: methods and applications*. Springer Science & Business Media.

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, *1*(1), 81-106.

Quinlan, J. R. (1987). Simplifying decision trees. *International journal of man-machine studies*, *27*(3), 221-234.

Rhyne, T. M. (2003). Does the difference between information and scientific visualization really matter? *IEEE Computer Graphics and Applications*, *23*(3), 6-8.

Rhyne, T. M., Tory, M., Munzner, T., Worcester, M. W., Johnson, C., y Laidlaw, D. H. (2003). Information and Scientific Visualization: Separate but Equal or Happy Together at Last. En *IEEE Visualization* (pp. 611-614).

Riedmiller, M. (1994). Advanced supervised learning in multi-layer perceptrons—from backpropagation to adaptive learning algorithms. *Computer Standars & Interfaces*, *16*(3), 265-278.

Roberts, D. W., y Costello, J. (2003). QSAR and Mechanism of Action For Aquatic Toxicity of Cationic Surfactants. *Molecular Informatics*, *22*(2), 220-225.

Rodgers, T., Leahy, D., y Rowland, M. (2005). Physiologically based pharmacokinetic modeling 1: Predicting the tissue distribution of moderate-to-strong bases. *Journal of Pharmaceutical Sciences*, *94*(6), 1259-1276.

Rogers, D. J., y Tanimoto, T. T. (1960). A computer program for classifying plants. *Science*, *132*, 1115-1118.

Roy, K., Kar, S., y Das, R. N. (2015). *A primer on QSAR/QSPR modeling: Fundamental concepts*. Springer.

Sadana, R., y Stasko, J. (2016). Designing Multiple Coordinated Visualizations for Tablets. *Computer Graphics Forum*, *35*(3), 261-270.

Sager, J. E., Yu, J., Raguenau-Majlessi, I., y Isoherranen, N. (2015). Physiologically based pharmacokinetic (PBPK) modeling and simulation approaches: a systematic review of published models, applications and model verification. *Drug Metabolism and Disposition*, *43*, 1823-1837.

Sahigara, F., Mansouri, K., Ballabio, D., Mauri, A., y Consonni, Viviana Todeschini, R. (2012). Comparison of Different Approaches to Define the Applicability Domain of QSAR Models. *Molecules*, *17*, 4791-4810.

Saíz-Urra, L., Pérez González, M., y Teijeira, M. (2006). QSAR studies about cytotoxicity of benzophenazines with dual inhibition toward both topoisomerases I and II: 3D-MoRSE descriptors and statistical considerations about variable selection. *Bioorganic & Medicinal Chemistry*, *12*(21), 7347-7358.

Schultz, W. T., Cronin, M. T. D., Netzeva, T. I., y Aptula, A. O. (2002). Structure-toxicity relationships for aliphatic chemicals evaluated with Tetrahymena pyriformis. *Chemical research in toxicology*, *15*(12), 1602-1609.

Schuur, J. H., Selzer, P., y Gasteiger, J. (1996). The Coding of the Three-Dimensional Structure of Molecules by Molecular Transforms and Its Application to Structure-Spectra Correlations and Studies of Biological Activity. *Journal of Chemical Information and Computer Sciences*, *36*(2), 334-344.

Soto, A. J., Cecchini, R. L., Vazquez, G. E., y Ponzoni, I. (2009). Multi-objective feature selection in QSAR using a machine learning approach. *QSAR & Combinatorial Science*, *28*(11), 1509–1523. https://doi.org/10.1002/qsar.200960053

Soto, A. J., Ponzoni, I., y Vazquez, G. E. (2009). Segregating confident predictions of chemicals' properties for virtual screening of drugs. *Distributed Computing, Artificial Intelligence, Bioinformatics, Soft Computing, and Ambient Assisted Living*, 1005-1012.

Soto, A. J., Vazquez, G. E., Strickert, M., y Ponzoni, I. (2011). Target-Driven Subspace Mapping Methods and Their Applicability Domain Estimation. *Molecular Informatics*, *30*(9), 779-789.

Srivastava, N., Hinton, G. E., Krizhevsky, A., Sutskever, I., y Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, *15*(1), 1929-1958.

Thomas, J. J., y Cook, K. A. (2005). *Illuminating the Path: The Research and Development Agenda for Visual Analytics*.

Todeschini, R., y Consonni, V. (2008). *Handbook of molecular descriptors*. Weinheim, Germany: John Wiley & Sons.

Tong, W., Xie, Q., Hong, H., Shi, L., Fang, H., y Perkins, R. (2004). Assessment of prediction confidence and domain extrapolation of two structure-activity relationship models for predicting estrogen receptor binding activity. *Enviromental health perspectives*, 1249-1254.

TSAR. (2000). TSAR. Ltd, Oxford Molecular.

Turkey, J. W. (1977). *Exploratory Data Analysis*.

Utracki, L. A., y Wilkie, C. A. (2002). *Polymer blends handbook*. Dordrecht: Kluwer academic publishers.

Vallero, D. A. (2007). *Fundamentals of Air Pollution* (4th ed. Ac). San Diego, CA, USA.

Van Krevelen, D. W. (2009). *Properties of Polymers* (volume 4). Elseiver.

Viswanandhan, V. N., Ghose, A. K., Revankar, G. R., y Robins, R. K. (1989). Atomic physicochemical parameters for three dimensional structure directed quantitative structure-activity relationships. 4. Additional parameters for hydrophobic and dispersive interactions and their application for an automated superposition of certain . *Journal of Chemical Information and Computer Sciences*, *29*(3), 163-172.

Viswanandhan, V. N., Reddy, R. M., Bacquet, R. J., y Erion, M. D. (1993). Assessment of Methods Used for Predicting Lipophilicity: Application to Nucleosides and Nucleoside Bases. *Journal of Computational Chemistry*, *14*(9), 1019-1026.

Vork, K., Carlisle, J., y Brown, J. P. (2013). Estimating Workplace Air and Worker Blood Lead Concentration using an Updated Physiologically-based Pharmacokinetic (PBPK) Model: Office of Environmental Health Hazard Assessment. *California Environmental Protection Agency*.

Wang, Y., y Witten, I. H. (1996). *Induction of model trees for predicting continuous classes*. Hamilton, New Zealand.

Ward, I. M., y Sweeney, J. (2012). *Mechanical properties of solid polymers.* John Wiley & Sons.

Weininger, D. (1988). SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of chemical information and computer sciences*, *28*(1), 31-36.

Wetzel, S., Klein, K., Renner, S., Rauh, D., y Oprea, T. (2009). Interactive exploration of chemical space with Scaffold Hunter. *Nature chemical*.

Williams, J., y Ralf, K. (2007). Volatile Organic Compounds in the Atmosphere: An Overview. *Volatile Organic Compounds in the Atmosphere*, 1-32.

Witten, I. H., Frank, E., Hall, M. A., y Pal, C. J. (2016). *Data Mining: Practical machine learning tools and techniques*. (M. Kaufmann, Ed.).

Wold, S., Esbensen, K., y Geladi, P. (1987). Principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, *2*(1-3), 37-52.

Wong, P. C., y Thomas, J. (2004, septiembre). Visual analytics. *IEEE Computer Graphics and Applications*, *24*(5), 20-21. https://doi.org/10.1109/MCG.2004.39

Woodruff, T. J., Burke, T. A., y Zeise, L. (2011). The Need for Better Public Health Decisions on Chemicals Released Into Our Environment, *30*(957-967).

Xiaofeng, M., y Xiang, C. (2013). Big data management: concepts, techniques and challenges [J]. *Journal of Computer Research and Development*, *1*, 98.

Yang, J., y Honavar, V. (1998). Feature subset selection using a genetic algorithm. *IEEE Intelligent Systems and their Applications*, *13*(2), 44-49.

Yap, C. W. (2011). PaDEL-Descriptor: An open source software to calculate molecular descriptors and fingerprints. *Journal of Computational Chemistry*, *32*(7), 1466-1474.

Zhang, H. (2004). A new nonlinear equation for the tissue/blood partition coefficients of neutral compounds. *Journal of Pharmaceutical Sciences*, *93*(6), 1595-1604.

Zhang, H., y Zhang, Y. (2006). Convenient Nonlinear Model for Predicting the Tissue/Blood Partition Coefficients of Seven Human Tissues of Neutral, Acidic, and Basic Structurally Diverse Compounds. *Journal of medicinal chemistry*, *49*(19), 5815-5829.

Zhang, Q., Hughes-Oliver, J. M., y Ng, R. T. (2009). A model-based ensembling approach for developing QSARs. *Journal of Chemical Information and Modeling*, *49*(8), 1857.