UNIVERSIDAD NACIONAL DEL SUR

TESIS DOCTORAL EN CIENCIAS DE LA COMPUTACIÓN

MODELADO PREDICTIVO DE SISTEMAS COMPLEJOS PARA INFORMÁTICA MOLECULAR: DESARROLLO DE MÉTODOS DE SELECCIÓN Y APRENDIZAJE DE CARACTERÍSTICAS EN PRESENCIA DE INCERTIDUMBRE

FIORELLA CRAVERO

BAHÍA BLANCA ARGENTINA

2020

PREFACIO

Esta tesis es presentada como parte de los requisitos para optar por el 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 e Ingeniería de la Computación de la Universidad Nacional del Sur, durante el período comprendido entre el 16 de Julio de 2015 y el 12 de Diciembre de 2019, bajo la dirección conjunta del Dr. Ignacio Ponzoni, Profesor Asociado de la Universidad Nacional del Sur e Investigador Independiente de CONICET, y la Dra. Mónica Fátima Díaz, Profesora Adjunta de la Universidad Nacional del Sur e Investigadora Adjunta de CONICET.

Fiorella Cravero
Bahía Blanca, 12 de diciembre de 2019

UNIVERSIDAD NACIONAL DEL SUR

SECRETARÍA GENERAL DE POSGRADO Y EDUCACIÓN CONTINUA

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

AGRADECIMIENTOS

Estas líneas, pensadas para expresar gratitud a quienes de una u otra manera fueron parte de la elaboración de esta tesis, no tendrían lugar sin la educación pública y el sistema de becas para investigación que tenemos en Argentina. Razón por la cual, en primer lugar, quiero agradecer a la Universidad Nacional del Sur, a la Comisión de Investigaciones Científicas de la Provincia de Buenos Aires (CIC), al Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), y a Planta Piloto de Ingeniería Química (PLAPIQUI) por brindarme los medios necesarios para el desarrollo de esta tesis, renovando mi compromiso para la defensa de la educación pública, laica, gratuita y de calidad, en todos los niveles.

Genuina y emocionadamente, quiero agradecerles al Dr. Ignacio Ponzoni y a la Dra. Mónica F. Díaz, mis directores, por la oportunidad brindada y el privilegio que tuve y tengo de aprender de ellos. Acompañaron y guiaron este camino con gran rigurosidad técnica y con una calidad y calidez humana infrecuentes. Los admiro profesionalmente y los aprecio como personas. Gracias. También, quiero expresar mi agradecimiento a todo el grupo de trabajo, al Dr. Gustavo E. Vázquez, a la Dra. M. Jimena Martínez y al Ing. Santiago A. Schustik, por sus incuestionables aportes.

Párrafo aparte para mis compañeros de oficina. Por la paciencia, por el aliento, por la confianza mutua, por hacer de la oficina un lugar seguro, por ser parte de mis días. Gracias.

A mi familia y a amigos, por transformar el caos en orden, por trascender mis días y ser parte de mi vida. Gracias infinitas.

Finalmente, quiero agradecer a los puenteros, a los chicos y a las vecinas de VTS, a los profes del apoyo escolar y a quienes se siguen sumando a El Puente; por la posibilidad de generar una pertenencia a algo que va más allá de una misma y esta agenda hegemónica de investigación, logrando una sensación de equilibrio y paz. Gracias por construir juntos el sueño colectivo.

Para Ana Lia, Leo, Ulises y
les amigues, que son familia.

RESUMEN

En la actualidad existe una necesidad creciente de guiar el descubrimiento *in silico* de nuevos polímeros industriales mediante enfoques de Aprendizaje Maquinal supervisado que identifiquen correlaciones estructura-propiedad a partir de la información contenida en bases de datos de materiales, donde cada uno de estos está caracterizado mediante Descriptores Moleculares (DMs). Estas correlaciones se conocen como Modelos de Relación Cuantitativa Estructura-Actividad/Propiedad (QSAR/QSPR, por las siglas en inglés de *Quantitative Structure-Activity/Property Relationship*) y pueden ser empleadas para predecir propiedades de interés previo a la etapa de síntesis química, contribuyendo de este modo a acelerar el diseño de nuevos materiales y reducir sus costos de desarrollo.

El modelado QSAR/QSPR ya ha sido ampliamente empleado en Informática Molecular para el Diseño Racional de Fármacos asistido por computadoras. Sin embargo, los materiales poliméricos son significativamente más complejos que las moléculas pequeñas como las drogas, dado que están integrados por colecciones de macromoléculas compuestas por miles de cadenas que, a su vez, se forman por la unión de cientos de miles de Unidades Repetitivas Estructurales (UREs). Estas cadenas poseen diferentes pesos moleculares (o largos de cadena) y, a su vez, aparecen con distintas frecuencias dentro de cada material. Este fenómeno, conocido como polidispersión, es la principal razón de que muchas aproximaciones informáticas desarrolladas para el diseño racional de fármacos no sean directamente aplicables, ni lo suficientemente efectivas, en el ámbito de la Informática de Polímeros.

El objetivo general de esta tesis es contribuir con soluciones para distintas cuestiones relativas a la representación computacional y algoritmia que surgen durante el modelado QSPR de propiedades de polímeros polidispersos de alto peso molecular, con especial énfasis en el tratamiento del problema de selección de descriptores moleculares. Las variaciones en la frecuencia de las cadenas de diferentes largos hacen que la descripción de la estructura de un material polimérico contenga incertidumbre, en contraste con lo que sucede en la caracterización estructural típica de una molécula pequeña. No obstante esto, debido a la complejidad de modelar esta incertidumbre, la mayoría de los estudios QSAR/QSPR han utilizado hasta ahora modelos moleculares simples y univaluados, es decir, calculan los descriptores moleculares para una única instancia de peso, de entre todas las posibles cadenas que conforman un material. En particular, la casi totalidad de estos estudios usan descriptores calculados sobre una única URE, sin tener en cuenta la polidispersión. En tal sentido, esta tesis propone investigar

distintas alternativas de selección y aprendizaje de características para modelado QSPR con incertidumbre, que exploren la efectividad de otras representaciones computacionales más realistas para los materiales poliméricos.

En primer lugar, se presenta una metodología híbrida que emplea tanto algoritmos de Selección de Características como de Aprendizaje de Características, a fin de evaluar la máxima capacidad predictiva que se puede alcanzar con la tradicional representación univaluada URE. En segundo lugar, se proponen nuevas representaciones univaluadas, basadas en pesos moleculares promedios, denominadas como modelos moleculares Mn y Mw, cuyas capacidades para inferir modelos QSPR son contrastadas con el modelo molecular URE.

La siguiente alternativa propuesta estudia una representación computacional trivaluada, basada en la integración de los modelos moleculares univaluados URE, Mn y Mw en una única base de datos, la cual permite capturar parcialmente el fenómeno de la polidispersión. Esta caracterización computacional logra mejorar la generalizabilidad de los modelos QSPR obtenidos durante el proceso aprendizaje supervisado, en comparación con los inferidos mediante enfoques de representación univaluados. Sin embargo, esta nueva representación sigue sin contemplar las frecuencias de aparición de los distintos largos de cadena dentro de un material.

Por último, como contribución final de esta tesis se propone una representación computacional multivaluada, basada en el perfil polidisperso real de un material, donde cada descriptor queda caracterizado por una distribución probabilística discreta. En este contexto, las técnicas de selección de características empleadas para representaciones univaluadas ya no resultan aplicables, y surge la necesidad de contar con algoritmos que permitan operar sobre este nuevo modelo molecular. Como consecuencia de esto, se presenta el diseño e implementación de un algoritmo para selección de características multivaluadas. Este nuevo método, $FS4RV_{DD}$ (como sigla de su nombre en inglés *Feature Selection for Random Variables with Discrete Distribution*), logra un desempeño prometedor en todos los escenarios experimentales ensayados en estas investigaciones.

ABSTRACT

Nowadays, there is an increasing need to lead the in silico discovery of new industrial polymers through supervised Machine Learning approaches that identify structure-property correlations from the information contained in material databases, where each of them is characterized by Molecular Descriptors (MDs). These correlations are known as Quantitative Structure-Activity/Property Relationship models (QSAR/QSPR). They can be used to predict desirable properties of new materials before the synthesis stage, contributing to accelerate the design of new materials and to reduce the associated development costs.

QSAR/QSPR modeling is widely used in Molecular Informatics for Computer-Aided Drug Design. However, polymeric materials are significantly more complex than small molecules such as drugs, since they are collections of macromolecules that consist of a large number of structural repetitive units (SRUs) linked together in thousands of chain-like structures. These chains have different molecular weights (or lengths) and, in turn, they appear with different frequencies within each material. This phenomenon, known as polydispersity, is the main reason why many approaches developed for rational drug design are neither directly applicable nor sufficiently effective in the field of Polymer Informatics.

The main objective of this thesis is to contribute with solutions for various issues related to computational representation and algorithm development that arise during the QSPR modeling of properties of high molecular weight polydisperse polymers, with special emphasis on the Feature Selection problem. Because of frequency variations in the different chain lengths, the characterization of the polymeric material structure contains uncertainty, in contrast with the typical structural characterization of a small molecule. However, to deal with the uncertainty that introduces the polydispersity of polymeric materials, most of the QSAR/QSPR studies, until now, have used simple and univalued molecular models, that is, they calculate the molecular descriptors for a single instance of weight among all the possible chains that constitute a material. In particular, most QSPR studies use descriptors calculated on a single SRU, regardless of polydispersity. In this context, the present thesis proposes to investigate different alternatives of Feature Selection and Feature Learning for QSPR modeling with uncertainty that explore the effectiveness of more realistic computational representations for polymeric materials.

First, a hybrid methodology that uses MDs from both Feature Selection and Feature Learning algorithms is presented to evaluate the maximum predictive capability the traditional univalued representation (URE) can achieved. Then, new univalued representations based on average molecular weights are proposed, called Mn molecular model and Mw molecular model, whose capabilities to infer QSPR models are contrasted with the URE molecular model ones.

The other alternative computational representation proposes is trivalued MDs, based on the integration of URE, Mn, and Mw univalued molecular models into a single database. This representation partially captures the polydispersity inherent to polymers. This computational characterization improves the generalizability of QSPR models obtained during the supervised learning process, compared to those inferred through univalued representation approaches. However, this new trivalued representation still does not contemplate the frequencies of appearance of the different chain lengths within a material.

Finally, this thesis contributes with a multivalued computational representation based on the actual polydisperse profile of a material, in which each descriptor is characterized by a probabilistic discrete distribution. In this context, the Feature Selection techniques used for univalued representations are no longer applicable, and there is a need for algorithms to deal with this new multivalued molecular model. To face this need, both the design and implementation of an algorithm for the selection of multivalued features are presented here. This new method is called Feature Selection for Random Variables with Discrete Distribution (FS4RVDD), and it achieves a promising performance in all the experimental scenarios tested in these investigations.

LISTADO DE PUBLICACIONES

1. PUBLICACIONES CON CONTRIBUCIONES CENTRALES DE LA PRESENTE TESIS:

1.a) Artículos en Revistas Científicas con Referato Indexadas en Scopus:

1. **Cravero F.**, Schustik S., Martínez M.J., Vázquez G., Díaz M., Ponzoni I. " Feature Selection for Polymer Informatics: Evaluating Scalability and Robustness of the $FS4RV_{DD}$ algorithm using Synthetic Polydisperse Datasets". ***Journal of Chemical Information and Modeling***. 60 (2), 592-603 **(2020)**. American Chemical Society Publications. ISSN: 15499596. [doi: 10.1021/acs.jcim.9b00867].

2. **Cravero F.**, Martínez M.J., Ponzoni I, Díaz M.F. "Computational modelling of mechanical properties for new polymeric materials with high molecular weight". ***Chemometrics and Intelligent Laboratory Systems***, *193*, 103851 **(2019)**. Elsevier. ISSN: 0169-7439. [doi: 10.1016/j.chemolab.2019.103851].

3. **Cravero, F.**, Schustik, S., Martínez, M.J., Barranco, C.D., Díaz, M.F., Ponzoni, I. "Computer-Aided Design of Polymeric Materials: Characterization of Databases for Prediction of Mechanical Properties under Polydispersity". ***Chemometrics and Intelligent Laboratory Systems***, Vol. 191, pp. 65-72 **(2019)**. Elsevier. ISSN: 0169-7439. [doi: 10.1016/j.chemolab.2019.06.006].

4. **Cravero F.**, Martínez M.J., Vázquez G., Díaz M., Ponzoni I. "Feature Learning applied to the Estimation of Tensile Strength at Break in Polymeric Material Design", ***Journal of Integrative Bioinformatics***. Vol. 13, No. 2, 286 **(2016)**. De Gruyter. ISSN: 1613-4516. [doi: 10.2390/biecoll-jib-2016-286].

1.b) Publicaciones en Conferencias Indexadas en SCOPUS

1. **Cravero, F.**, Schustik, S., Martínez, M.J., Barranco, C.D., Díaz, M.F., Ponzoni, I. "Feature Selection and Polydispersity Characterization for QSPR Modelling: Predicting a Tensile Property". In: Fdez-Riverola F., Saberi Mohamad M., Rocha M., De Paz J. Gonzales P., (Eds.) 12th International Conference on Practical Applications of Computational Biology & Bioinformatics. ***Advances in Intelligent Systems and Computing***, Vol. 803, pp. 43-51 **(2019)**. Springer-Verlag. ISSN: 2194-5357. [doi: 10.1007/978-3-319-98702-6_6].

2. **Cravero, F.**, Schustik, S., Martínez, M.J., Díaz, M., Ponzoni, I. "$FS4RV_{DD}$: A feature selection algorithm for random variables with discrete distribution". In: Medina J., Ojeda-Aciego M., Verdegay J., Perfilieva I., Bouchon-Meunier B., Yager R. (Eds.) Information Processing and Management of Uncertainty in Knowledge-Based Systems. Applications. IPMU 2018. ***Communications in Computer and Information Science***, Vol. 855, pp. 211-222, **(2018)**. Springer-Verlag. ISSN: 1865-0929. [doi: 10.1007/978-3-319-91479-4_18].

3. **Cravero F.**, Martínez M.J., Vazquez G.E., Díaz M.F., Ponzoni I. "Intelligent Systems for Predictive Modelling in Cheminformatics: QSPR Models for Material Design using

Machine Learning and Visual Analytics Tools". In: Saberi Mohamad M., Rocha M., Fdez-Riverola F., Domínguez Mayo F., De Paz J. (eds) 10th International Conference on Practical Applications of Computational Biology & Bioinformatics. ***Advances in Intelligent Systems and Computing***, Vol. 477, pp. 3-11 **(2016)**. Springer-Verlag. ISSN: 2194-5357. [doi: 10.1007/978-3-319-40126-3_1].

1.c) Manuscritos en Enviados y/o en Proceso de Redacción:

1. Schustik, S., **Cravero, F.**, Martínez, M. J., Ponzoni, I., Díaz, M. F. PolyMaS: A new tool for computational polymerization of structural repetitive units. ***Artículo enviado a una revista del área de Quimioinformática. Actualmente en revisión.*** (2019)

2. PUBLICACIONES CON OTROS GRUPOS CON RESULTADOS RELACIONADOS CON LOS CAPÍTULOS INTRODUCTORIOS DE LA PRESENTE TESIS (CAPÍTULOS 3 Y 4):

2.a) Artículos en Revistas Científicas con Referato Indexadas en Scopus:

1. Ponzoni I., Sebastian-Pérez V., Martínez M.J., Roca C., De la Cruz, C., **Cravero F.**, Vazquez, G.E., Páez J.A., Díaz M.F., Campillo N.E**.** "QSAR Classification Models for Predicting the Activity of Inhibitors of Beta-Secretase (BACE1) Associated with Alzheimer's Disease". ***Scientific Reports***, Vol. 9:9102, **(2019)**. Nature Pub. Group. ISSN: 2045-2322. [doi: 10.1038/s41598-019-45522-3].

2. Ponzoni I., Sebastian-Pérez, Requena C., Roca C., Martínez M.J., **Cravero F.**, Díaz M.F., Páez J.A., Gomez Arrayas R., Adrio J., Campillo N.E. "Hybridizing Feature Selection and Feature Learning Approaches in QSAR Modeling for Drug Discovery", ***Scientific Reports***. Vol. 7:2403, **(2017)**. Nature Pub. Group. ISSN: 2045-2322. [doi: s41598-017-02114-3].

2.b) Publicaciones en Conferencias Internacionales Indexadas en SCOPUS

1. **Cravero F.**, Martínez M.J., Díaz M.F., Ponzoni I. "QSAR Classification Models for Predicting Affinity to Blood or Liver of Volatile Organic Compounds in e-Health". In: Rojas I., Ortuño F. (Eds.) Bioinformatics and Biomedical Engineering. IWBBIO 2017. ***Lecture Notes in Computer Science***, Vol. 10209, pp. 424-433, **(2017)**. Springer-Verlag. ISSN: 0302-9743. [doi: 10.1007/978-3-319-56154-7_38].

2. Palomba, D., Martínez, M.J., **Cravero, F.**, Soto, A.J., Vazquez, G.E., Ponzoni, I., Díaz, M.F. "Prediction of mechanical properties of tensile test for linear polymers. QSPR modeling with computational intelligence and interactive visual analysis" | "Predicción de propiedades mecánicas del ensayo de tensión para polímeros lineales. Modelado QSPR con inteligencia computacional y análisis visual interactivo", ***Journal of the Argentine Chemical Society***, Vol. 101(1-2), pp. 137-147. Asociación Argentina de Química, **(2014)**. ISSN: 0365-0375. [link: http://aqa.org.ar/images/anales/pdf101/][1].

[1] Número especial dedicado al Congreso Argentino de Química 2014.

3. OTRAS PUBLICACIONES DE ARTÍCULOS COMPLETOS Y RESÚMENES EN CONFERENCIAS RELACIONADAS CON ESTA TESIS (NO INDEXADAS EN SCOPUS):

3.a) En Conferencias Internacionales:

1. **Cravero F.**, Schustik S., Martínez M.J., Ponzoni I., Díaz M.F. Macro Approach to Molecular Modelling of Linear Polymers Applied to Estimation of Tensile Modulus for New Materials Development (1 pag.). *VIII International Symposium on Materials* (Materias 2017). Aveiro, Portugal. (Abril, **2017**). Resumen.

2. Martínez, M.J., **Cravero F.**, Díaz M.F., Ponzoni I. QSPR Modeling Applied to High Molecular Weight Polymers: Ductility Characterization from Elongation at Break (1 pag.) *VIII International Symposium on Materials* (Materias 2017). Aveiros, Portugal. (Abril, **2017**). Resumen.

3. **Cravero F**., Vázquez G.E., Ponzoni I., Díaz Mónica F. Modelado molecular de materiales poliméricos en quimioinformatica. *Simposio Argentino de Materiales* (SAM-CONAMET 2016). Córdoba, Argentina. (Noviembre, **2016**). Resumen.

4. Palomba, D.; **Cravero, F.**; Vázquez, G.E.; Díaz, M.F. Prediction of tensile modulus for linear polymers applied to new materials development. *Congreso Internacional de Metalurgia y Materiales* (SAM-CONAMET) / Iberomat / Simposio Materia 2014. Santa Fe, Argentina. (Octubre, **2014**). Artículo breve.

5. Palomba, D.; **Cravero, F.**; Vázquez, G.E.; Díaz, M.F. Prediction of tensile strength at break for linear polymers applied to new materials development. *Congreso Internacional de Metalurgia y Materiales* (SAM-CONAMET) / Iberomat / Simposio Materia 2014. Santa Fe, Argentina. (Octubre, **2014**). Artículo breve.

3.b) En Conferencias Nacionales

1. Schustik, S.A., **Cravero F.**, Martinez, M.J., Ponzoni, I., Diaz, M.F. Informática de polímeros aplicada a la estimación de la ductilidad en un nuevo material a partir de la relación de propiedades derivadas del ensayo de tensión. 32 *Congreso Argentino de Química*, Buenos Aires, Argentina (Marzo, **2019**). Artículo breve.

2. **Cravero F.**, Schustik S., Martínez M.J., Ponzoni I., Díaz Mónica F. Herramienta computacional para testeo virtual de propiedades mecánicas de polímeros. Etapa de ensamble y prueba. *Simposio Argentino de Polímeros* (SAP 2017). Los Cocos, Córdoba, (Octubre, **2017**). Resumen.

3. **Cravero F.**, Martínez M.J., Vázquez G.E., Ponzoni I., Díaz M.F. Predicción de curvas teóricas de distribución de peso molecular de resinas 31° *Congreso Argentino de Química*, Buenos Aires, Argentina (Octubre, **2016**). Artículo breve.

4. **Cravero F.**, Martínez M.J., Vázquez G.E., Ponzoni I., Díaz M.F. Representación de la estructura molecular de polímeros sintéticos de alto peso. 31° *Congreso Argentino de Química*, Buenos Aires, Argentina (Octubre, **2016**). Artículo breve.

5. **Cravero F.**, Ponzoni I., Díaz M.F. Informática Molecular: Polímeros modelados por Distribución de Pesos. III *Congreso Internacional de Ciencia y Tecnología de la Provincia de Buenos Aires* (CICyT CIC **2016**), La Plata – Argentina (Septiembre, **2016**). Resumen.

6. **Cravero F.**, Martínez M.J., Ponzoni I., Vazquez G.E., Díaz M.F. Desarrollo de modelos QSPR asistido por técnicas de analítica visual para la predicción de propiedades mecánicas de polímeros lineales. *Simposio Argentino de Polímeros* (SAP 2015), Santa Fe, (Octubre **2015**). Artículo breve.

7. **Cravero F.**, Martínez M.J., Ponzoni I., Vazquez G.E., Díaz M.F. Predicción del Módulo Elástico Para Polímeros Lineales aplicando Analítica Visual y Aprendizaje Automático. *Simposio Argentino de Polímeros* (SAP **2015**), Santa Fe, (Octubre, **2015**). Artículo breve.

8. **Cravero F.**, Vazquez G.E., Díaz M.F., Ponzoni I. Modelado QSPR de propiedades mecánicas de materiales poliméricos empleando técnicas de reducción de variables basadas en algoritmos de aprendizaje automático. VIII *Congreso Argentino de Ingeniería Química* (CAIQ **2015**), Ciudad Autónoma de Buenos Aires, Argentina, (Agosto, **2015**). Artículo completo.

9. **Cravero F.**, Martínez M.J., Díaz M.F., Vazquez G.E., Ponzoni I. An integral framework for QSAR Modelling using Computational Intelligence and Visual Analytics. VI *Congreso Argentino de Bioinformática y Biología Computacional* (6CAB2C). Bahía Blanca, Argentina (Octubre, **2015**). Resumen.

10. **Cravero F.**, Díaz, M.F.; Ponzoni, I. Modelado Predictivo: Nuevos Descriptores Moleculares para Polímeros. II *Congreso Internacional Científico y Tecnológico de la Provincia de Buenos Aires* (2 ConCyT **2015**). La Plata, Argentina, (Octubre, **2015**). Resumen.

11. Martínez M.J.; **Cravero F.**; Palomba D.; Soto A.J.; Díaz M.F.; Vázquez, GE. Ponzoni, I. Feature Selection in Molecular Informatics: Improving QSAR/QSPR Modeling by Computational Intelligence Approaches and Interactive Visual Analysis. *Simposio Argentino de Inteligencia Artificial* 2014 (43JAIIO). Ciudad Autónoma de Buenos Aires, Argentina (Septiembre, **2014**). Resumen.

12. Martínez, M. J; **Cravero, F.**; Vázquez, G.E.; Díaz, M.F.; Soto, A.J.; Ponzoni, I. Interactive Visual Analysis Methodology for Improving Descriptor Selection in QSPR: First Steps. V Congreso Argentino de Bioinformática y Biología Computacional (5CAB2C). San Carlos de Bariloche, Argentina (Septiembre, **2014**). Resumen.

ÍNDICE

CAPÍTULO 1: INTRODUCCIÓN

CAPÍTULO 2: APRENDIZAJE MAQUINAL

CAPÍTULO 3: INFORMÁTICA MOLECULAR

CAPÍTULO 4: INFORMÁTICA DE POLÍMEROS

CAPÍTULO 5: MODELADO QSPR CON DMS UNIVALUADOS

CAPÍTULO 6: MODELADO QSPR CON DMS TRIVALUADOS

CAPÍTULO 7: SELECCIÓN DE DMS MULTIVALUADOS

CAPÍTULO 8: CONCLUSIONES

REFERENCIAS Y ANEXOS

INTRODUCCIÓN

CAPÍTULO 1

Este capítulo resume lo que se tratará a lo largo de esta tesis para que el lector encuentre más fácil la comprensión de la misma. Se comienza con los aspectos generales para luego plantear los objetivos. Se continúa con un marco teórico sintético, el estado del arte, la proyección y direccionamiento del trabajo de esta tesis. Finalmente, se describe la organización y en forma breve el contenido de cada uno de sus 8 capítulos.

1.1. ASPECTOS GENERALES

Esta tesis doctoral en Ciencias de la Computación desarrollada bajo la dirección conjunta de la Dra. Mónica F. Díaz, especialista en química y materiales poliméricos, y el Dr. Ignacio Ponzoni, especialista en aprendizaje maquinal y bioinformática, está enmarcada en una línea de trabajo que los doctores vienen desarrollando hace más de 10 años y tiene como objetivo final diseñar algoritmos de aprendizaje maquinal especialmente ideados para predecir propiedades ADMET de fármacos y propiedades mecánicas de materiales poliméricos de alto peso molecular. Este último punto, es central para esta tesis ya que actualmente es el que presenta mayor desafío y originalidad al involucrar variables con polidispersión (familias de pesos moleculares), problema que no ha sido abordado en la literatura hasta ahora.

La Informática Molecular constituye una disciplina emergente dedicada al desarrollo de técnicas computacionales que facilitan el análisis y la comprensión de los principios que rigen a las moléculas y a sus interacciones [Baumann *et al.,* 2011]. Este nuevo campo de investigación, derivado de la Bioinformática y la Quimioinformática, posee múltiples áreas de aplicación que van desde el descubrimiento de drogas y fármacos [Gola *et al.*, 2006], hasta el diseño de nuevos materiales poliméricos [Afantitis *et al.*, 2005].

En Quimioinformática, uno de los enfoques más desarrollados es el modelado de la Relación Cuantitativa Estructura Actividad/Propiedad (QSAR/QSPR por sus siglas en inglés para *Quantitative Structure - Activity/Property Relationship*). En el diseño y descubrimiento racional de fármacos y pequeñas moléculas, la predicción

de actividades y propiedades por metodología QSAR/QSPR lidera el desarrollo de algoritmos [Nosengo, 2016]. La tarea de diseñar modelos QSAR/QSPR enfrenta algunos desafíos intrínsecos como determinar el conjunto de descriptores moleculares óptimos a utilizar en el modelo. Esto se conoce como Selección de Características (*Feature Selection*). De este campo, se han tomado herramientas computacionales y se las ha adaptado para el diseño de nuevos materiales poliméricos, emergiendo así la Informática de Polímeros [Adams, 2010]. Sin embargo, esta última requiere una adecuación en cada uno de los pasos de la metodología QSAR/QSPR, debida al gran tamaño de las moléculas y también a su polidispersión asociada. Razón por la cual, los materiales poliméricos constituyen un caso particular de sistemas complejos.

El foco principal de esta tesis es la investigación de técnicas de aprendizaje maquinal aplicadas al modelado QSAR/QSPR, más específicamente al proceso de análisis y selección de descriptores moleculares. A continuación, se detallan el objetivo general de esta tesis y sus objetivos específicos asociados. Luego, se describe brevemente el marco conceptual en el que se desarrolla este trabajo de tesis, continuando, con la organización de la misma mediante una resumida descripción de lo desarrollado en cada capítulo.

1.1.1. OBJETIVO GENERAL

El objetivo general de esta tesis es contribuir con soluciones para distintos problemas que surgen en el modelado QSPR de propiedades de polímeros polidispersos de alto peso molecular, las cuales resultan relevantes para el diseño computacional de materiales poliméricos. En particular, se aborda el diseño de nuevas representaciones computacionales para materiales poliméricos y el desarrollo de algoritmos de Aprendizaje Maquinal (*Machine Learning*), especialmente ideados para atacar el problema de Selección de Descriptores (*Feature Selection*), que capturen la polidispersión de estas macromoléculas.

1.1.2. OBJETIVOS ESPECÍFICOS

Los objetivos específicos de esta tesis se listan más abajo. Cumpliéndolos se espera que la presente tesis logre contribuciones relevantes en las Ciencias de la Computación, tanto en el área del Aprendizaje Maquinal como en el ámbito de la Informática de Polímeros.

- Explorar, diseñar y testear algoritmos para la transformación y reducción del espacio de descriptores de un conjunto de moléculas, usando enfoques de aprendizaje maquinal.

- Representar computacionalmente la polidispersión de polímeros de alto peso molecular con curvas teóricas a partir de pesos promedios (Mn y Mw) que definen al material.
- Evaluar el desempeño de los modelos QSPR, inferidos para la menor instancia de peso (URE), en instancias mayores de pesos (Mn y Mw); y viceversa.
- Desarrollar modelos QSPR para la inferencia de propiedades mecánicas de polímeros de alto peso, usando enfoques de aprendizaje maquinal que permitan modelar la entrada mediante descriptores moleculares multivaluados.
- Diseñar un algoritmo de selección de características para la inferencia de propiedades mecánicas de polímeros de alto peso, usando enfoques de aprendizaje maquinal que permitan como entrada variables caracterizadas como distribuciones discretas de probabilidad.
- Integrar todas las herramientas y conocimientos desarrollados en estas investigaciones en una arquitectura de software que asista a científicos en el descubrimiento, análisis y estudio de propiedades moleculares en contextos donde existe incertidumbre en los valores de los descriptores moleculares.
- Testear y validar estas estrategias en la predicción de propiedades moleculares de relevancia para el diseño de nuevos materiales poliméricos.

1.2. MARCO TEÓRICO

Las primeras civilizaciones disponían de una cantidad reducida de materiales, ya que estos solo eran naturales. Hace apenas un poco más de 200 años comenzó a entenderse la relación entre los elementos estructurales de los materiales y sus propiedades macromoleculares. Este conocimiento permitió modificar o adaptar las características de los materiales para proporcionar mayor confort a las generaciones modernas, a través del desarrollo de nuevos materiales y de tecnologías avanzadas. El diseño y la síntesis de nuevos materiales con propiedades específicas y novedosas, han resultado en uno de los campos más dinámicos de la ciencia moderna. Sin embargo, el tiempo promedio para alcanzar un nuevo material se estima entre 10 y 20 años desde la investigación inicial hasta su implementación y primer uso [Kutz, 2002; Holdren, 2011].

En los últimos 50 años, las aproximaciones computacionales se han vuelto esenciales en este campo, sin embargo, los tiempos de diseño aún no han experimentado una disminución notable, especialmente para los materiales poliméricos. El enfoque típico en el diseño de nuevos materiales ha sido empírico, con una modalidad "lineal" (Figura 1.1) que incluye varias etapas consecutivas que generalmente llevan a cabo equipos científicos y/o de ingeniería de diferentes

instituciones con poca retroalimentación entre ellos, lo que podría ralentizar el proceso total [Kutz, 2002; Holdren, 2011].

FIGURA 1. 1 RESUMEN GRÁFICO DE LAS ETAPAS DE DESARROLLO DE NUEVOS MATERIALES: MÉTODO TRADICIONAL (LINEAL) VS MÉTODO CON INCLUSIÓN DE ETAPA DE DISEÑO Y TESTEO *IN SILICO* (NO LINEAL).

Actualmente se avanzó mucho en el conocimiento de las relaciones entre la estructura molecular de un material y sus propiedades, lo que conduce a la capacidad de predecir computacionalmente las propiedades del material, previo a su síntesis. Además, en 2011 surgió oficialmente la Iniciativa Genoma de Materiales (*Materials Genome Initiative*) que busca potenciar la combinación sinérgica del experimento, la teoría y la informática para acelerar el ritmo del descubrimiento y diseño de materiales [Adams & Murray-Rust, 2008]. El desarrollo de nuevas entidades poliméricas se convierte en una prioridad estratégica, ya que estas tienen cada vez mayor implicancia en las interfaces de varias disciplinas científicas: polímeros y medicina, polímeros y alimentos, etc. [Adams, 2010].

La Informática Molecular constituye una disciplina dedicada al desarrollo de técnicas computacionales que facilitan el análisis y la comprensión de los principios que rigen a las moléculas y a sus interacciones, entre otras funciones. En particular, la técnica de QSAR/QSPR (*Quantitative Structure-Activity/Property Relationship*) relaciona de manera cuantitativa parámetros específicos de la estructura de la molécula (descriptores moleculares) con la actividad o propiedad objetivo (target) que se desea estudiar. Las herramientas predictivas comenzaron a utilizarse desde mediados del siglo pasado [Hansch *et al.*, 1962] y han evolucionado en la complejidad de sus algoritmos y metodologías. La técnica QSAR/QSPR es uno de

los enfoques más empleados para modelar las propiedades físicas y biológicas de productos químicos.

Los modelos QSAR se aplican para evaluar impactos potenciales de químicos y materiales en sistemas ecológicos y en la salud humana. Una de sus utilidades más conocidas, y que mayor desarrollo de algoritmia presenta es la predicción de las cualidades o propiedades ADMET (*Absorption, Distribution, Metabolism, Excretion and Toxicity*) de fármacos y, aunque menos conocido para COVs (*Volatic Organic Compounds*) [Cherkasov *et al*., 2014]. Usar métodos QSAR/QSPR es uno de los pasos claves en el objetivo de reducir considerablemente el costo del proceso de descubrimiento y diseño racional de drogas. Esto se debe a que ayuda a detectar, en etapas tempranas del desarrollo, la mayoría de aquellos compuestos que no serán viables en etapas futuras.

En lo que respecta al diseño de nuevos materiales, la predicción de las propiedades de un polímero de diseño, con métodos *in silico*, se ha convertido en un paso relevante durante las primeras etapas del desarrollo de nuevos materiales con propiedades específicas. Los polímeros se definen como macromoléculas compuestas por cientos de miles, y más, de Unidades Repetitivas Estructurales (UREs) que se repiten a lo largo de las cadenas que lo componen. Por ende, son macromoléculas significativamente más complejas y difusas que las habitualmente estudiadas (pequeñas moléculas como fármacos).

Por lo tanto, las soluciones informáticas que se han desarrollado, y continúan desarrollándose para el diseño de drogas, no son lo suficientemente efectivas para la Informática de Polímeros, la cual requiere de sus propios enfoques. Varios autores han realizados modelos QSPR teóricos que utilizan descriptores moleculares basados en UREs para predecir la Temperatura de Transición vítrea (Tg). Esta propiedad térmica, una de las más estudiadas en la literatura, está relacionada con el rendimiento mecánico y la procesabilidad del material. Probablemente como se esperaba, estos estudios muestran que el monómero tiene cierta información sobre la temperatura de transición vítrea y la estructura del polímero correspondiente. Sin embargo, el principal problema de utilizar la URE o el monómero para realizar una predicción es que se ignora tanto la historia del polímero como el peso molecular, los cuales tienen una fuerte dependencia con la Tg [Adams, 2010].

Aprender de los datos ha generado cambios de paradigmas en una multitud de disciplinas. El primer paso en el aprendizaje maquinal molecular es codificar la estructura de la molécula en una forma que sea apta para este, y es aquí donde se enfoca gran cantidad de investigación actualmente. Una representación útil codifica

características que son relevantes y eficientes, para evitar la maldición de la dimensionalidad. En el aprendizaje maquinal, el método que se utiliza parecería no ser el factor limitante, siempre que se haga su debida diligencia, se exploren varias clases de métodos y se optimicen sus hiperparámetros. El componente que determina el éxito es la caracterización utilizada, en el caso de la Informática de Polímeros, es la caracterización o representación estructural del material. El genoma de polímeros persigue el objetivo de lograr una fácil representación que pueda correlacionarse con las propiedades y la posibilidad de estimarlas como una función del genoma [Mannodi-Kanakkithodi *et al.*, 2018].

Es importante considerar que la mayoría de los polímeros comerciales no presentan un único peso molecular, sino una distribución de valores denominado polidispersión de pesos moleculares. Como consecuencia la caracterización de estos materiales requiere computar distintas instancias de peso molecular derivadas de la URE y no existen antecedentes en la literatura del estudio de este problema. Para los polímeros la representación de la estructura por tabla de conexión no es satisfactoria, porque estos son colecciones de macromoléculas polidispersas [Adams, 2010]. Por lo tanto, la representación de los valores de los descriptores moleculares calculados para estas representaciones ya no puede ser descripta con un único valor real, sino que esta debe ser multivaluada y puede ser representada como distribuciones probabilísticas discretas.

Se pueden calcular miles de descriptores moleculares a partir de la estructura, pero solo algunos serán los adecuados para establecer la relación con la propiedad. Para elegir los que mejor se ajusten, se aplican métodos de Selección de Características o Aprendizaje de Características (también conocido como Extracción de Características). La Selección de Características es un campo del Aprendizaje Maquinal y del Reconocimiento de Patrones que consiste en reducir la dimensionalidad de los datos al eliminar aquellas características que son ruidosas, redundantes o irrelevantes para el problema en estudio. Para esta propuesta de caracterización multivaluada de los descriptores moleculares de los materiales para el modelado QSPR, se requiere realizar una Selección de Características teniendo en cuenta dichas distribuciones. El desafío reside entonces, en el desarrollo de un método Selección de Características capaz de trabajar con representaciones basadas en la polidispersión.

1.3. ORGANIZACIÓN DE LA TESIS

En la presente tesis, en cuanto a Informática Molecular, se desarrollaron algoritmos y metodologías para el modelado QSAR basados en Aprendizaje

Maquinal para el estudio de pequeñas moléculas como drogas y compuestos orgánicos volátiles (COVs). Con respecto a Informática de Polímeros, primero se desarrollaron estrategias computacionales para inferir modelos QSPR para la predicción de propiedades mecánicas, donde los materiales son caracterizados a través de su Unidad Repetitiva Estructural (URE). Luego, en lo que se consideran los aportes más originales, se infirieron modelos QSPR basados en descriptores moleculares (DMs) trivaluados y, posterioremente se abordó el modelado computacional de polímeros teniendo en cuenta su rasgo más descriptivo e importante: la polidispersión, haciendo foco en la Selección de Características. Esta tesis se encuentra organizada en 8 capítulos. A continuación, se describe brevemente la estructura de cada uno de ellos.

Capítulo 1: Introducción

El Capítulo 1 es la introducción de esta tesis y presenta los aspectos generales estudiados en la misma y los objetivos perseguidos. Dado que los conceptos específicos necesarios son revisados en cada capítulo a medida que estos son demandados, en este capítulo sólo se presenta un resumido marco conceptual genérico del área de investigación específica en que se enmarca la tesis. Finalmente, se describen el alcance de cada capítulo según la organización que presenta esta tesis. Un esquema gráfico del contenido de este capítulo se presenta en la Figura 1.2.

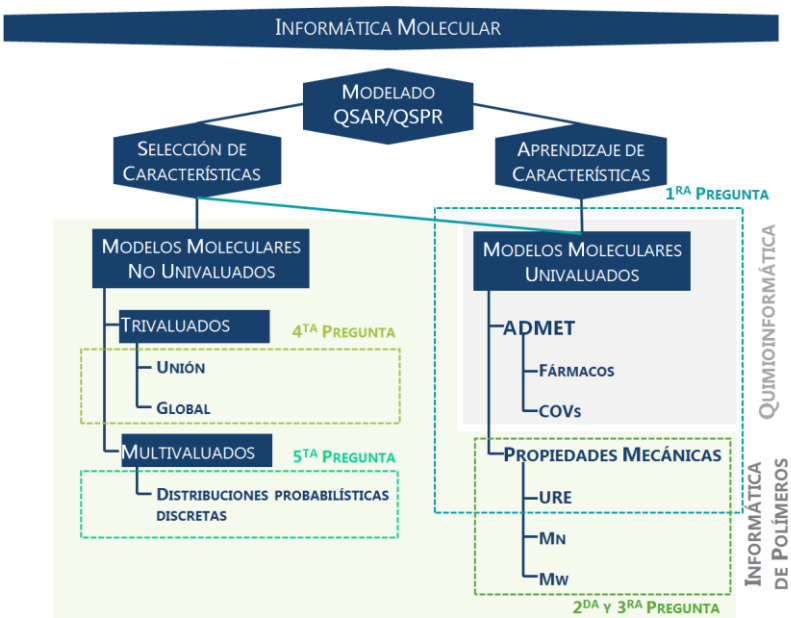


FIGURA 1. 2. ESQUEMA DE LA ORGANIZACIÓN DE LA TESIS PRESENTE EN EL CAPÍTULO 1.

Capítulo 2: Aprendizaje Maquinal

En el Capítulo 2 se presentan brevemente algunos conceptos centrales relativos al Aprendizaje Maquinal (*Machine Learning*) que son empleados en esta tesis. Es importante aclarar que no se busca hacer una revisión exhaustiva de esta amplia área del conocimiento, sino simplemente establecer los límites del marco teórico-conceptual dentro del cual se realizaron las investigaciones de esta tesis. Un esquema gráfico del contenido de este capítulo se presenta en le Figura 1.3. Aquel lector con experticia en el área de Aprendizaje Maquinal puede simplemente avanzar hacia los siguientes capítulos, si así lo deseara, dado que se trata de conceptos básicos y bien establecidos en la disciplina.

FIGURA 1. 3. ESQUEMA GRÁFICO DE LOS CONCEPTOS DESARROLLADOS EN EL CAPÍTULO 2.

Capítulo 3: Informática Molecular

El Capítulo 3 está dedicado a la Informática Molecular. En este capítulo se hace un recorrido por los principales hitos en la representación de moléculas. Se revisan conceptos básicos en el área, como los relativos al modelado QSAR/QSPR, para luego enfocarse en la descripción de las técnicas utilizadas en esta tesis, empleadas sobre pequeñas moléculas como fármacos y COVs. En la Figura 1.4 se presenta un esquema gráfico del proceso de modelado QSAR/QSPR, tema central de este capítulo. Estas primeras experiencias sentaron las bases para el desarrollo de las contribuciones orientadas a la Informática de Polímeros.

Figura 1. 4. Esquema Gráfico del Proceso de modelado QSAR/QSPR

Capítulo 4: Informática de Polímeros

En el Capítulo 4 se revisan los conceptos básicos de Informática de Polímeros y Química de Polímeros, necesarios para la comprensión de esta tesis. Un esquema gráfico del objetivo que persigue la Informática de polímeros, instanciado para los contenidos de este capítulo, se muestra en la Figura 1.5. Se presenta la base de datos de polímeros usada en este y los siguientes capítulos para la predicción de propiedades mecánicas. El estado de arte del área es resumido aquí. Además, se presentan las experimentaciones, resultados y conclusiones parciales que se obtuvieron al utilizar el modelo molecular URE, según lo realizado hasta el momento en la literatura para el modelado QSPR de polímeros. Las contribuciones de este capítulo abordan la primera pregunta de investigación: *¿Puede el aprendizaje de características empleadas en enfoques QSAR resultar de utilidad en el contexto de Informática de Polímeros? ¿Qué sucede si combinamos las ventajas del aprendizaje de características con las de la selección de características para la inferencia de modelos QSPR en polímeros de alto peso molecular?*

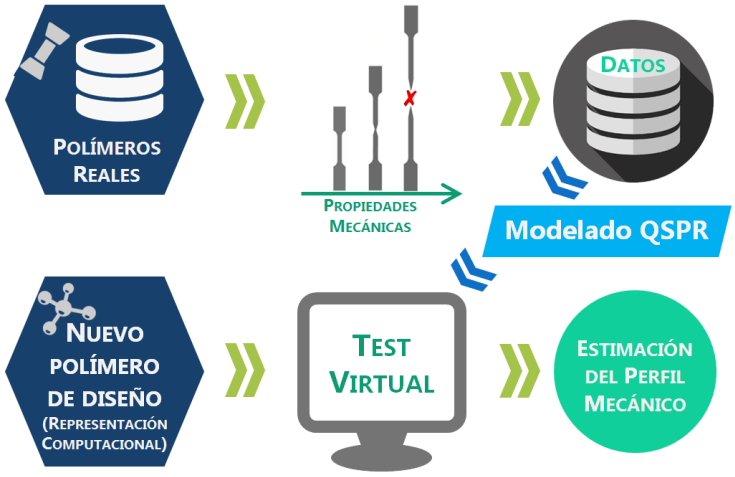


Figura 1. 5. Resumen gráfico del proceso de predicción de propiedades mecánicas para polímeros a través del modelado QSPR como cribado o test virtual.

Capítulo 5: Modelado QSPR con DMs univaluados

El capítulo 5 presenta la primera propuesta alternativa al uso de URE como la representación computacional de polímeros. Sin embargo, sigue empleando una única instancia de peso, es este caso Mn y Mn (peso promedio en número y en peso). En la Figura 1.6 se presenta un resumen del contenido de este capítulo en forma de esquema resaltando la idea de comprobación bidireccional de la representación estructural propuesta para polímeros. En este capítulo se responden la segunda y la tercera preguntas de investigación: ¿Son los modelos QSPR basados en el modelo molecular URE efectivos cuando se testean sobre modelos moleculares de alto peso? Los descriptores moleculares que fueron seleccionados en los modelos QSPR basados en el modelo molecular URE ¿pueden resultar de utilidad para inferir nuevos modelos QSPR a partir de base de datos de otras instancias univaluadas de representación de mayor peso que URE (Mn y Mw)?, y ¿Son los modelos QSPR basados en el modelo molecular Mn efectivos cuando se testean sobre modelos moleculares de otro peso (URE y Mw)? Los descriptores moleculares que fueron seleccionados en los modelos QSPR basados en el modelo molecular Mn ¿pueden resultar de utilidad para inferir nuevos modelos QSPR a partir de bases de datos de otras instancias univaluadas de representación (URE y Mw)? ¿Son los modelos QSPR basados en el modelo molecular Mw efectivos cuando se testean sobre modelos moleculares de otro peso (URE y Mn)? Los descriptores moleculares que fueron seleccionados en los modelos QSPR basados en el modelo molecular Mw ¿pueden resultar de utilidad para inferir nuevos modelos QSPR a partir de bases de datos de otras instancias univaluadas de representación (URE y Mn)?

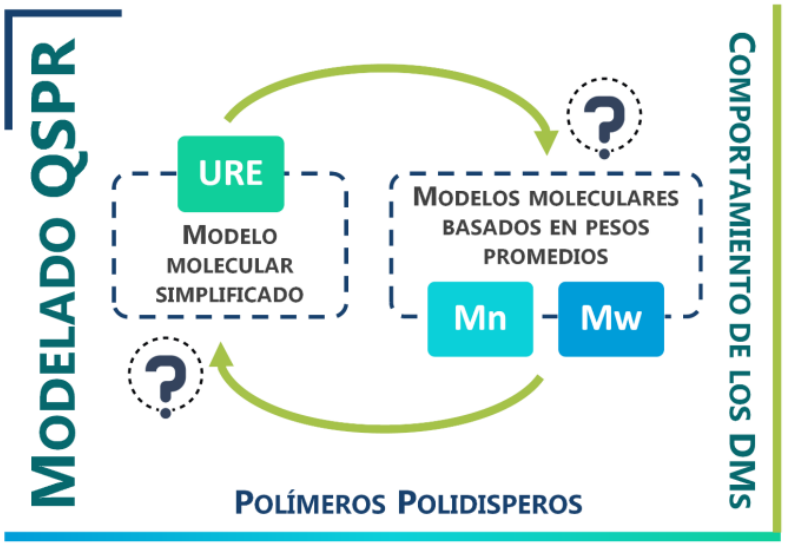


FIGURA 1. 6. ESQUEMA GRÁFICO DE LA COMPROBACIÓN BIDIRECCIONAL SOBRE LA EFECTIVIDAD DE LAS REPRESENTACIONES ESTRUCTURALES PROPUESTAS PARA POLÍMEROS POLIDISPERSOS DE ALTO PESO MOLECULAR.

Capítulo 6: Modelado QSPR con DMs trivaluados

En el Capítulo 6 se continúa con el modelado QSPR de materiales poliméricos, esta vez utilizando descriptores moleculares trivaluados (en URE, Mn y Mw) y esto constituye la segunda propuesta alternativa de representación computacional de materiales poliméricos que tiene en cuenta los pesos promedios que describen la polidispersión. De acuerdo a los resultados obtenidos en el capítulo anterior, una única instancia de peso molecular para representar a un polímero no resulta suficiente, se necesita caracterizarlo con información estructural descriptiva correspondiente a distintos largos de cadena polimérica. Se generaron las bases de datos, los experimentos y se presentan los resultados orientados a responder la cuarta pregunta de investigación (Figura 1.7): ¿Existen modelos moleculares basados en sus pesos moleculares promedios, de los materiales, que den como resultado modelos QSPR predictivos más precisos que los obtenidos por modelos moleculares URE? ¿Es aconsejable integrar en una única base de datos los descriptores moleculares correspondientes a modelos moleculares de los diferentes pesos característicos relacionados con las curvas de distribución de peso molecular de los materiales?

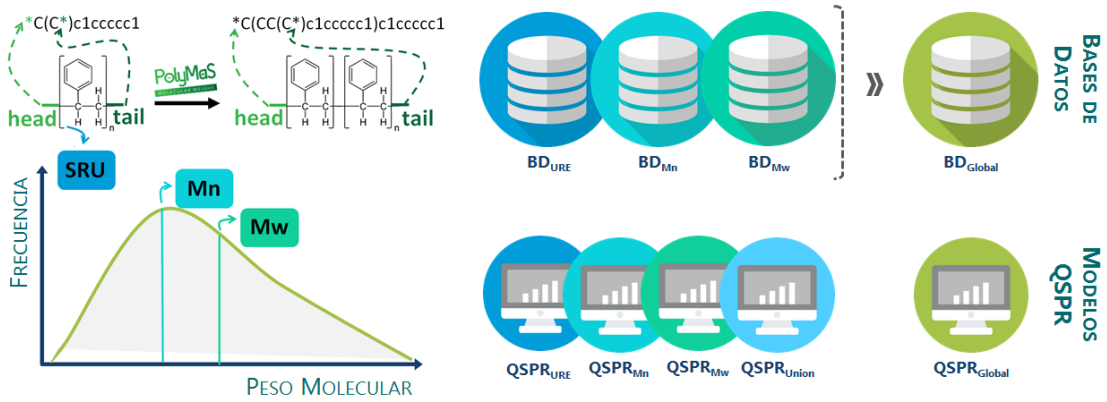


FIGURA 1. 7. ESQUEMA DE LA CONSTRUCCIÓN DE LA BASE DE DATOS TRIVALUADA Y LOS DIFERENTES MODELOS QSPR INFERIDOS PARA RESPONDER A LA CUARTA PREGUNTA DE INVESTIGACIÓN.

Capítulo 7: Selección de Características multivaluadas

En el Capítulo 7 se presenta la tercera propuesta alternativa de representación computacional para el modelado QSPR de materiales poliméricos. Aquí, los descriptores moleculares están representados mediante distribuciones discretas probabilísticas derivadas de la curva de polidispersión de pesos moleculares (descriptores multivaluados). A partir de esta representación se plantea la quinta, y última, pregunta de investigación: *¿Es posible identificar con más precisión los DMs más relevantes usando un algoritmo de Selección de Características multivaluadas que usando enfoques tradicionales sobre representaciones univaluadas*? Para responderla, se propuso un nuevo algoritmo de Selección de Características ideado para operar sobre DMs representados mediante distribuciones probabilísticas y se diseñaron bases de datos sintéticas que permitieran evaluar el rendimiento del mismo y responder a la pregunta de investigación (Figura 7.8).

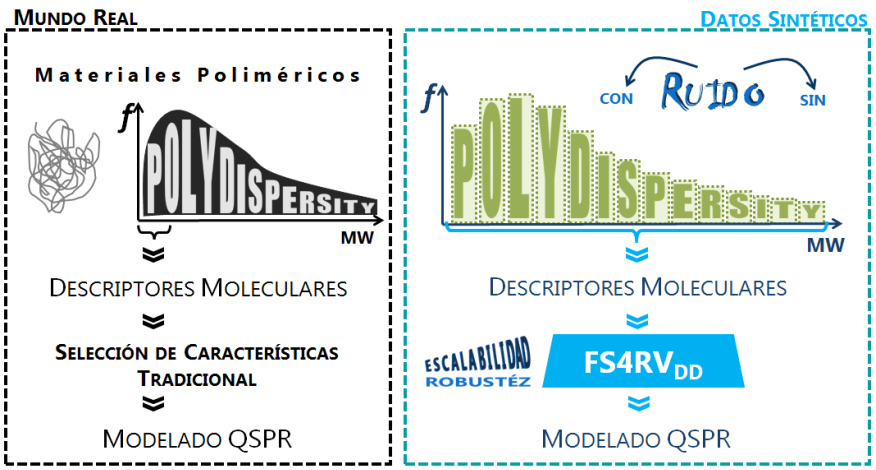


FIGURA 1. 8. ESQUEMA GRÁFICO DE LA DIFERENCIA ENTRE EL TRATAMIENTO DE LA REPRESENTACIÓN COMPUTACIONAL UNIVALUADA Y MULTIVALUADA.

Capítulo 8: Conclusiones y Trabajos Futuros

El capítulo 8 es el último de la presente tesis, y en él se describen las conclusiones que se obtuvieron de las tareas de investigación realizadas. Se presentan las dificultades sorteadas, aquellas aún por resolver y las contribuciones centrales realizadas. Además, se delinean algunas ideas que se desprenden de este trabajo de tesis y, que también, pueden constituir nuevas líneas de investigación.

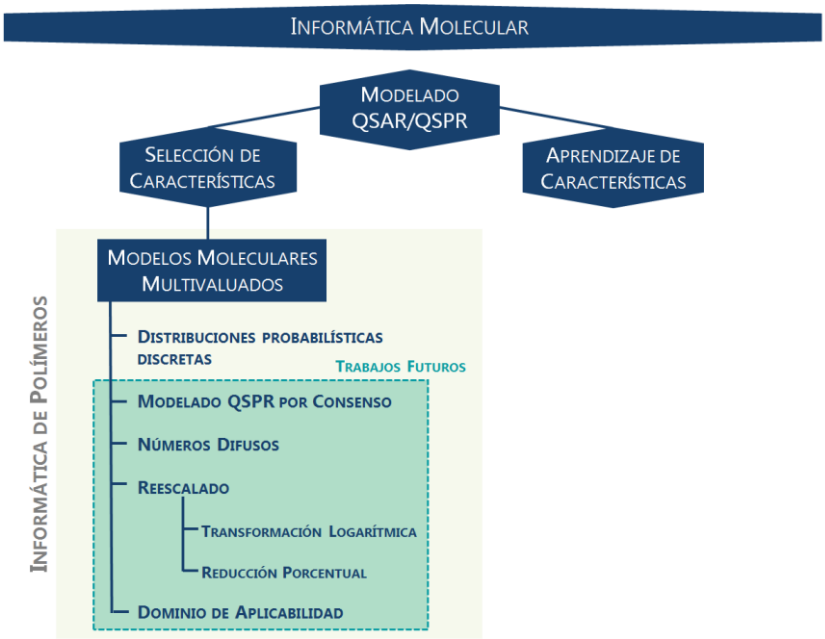


FIGURA 1. 9. RESUMEN GRÁFICO DE LOS TRABAJOS FUTUROS MÁS IMPORTANTES QUE PUEDEN DESPRENDERSE DE ESTA TESIS.

APRENDIZAJE MAQUINAL

CAPÍTULO 2

En el presente capítulo se describirán brevemente algunos conceptos centrales relativos al Aprendizaje Maquinal empleados en esta tesis. Es importante aclarar que no se pretende realizar una revisión exhaustiva de esta amplia área del conocimiento, sino simplemente establecer los límites del marco teórico dentro del cual se realizaron las investigaciones de la tesis. Por otra parte, consideramos que el lector con experticia en el área puede simplemente avanzar hacía los siguientes capítulos, si así lo deseara, dado que se trata de conceptos básicos y bien establecidos en la disciplina.

2.1. CONCEPTOS DE APRENDIZAJE MAQUINAL

Aprender es adquirir conocimientos diversos, de cualquier tipo. Está relacionado con la acción de perseguir y atrapar algo. Esta acción de captura de conocimientos se da mediante un *proceso de aprendizaje*, donde dichos conocimientos son obtenidos mediante el estudio o la experiencia de distintas situaciones. El *aprendizaje* ocurre mediante la experiencia con el entorno, en un contexto cultural y social, en el que se combinan los nuevos conocimientos con estructuras cognitivas anteriores. El término aprendizaje, y todo lo que implica, ha sido objeto de estudio de diversas áreas del conocimiento [Russell y Norvig, 2004]. Una proyección del aprendizaje, en el sentido de un desprendimiento como objeto de estudio, es el *Aprendizaje Maquinal*. Este aprendizaje como una estrategia de reconocimiento de patrones es lo que se intenta definir en este capítulo desde las Ciencias de la Computación.

La capacidad de aprender es inherente a la inteligencia, por lo que el Aprendizaje Maquinal (AM) es un área central de la Inteligencia Artificial (IA). La Inteligencia Artificial tuvo sus inicios en los años cincuenta, y su objetivo era desarrollar máquinas capaces de pensar como humanos y construir entidades inteligentes [Russell y Norvig, 2004; Buchanan, 2005]. Con el tiempo, los objetivos se volvieron más concretos y más prácticos. La IA es un campo potencialmente universal y para dar respuestas a las diversas áreas, se han originado disciplinas específicas como la Minería de Datos, la Robótica o el Aprendizaje Maquinal. Este último, se enfoca en el razonamiento probabilístico, la investigación basada en la

Estadística, la recuperación de información y, cada vez más, en el reconocimiento de patrones [Witten *et al.,* 2016].

Otros nombres comúnmente utilizados para nombrar el Aprendizaje Maquinal son: Aprendizaje Automático, Aprendizaje de Máquina o Aprendizaje Computacional, aunque popularmente es conocido por su nombre en inglés *Machine Learning*. Tiene como objetivo la creación de algoritmos con la capacidad de mejorar su desempeño a través de la experiencia, es decir, algoritmos que puedan inducir modelos automáticamente. Aprendizaje Maquinal se refiere a la capacidad de un sistema de software para generalizar en base a la experiencia pasada, y utilizar estas generalizaciones para proporcionar respuestas a preguntas relacionadas con nuevos datos que el sistema nunca antes había incorporado [Witten *et al.,* 2016].

En este contexto, el concepto de aprendizaje está relacionado con el reconocimiento o extracción de un conjunto de patrones que permitan desde llegar a conclusiones acerca del fenómeno observado, o generalizar de alguna forma las observaciones individuales obtenidas, hasta determinar reglas que informen acerca de la estructura que estas observaciones presentarían bajo cualquier supuesto adicional o cambio de condiciones. Es decir, ese patrón (o patrones) aprendido, puede tomar diferentes formas para representar las relaciones observadas, que van desde el reconocimiento de bloques de datos que sean coherentes con el fenómeno, hasta leyes generales que predigan las condiciones futuras del objeto (Figura 2.1).

Figura 2.1. Esquema del proceso simplificado de Aprendizaje Maquinal.

Generalmente, se crean modelos o programas para resolver diversos problemas. Aunque algunos problemas son difíciles de formalizar porque no se tiene demasiada información o no se cuenta con expertos en el dominio para guiar el desarrollo, los algoritmos de aprendizaje permiten generar automáticamente

modelos, a partir de datos, para resolver estos tipos de problemas. La relevancia del Aprendizaje Maquinal y la adopción general de esta tecnología, se deben a factores como la disponibilidad de una gran cantidad de datos de todo tipo y de computadoras con gran capacidad de almacenamiento y velocidad de procesamiento, además de la facilidad de utilizar sus herramientas y métodos para resolver problemas reales.

Según Tom Michael Mitchell: "se dice que un programa de computadora aprende de la experiencia E con respecto a una clase de tareas T y medida de desempeño D, si su desempeño en las tareas en T, medidas con D, mejoran con la experiencia E." [Mitchell, 1997]. Los objetivos principales del Aprendizaje Maquinal son abordar y resolver problemas prácticos. El enfoque de desarrollo de los métodos de Aprendizaje Maquinal incluye aprender de los datos de entradas, evaluar y optimizar los resultados del modelo y, por ende, depende en gran medida de los datos, ya que cuanto más y mejores sean estos, más eficiente será el aprendizaje logrado.

Con el surgimiento de las nuevas tecnologías, y los dispositivos asociados, se cree que una gran cantidad de datos se creará en los próximos lustros. De hecho, el 90% de los datos actuales se crearon en los últimos años [Al-Jarrah *et al.,* 2015]. La adopción masiva del Aprendizaje Maquinal se debe también al aumento de la disponibilidad y democratización de las herramientas de software libre, desarrolladas tanto desde las comunidades científico-educativas, como de gigantes tecnológicos como Google con Tensorflow[1] como mayor ejemplo. El universo de librerías y de entornos de trabajo (*frameworks)* disponibles en el área crecen rápidamente, siendo los lenguajes más empleados Python y R. Python se considera en el primer lugar de los lenguajes utilizados para el desarrollo de IA debido a su simplicidad, mientras que R es uno de los lenguajes y entornos más efectivos para analizar y manipular los datos con fines estadísticos. Por supuesto, existen muchas alternativas como, por ejemplo, C, C++, Java, JavaScript, Julia, entre otros.

En muchas ocasiones el campo de acción del Aprendizaje Maquinal se solapa con el de Minería de Datos (*Data Mining*), ya que ambas disciplinas se enfocan en el análisis de datos. También, está relacionado con Datos Masivos (*Big Data)*, ya que es capaz de asimilar una amplia gama de datos, percibiéndolos como una lista de ejemplos prácticos. El objetivo principal de un algoritmo de aprendizaje es desempeñarse con precisión y exactitud, tanto en tareas que le son familiares,

[1] TensorFlow es una biblioteca de software de código abierto para el desarrollo y entrenamiento de modelos de Aprendizaje Maquinal. Web: https://www.tensorflow.org/

como en escenarios nuevos o imprevistos. Existen metodologías generales para producir un aprendizaje de forma automática y métodos para medir el grado de éxito o fracaso de un aprendizaje.

2.2. División de los Datos

El Aprendizaje Maquinal es una estrategia de aprendizaje de una solución a partir de patrones presentes en los datos. La calidad y la cantidad de estos datos son entonces, fundamentales para obtener un aprendizaje que pueda considerarse exitoso. Cuando se cuenta con un conjunto de datos depurados, lo primero que debe realizarse es dividirlos en diferentes grupos, que tendrán diferentes tareas asignadas durante el proceso de aprendizaje y evaluación del mismo (Figura 2.2).

Figura 2. 2 Esquema de la división de los datos originales en los conjuntos de Entrenamiento, Validación y Prueba.

2.2.1. Conjunto de Datos de Entrenamiento

Para el proceso de aprendizaje o fase de entrenamiento, se selecciona un subconjunto del total de los datos disponibles, generalmente, equivalente al 60% o al 80%, conocido como *conjunto de datos de entrenamiento* o simplemente *conjunto de entrenamiento* (*training dataset*). Estos datos son utilizados como ejemplos, para entrenar o *enseñar* al algoritmo a encontrar patrones o relaciones en los datos. Los algoritmos de aprendizaje realizan predicciones o toman decisiones basadas en datos de entrenamiento mediante la construcción de un modelo matemático. Es decir, el modelo ajusta los parámetros a partir de ese conjunto de datos de entrada. Para que el modelo generalice (pueda ser aplicado a otros datos) de forma correcta, el conjunto de entrenamiento debe ser lo suficientemente grande como para generar resultados significativos desde el punto de vista estadístico. Además, la muestra (conjunto de entrenamiento) debe ser representativa de todo el conjunto original de datos.

2.2.2. CONJUNTO DE DATOS DE VALIDACIÓN

En la fase de validación se realiza el ajuste de parámetros para optimizar el modelo. El objetivo es generar un modelo que sea preciso con los nuevos datos y no solo con aquellos que está utilizando para construirlo. El *conjunto de datos de validación* o *conjunto de validación* se utiliza para ajustar los parámetros del modelo, y está constituido por aproximadamente el 20% del total de los datos. Este conjunto permite realizar una evaluación imparcial del ajuste del modelo obtenido del conjunto de datos de entrenamiento, mediante la variación de los parámetros, en pruebas sucesivas.

2.2.2.a. VALIDACIÓN CRUZADA

La fase de validación cruzada (cross-validation o *k-fold cross-validation*), al igual que la fase de validación, tienen como objetivo probar la capacidad del modelo para predecir nuevos datos que no se usaron en la fase de entrenamiento. La particularidad de la validación cruzada es que implica dividir la base de datos en k segmentos (*folds*), submuestras o pliegues complementarios, entrenar al modelo con un subconjunto $(k-1)$ y validarlo en el otro subconjunto (k). Es decir, se utilizan la mayoría de los segmentos para el entrenamiento y el segmento restante para la validación (Figura 2.3). Este proceso se repite k veces para reducir la variabilidad del rendimiento, se realizan múltiples iteraciones utilizando las diferentes particiones, y los resultados de cada validación se combinan a lo largo de las iteraciones para dar una estimación final del rendimiento del modelo. Por esta razón, puede interpretarse al conjunto completo de datos como de entrenamiento y prueba a la vez, porque lo que se realiza es la prueba con un subconjunto y el resto se usa para el entrenamiento, la cantidad de veces que sea necesario para que cada subconjunto haya actuado como entrenamiento y prueba.

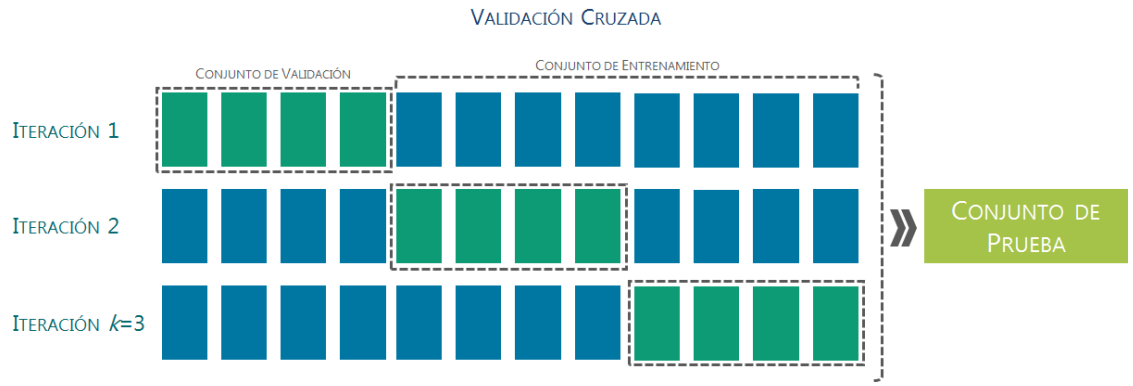


FIGURA 2. 3 ESQUEMA DEL PROCESO DE VALIDACIÓN CRUZADA.

La validación cruzada de k-pliegues es una técnica no exhaustiva, es decir, no se computan todas las formas de dividir la muestra original, sino que se la divide aleatoriamente en k submuestras de igual tamaño. No existe un estándar para la elección de k, aunque en la práctica lo más común es la utilización de $k = 10$, conocida como Validación Cruzada de 10 iteraciones (*10-fold cross-validation*). Sin embargo, cuando la muestra es pequeña se recomienda utilizar $k = 1$, y la técnica toma el nombre de Validación Cruzada Dejando Uno Fuera, conocida generalmente por su denominación en inglés *Leave One Out Cross Validation* (LOOCV). En esta técnica una sola instancia es utilizada como dato de prueba y el resto, como datos de entrenamiento en cada una de las iteraciones, que se repite n veces, siendo n el número de muestras originales (Figura 2.4).

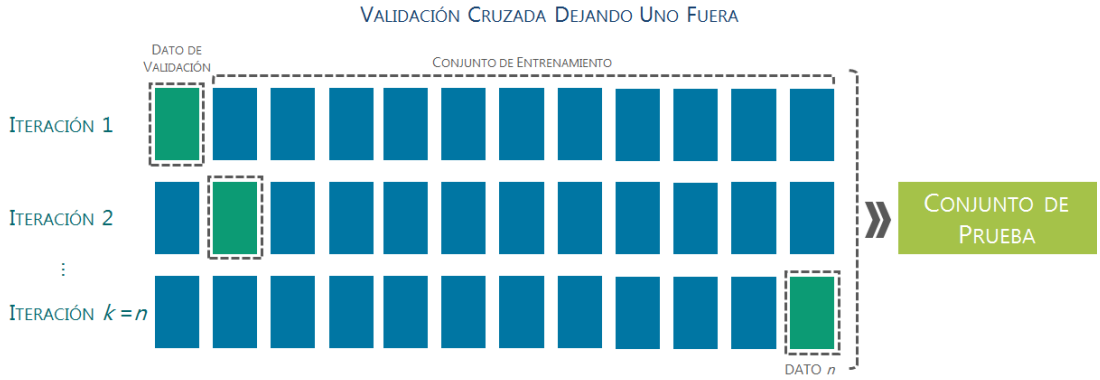


FIGURA 2. 4 ESQUEMATIZACIÓN GRÁFICA DEL PROCESO DE VALIDACIÓN CRUZADA DEJANDO UNO AFUERA.

Una de las principales desventajas que presenta este método es el alto costo computacional. Mientras que una de sus ventajas es que la estimación del error tiende a no ser variable, y también, que este no está sobreestimado. Aun así, es recomendable apartar un porcentaje de la totalidad de los datos, generalmente un 20%, y que no sean utilizados en este proceso, para realizar una validación externa, es decir, un proceso en el que no se involucraran datos que fueron utilizados para el entrenamiento del modelo. Resumiendo, por un lado, existe el conjunto de entrenamiento y, por el otro, el conjunto de validación; entre ambos representan aproximadamente el 80% del conjunto total de los datos. Cuando el conjunto de datos de validación no existe como tal, se recomienda el uso de métodos de validación cruzada (también conocido como validación interna), donde el conjunto de datos de entrenamiento (que en este caso puede ascender generalmente hasta un 80%) es dividido en k pliegues y cada uno de esos pliegues funciona como datos de validación en una de las interacciones y en las demás, forma parte del conjunto de entrenamiento. El 20% restante de los datos se conoce como conjunto de prueba y se utiliza en lo que se conoce como proceso de validación externa.

Ambos procedimientos de testeo proporcionan información diferente y complementaria en lo que respecta a la evaluación de un modelo.

2.2.3. CONJUNTO DE DATOS DE PRUEBA

La fase prueba o testeo, representa una parte fundamental en un experimento. Permite determinar si el modelo ha *aprendido* patrones o relaciones lo suficientemente sensibles, evalúa la capacidad del mismo para predecir la variable objetivo y si pueden generalizarse los resultados o patrones aprendidos. Es necesario contar con datos que no hayan participado en ninguna de las fases anteriores (fase de entrenamiento y fase de validación). Estos datos se conocen como conjuntos de prueba o conjuntos de testeo, también llamado conjuntos de validación externa, y generalmente equivale al 20% de los datos iniciales. Sobre este conjunto, se aplica el modelo para calcular diferentes métricas estadísticas que establezcan el grado de generalización en las diferentes instancias, evalúen el rendimiento del modelo final y garanticen que haya aprendido los patrones correctos de los datos y que no sea demasiado sensible al ruido, es decir, que no haya demasiadas perturbaciones que tiendan a enmascarar la información.

A menudo se espera que los datos del conjunto de testeo sean representativos de los nuevos datos, que el modelo deberá predecir en el futuro. Para lidiar con este problema una de las técnicas que existen es el Dominio de Aplicación que determina el espacio de información en el que un modelo puede generar una predicción confiable.

2.3. TIPOS DE MODELOS

Los modelos aprendidos a través de Aprendizaje Maquinal, identifican patrones o relaciones que pueden servir para predecir resultados. Dependiendo del tipo de resultado, es decir, de cuál es el número y la clase de la variable de salida puede clasificarse a los modelos en tres grandes grupos. A saber, cuando la clase es continua se trata de un modelo de regresión, en cambio, cuando la clase de salida es discreta se llama modelo de clasificación, y cuando lo importante es el orden óptimo de los datos de salida, se refiere un modelo de ranking (Figura 2.5).

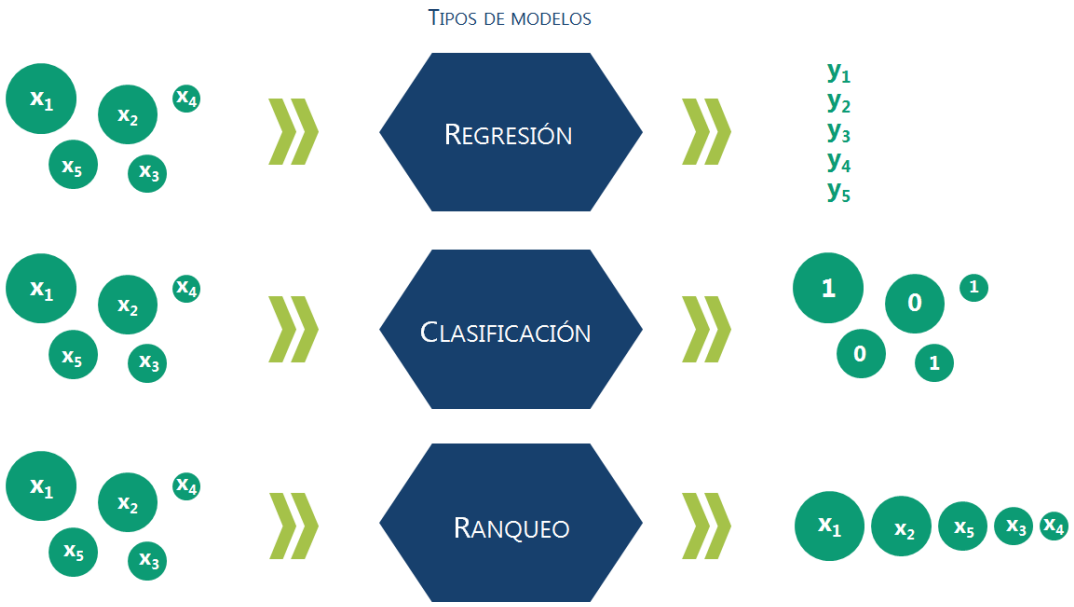


FIGURA 2. 5 DIFERENTES TIPOS DE MODELOS DE APRENDIZAJE AUTOMÁTICO SEGÚN LA SALIDA QUE PERMITEN.

2.3.1. REGRESIÓN

Los modelos de regresión, se utilizan para predecir el valor de una variable continua, un valor real. También son conocidos como modelos de estimación, ya que en estadística regresión hace referencia al proceso por el cual se estima la relación entre variables. Consiste en un modelo matemático, usado para aproximar la relación de dependencia entre una variable dependiente y, las variables independientes x_i y un término independiente b. Este tipo de modelo puede ser expresado en forma de recta, el término regresión lineal se emplea para distinguirlo del resto de técnicas de regresión que utilizan modelos matemáticos basados en otras funciones matemáticas.

2.3.2. CLASIFICACIÓN

Los modelos de regresión están muy relacionados con los modelos de clasificación. La regresión se aplica cuando la *clase* a predecir se compone de valores numéricos continuos. En cambio, modelo de clasificación hace referencia a la predicción de clases discretas prefijadas. Si solo están permitidas dos clases posibles, se llama clasificación binaria. Esta es el método más simple de clasificación, donde se clasifica generalmente en 1 o 0 a los datos de entrada. En cambio, si existen más de dos clases o categorías, es una extensión de la clasificación binaria y se trata de clasificación multiclase o clasificación múltiple.

2.3.3. RANKING

Los modelos de ranqueo o simplemente ranking, intentan predecir el orden óptimo de un conjunto de objetos según una relevancia definida previamente. Generan listas de estos objetos con un orden parcial especificado, mediante permutaciones de los elementos en listas nuevas. Este orden es inducido generalmente mediante una puntuación ya sea numérica u ordinal. Los modelos de ranqueo son ampliamente utilizados, uno de los ejemplos más usados para describirlos es el de los buscadores de internet, que devuelven en una lista ordenada los recursos encontrados como respuesta a una búsqueda de un usuario.

2.4. MÉTRICAS ESTADÍSTICAS

Los algoritmos inteligentes, al igual que la inteligencia humana, no están exentos de equivocación [Witten *et al.,* 2017], por lo cual es necesario medir en cierto grado la confianza, por ejemplo, con que los algoritmos de aprendizaje pueden realizar una predicción. Las métricas estadísticas permiten evaluar diferentes aspectos de los algoritmos de aprendizaje. Para entender el comportamiento de estos, existen dos conceptos claves en el Aprendizaje Maquinal: sobreajuste (*overfitting*) y subajuste (*underfitting*). El sobreajuste se da cuando el modelo mapea casi perfectamente la tendencia de los datos de entrenamiento, pero falla a menudo en la generalización de nuevos registros (Figura 2.6.c.). Para enfrentar este inconveniente pueden reducirse el número de variables que definen el patrón a replicar o eliminar aquellas que sean no relevantes. El subajuste se da cuando el modelo mapea pobremente la tendencia de los datos de entrenamiento, con lo cual también tiene problemas con la generalización de nuevos registros (Figura 2.6.a.). Para combatirlo puede aumentarse el número de variables relevantes o, si es posible, incluir un número mayor de registros (ejemplos o instancias) al conjunto de entrenamiento.

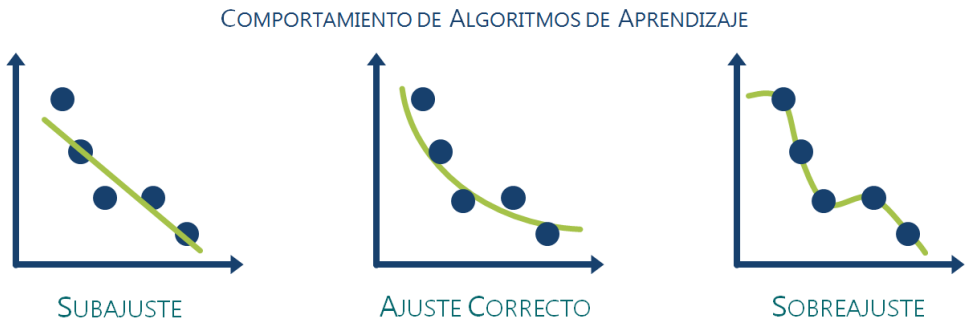


FIGURA 2. 6. ESQUEMA DEL COMPORTAMIENTO DE LOS ALGORITMOS DE APRENDIZAJE AUTOMÁTICO.

Existen métricas estadísticas que permiten evaluar un modelo y compararlo con otros. Dependiendo el tipo de modelo y lo que midan, estas métricas pueden ser variables cuantitativas o cualitativas. Estas métricas comparan los valores reales con sus predicciones o estimaciones, con el fin de estimar la *distancia* a la que se encuentran una de la otra. Lo hacen de diferentes maneras, empleando valores absolutos, relativos, raíces cuadradas, entre otras herramientas matemáticas. Por ejemplo, las raíces hacen que las métricas sean más sensibles a valores atípicos o valores extremos (*outliers*). A continuación, se describen brevemente las distintas métricas de error utilizadas para reportar el rendimiento de los modelos inferidos en esta tesis:

- MSE: Error Cuadrático Medio (*Mean Squared Error*) es un estimador que mide el promedio de los errores cuadráticos de los N errores, donde N es el número de muestras, es decir, la diferencia cuadrática promedio entre el valor observado (y_i) y el valor calculado ($\hat{y}_i$), donde la media aritmética de los valores observados es $\bar{y}$ y la media aritmética de los valores calculados es $\overline{\hat{y}_i}$.

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

- RMSE: Error Cuadrático Medio (*Root Mean Squared Error*) es el promedio de las diferencias entre el valor real y el valor predicho al cuadrado.

$$RMSE = \left(\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2\right)^{1/2} = (MSE)^{1/2}$$

- MAE: Error Absoluto Medio (*Mean Absolute Error*) es el promedio de las diferencias entre el valor real y el valor predicho.

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|^2$$

- RAE: Error Relativo (*Relative Absolute Error*), es el cociente entre el error absoluto y el valor exacto.

$$RAE = \frac{\sum_{i=1}^{N}|y_i - \hat{y}_i|}{\sum_{i=1}^{N}|\hat{y}_i - \overline{\hat{y}_i}|}$$

- RRSE: Error Cuadrático Relativo (*Root Relative Squared Error*) es el error cuadrado total normalizado por el error cuadrado total del estimador simple.

$$RRSE = \left(\left(\sum_{i=1}^{N} (y_i - \hat{y}_i)^2 \right) / N \right)^{1/2}$$

- MAPE: Error Porcentual Medio Absoluto (*Mean Absolute Porcentual Error*) es el promedio del error absoluto o la diferencia entre el valor real y el valor predicho, expresado como un porcentaje de los valores reales. Brinda información en términos porcentuales y no, en unidades.

$$MAPE = \frac{100\%}{N} \sum_{i=1}^{N} \left| \frac{y_i - \hat{y}_i}{y_i} \right|$$

- R^2: Coeficiente de Determinación (*Determination Coefficient*) es un estadístico que determina la calidad del modelo para replicar los resultados y sirve para explicar la proporción de variación de los resultados que puede darse por el modelo.

$$R^2 = 1 - \frac{1}{N} \left(\frac{\sum_{i=1}^{N} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{N} (y_i - \bar{y}_i)^2} \right)$$

- CC%: Porcentaje de Casos Correctamente Clasificados (*Correctly Classified Percentage*) es el porcentaje del total de los casos para los que la predicción fue realizada correctamente. Esta métrica también es comúnmente conocida como Precisión (*Accuracy*). No es considerada una métrica útil cuando las clases con las que se trabajan tienen tamaños muy diferentes (desbalanceadas).

$$Precisión\ (\%CC) = (Predicciones\ correctas)/(Número\ total\ de\ Predicciones)$$

- KAPPA: Kappa (K) indica el grado de acuerdo que existe por encima del esperado por azar. Es decir, es una medida de qué tan cerca las instancias clasificadas por el clasificador de aprendizaje automático (clase predicha) coinciden con los datos etiquetados (clase real). Es un estadístico que se usa para medir la confiabilidad de un clasificador con otro en particular o con el azar. Puede tomar valores entre 0 y 1, si las clasificaciones están en total acuerdo entonces $K = 1$, si están en total desacuerdo $K = 0$. La fórmula para medir el estadístico Kappa se indica a continuación, donde $\Pr(a)$ es el acuerdo observado relativo entre los clasificadores (precisión) y $\Pr(e)$ es es la probabilidad hipotética de acuerdo por azar (por ejemplo, si se trata de dos clases corresponde al 50%, si son tres clases a 33.33%, etc.):

$$K = \Pr(a) - \Pr(e)/1 - \Pr(e)$$

- Matriz de confusión: (*Confusion Matrix*) es una herramienta de visualización del desempeño de un algoritmo de clasificación (Figura 2.7). Cada columna representa el número de predicciones de cada una de las clases, y cada fila representa las instancias de la clase real. De acuerdo a qué lugar en la matriz ocupan las predicciones representan: Verdadero Positivo (*True Positive*, TP) cuando existe una coincidencia positiva, Falso Positivo (*False Positive*, FP) para un error tipo I, verdadero negativo (*True Negative,* TN) cuando el rechazo es correcto, y falso negativo (*False Negative,* FN) para un error tipo II.

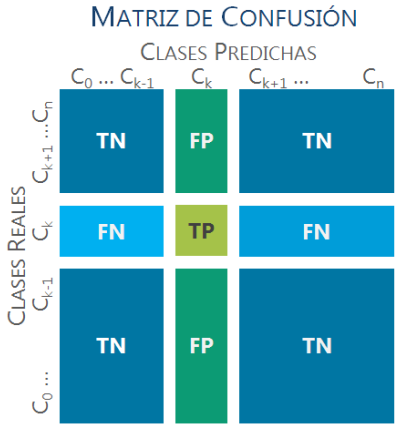


FIGURA 2. 7. REPRESENTACIÓN DE UNA MATRIZ DE CONFUSIÓN MULTICLASE.

- Especificidad: (*Specificity*) al igual que la sensibilidad, es una medida estadística del rendimiento de una clasificación. Es la tasa de TN y mide la proporción de negativos reales que se identifican correctamente como tales. Para definir esta métrica, y las que siguen, utilizaremos un lenguaje común donde a será el número de predicciones correctas de clase negativa (TN), b será el número de predicciones incorrectas de clase positiva (FP), c será el número de predicciones incorrectas de clase negativa (FN) y d será el número de predicciones correctas de clase positiva (TP). Para facilitar la comprensión la Figura 2.8 ilustra dichas convenciones.

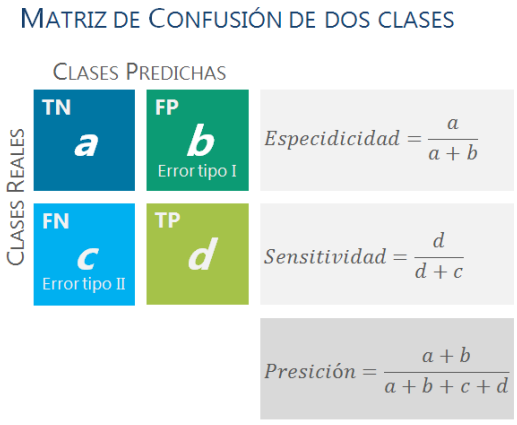


Figura 2. 8. REPRESENTACIÓN DE UNA MATRIZ DE CONFUSIÓN 2X2.

$$\boldsymbol{Especificidad = \frac{a}{a+b}}$$

- Sensibilidad: (*Sensitivity*) es la tasa de TP y mide la proporción de positivos reales que se identifican correctamente como tales.

$$Sensitividad = \frac{d}{d+c}$$

- Coeficiente de Correlación de Mathew: (*Matthews Correlation Coefficient*) es una medida de la calidad de las clasificaciones. Generalmente es considerado una medida equilibrada y puede ser utilizado en conjuntos de datos muy desbalanceados. El CCM es, en esencia, un coeficiente de correlación entre las clasificaciones observadas y predichas; devuelve un valor entre −1 y +1. Un coeficiente de +1 representa una predicción perfecta, 0 no es mejor que la predicción aleatoria y -1 indica un desacuerdo total entre la predicción y la observación.

$$CCM = \frac{(TP * TN) - (FP * FN)}{[(\mathrm{TP} + \mathrm{FP}) * (\mathrm{FN} + \mathrm{TN}) * (\mathrm{FP} + \mathrm{TN}) * (\mathrm{TP} + \mathrm{FN})]^{1/2}}$$

- Tasa de No Error, conocido por las siglas en inglés NER para Non-Error Rate o también llamada Precisión Equilibrada (Balanced Accuracy). El valor de NER es la media aritmética de la sensibilidad de clase, es decir, el promedio de los porcentajes de muestras que se clasificaron correctamente para cada clase.

$$NER = (Especificidad + Sensitiviad)/2$$

- Área de la curva de ROC: Área de la Curva de Característica Operativa del Receptor (*Average Receiver Operating Characteristic Curve Area*) es utilizada para comparar el rendimiento entre modelos. Implica el espacio definido por la tasa de TP como eje de las abscisas, y la tasa de FP como eje de las ordenadas. A mayor área, mejor será la predicción [Hanley, 1982].

$$Tase\ de\ Falsos\ Positivos = \frac{b}{a+b}$$

$$Tasa\ de\ Falsos\ Negativos = \frac{c}{c+d}$$

2.5. ENFOQUES DE APRENDIZAJE

El Aprendizaje Maquinal está formado por un conjunto de algoritmos que aprenden y resuelven problemas gracias a la experiencia, es decir, reconociendo patrones o relaciones en el conjunto de datos de entrenamiento. De acuerdo a la forma en que estos algoritmos *aprenden* pueden clasificarse en distintas categorías. La clasificación más conocida se basa en las diferencias del conocimiento *a priori* que se tiene. En el enfoque supervisado se conocen previamente los datos de salida deseada, mientras que el enfoque no supervisado está caracterizado por la ausencia de ese conocimiento previo. A continuación, se presenta una explicación más detallada de estos enfoques.

2.5.1. APRENDIZAJE SUPERVISADO

La experiencia a través de la cual los algoritmos de Aprendizaje Maquinal aprenden debe ser generada. Esto se consigue a través de ejemplos con los que se entrena al algoritmo, para luego aplicarlos sobre otras instancias distintas. Cuando ese conjunto de ejemplos está formado por datos previamente etiquetados y clasificados, es decir, se sabe a qué grupo, valor o categoría pertenecen, se trata de un Aprendizaje por Enfoque Supervisado (*Supervised Machine Learning*). Estos ejemplos funcionan como entrada del modelo para que pueda aprender o entrenarse a partir de ellos. Este enfoque consta como mínimo de dos fases: entrenamiento y prueba.

Como se mencionó antes, en el aprendizaje supervisado, para un conjunto de datos de entrada, se conocen *a priori* los datos deseados de salida. Entonces, una vez que el modelo aprendió una regla general, es capaz de asignar entradas a la salida. La estructura de datos de un aprendizaje por Enfoque Supervisado se compone generalmente de los datos de entrada que se encuentran en formato de tabla, cada fila representa una instancia, es decir, un ejemplo y cada columna un atributo o característica. Además, se compone por otra columna, habitualmente ubicada en último lugar (o separada de la tabla), que representa la variable objetivo (*target)* (Figura 2.9).

FIGURA 2. 9 ESQUEMA DEL PROCESO DE APRENDIZAJE DE UN ALGORITMO DE APRENDIZAJE MAQUINAL SUPERVISADO.

2.5.2. APRENDIZAJE NO SUPERVISADO

En el Aprendizaje por Enfoque no Supervisado (*Unsupervised Machine Learning*), no se conocen *a priori* los datos de salida para el conjunto de datos de entrenamiento. El proceso de aprendizaje se lleva a cabo sobre un conjunto de ejemplos formados únicamente por entradas al sistema, sin conocer su correcta clasificación (salida). Este tipo de enfoque tiene como objetivo encontrar una estructura en las entradas, se busca que el modelo sea capaz de reconocer patrones para poder etiquetar las nuevas entradas. Generalmente, se utiliza para generar grupos coherentes (*cluster*) de las instancias, en función de las relaciones que se establecen entre las variables de cada una de ellas (Figura 2.10).

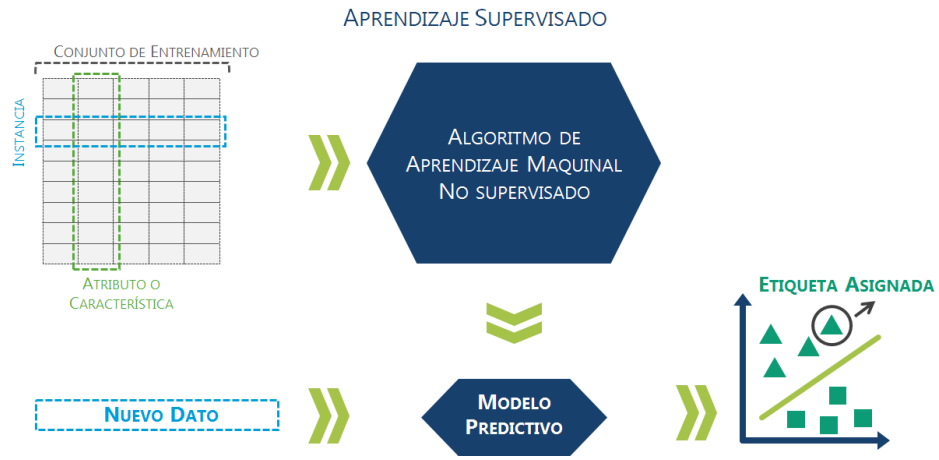


FIGURA 2. 10 ESQUEMA DEL PROCESO DE APRENDIZAJE DE UN ALGORITMO DE APRENDIZAJE MAQUINAL NO SUPERVISADO.

2.5.3. MÁS ALLÁ DEL APRENDIZAJE (NO) SUPERVISADO

El Aprendizaje totalmente Supervisado, utiliza una gran cantidad de datos de entrenamiento que, necesariamente, deben estar curados. La cantidad de datos de entrenamiento requeridos, para algunas aplicaciones, pueden generar problemas de escalabilidad prohibitivos. Una estrategia que podría lidiar con esto es el Aprendizaje de Enfoque Semisupervisado, ya que los algoritmos dentro de este enfoque cuentan con datos etiquetados, se conoce *a priori* la salida, así como con datos no etiquetados. Esta alternativa nos brinda la ventaja de los enfoques supervisados y reduce la necesidad de contar con la totalidad de los datos etiquetados.

Otra estrategia, son los algoritmos de Aprendizaje por Enfoque de Refuerzo (*Reinforcement Machine Learning)*, utilizados para resolver problemas no supervisados que solo reciben re-alimentaciones del entorno o *recompensas*. Estos algoritmos generan modelos enfocados en maximizar su recompensa. Aprenden *observando* el mundo que los rodea, interactuando con un entorno dinámico con un flujo de información continuo, realizando un proceso de prueba y error, reforzando las acciones por las que recibe una respuesta positiva (Figura 2.11).

FIGURA 2. 11 ESQUEMA DEL PROCESO DE APRENDIZAJE DE UN ALGORITMO DE APRENDIZAJE MAQUINAL POR REFUERZO.

2.6. MÉTODOS DE APRENDIZAJE

Los algoritmos de Aprendizaje Maquinal utilizan métodos computacionales para aprender información directamente de los datos, sin necesidad de depender de una ecuación predeterminada como modelo. Algunos métodos solo pueden ser utilizados en cierto enfoque de aprendizaje para un tipo de modelo, mientras que otros son más versátiles. Existe una amplia variedad de métodos de aprendizaje

desarrollados a lo largo de las últimas décadas. En esta sección solo nos limitaremos, en particular, a comentar sintéticamente aquellas técnicas que fueron empleadas a lo largo de la presente tesis. Para un mayor detalle sobre cómo funcionan las mismas, se recomienda la lectura de los artículos o libros referenciados en cada caso. Es necesario aclarar, que la amplia mayoría de las implementaciones de los métodos utilizados para el desarrollo de esta tesis fueron provistas por WEKA[2] [Frank *et al.,* 2004; Hall *et al.*, 2009; Witten *et al.*, 2016].

2.6.1. REGRESIÓN LINEAL

La regresión simple establece la relación entre variables independientes y dependientes ajustando los modelos a la ecuación de la recta ($y = a * x + b$). Este tipo de regresión se conoce como *regresión lineal simple*. También existe, *la regresión lineal múltiple* que se caracteriza por tener más de una variable independiente [Brownlee, 2016].

2.6.2. REDES NEURONALES

Las Redes Neuronales (*Artificial Neural Networks*) son uno de los algoritmos bioinspirados más conocidos. Los primeros trabajos en Inteligencia Artificial se enfocaban en la creación de redes neuronales artificiales y, actualmente, siguen siendo una de las tecnologías de tendencia. Christopher Michael Bishop en su libro "Redes neuronales para el reconocimiento de patrones" hace un recorrido por este tipo de método de Aprendizaje Maquinal enfocado al reconocimiento de patrones, el lector puede consultarlo para más detalles [Bishop, 1995]. Los tipos de redes empleadas en esta tesis se describen brevemente a continuación. En este punto es necesario aclarar que cuando nos refiramos, en los próximos capítulos, simplemente a "Redes Neuronales" estaremos haciendo referencia a redes del tipo Perceptron Multicapa; caso contrario será debidamente especificado el tipo de red empleada.

PERCEPTRON SIMPLE

El aprendizaje es estudiado y analizado por numerosas, y combinadas, disciplinas. Tanto es así, que el primer modelo simple de neurona artificial, basado en el modelo de McCulloch (neurocientífico) y Pitts (lógico del campo de la neurociencia computacional) [McCulloch & Pitts, 1943], fue propuesto por el psicólogo Frank Ronsenblant [Rosenblantt, 1958]. Este modelo se llama *Perceptrón*

[2] WEKA es un entorno de propósito general para la minería de datos que contiene una amplia colección de algoritmos de aprendizaje automático. Web: https://ai.waikato.ac.nz/weka/

Simple y está constituido por un conjunto de sensores (neuronas) de entrada que reciben los patrones (o datos) de entrada y una neurona de salida. Es decir, está compuesto solo por dos capas de neuronas. El modelo es capaz de clasificar en dos clases o categorías, de forma automática, determinando la ecuación del plano discriminante [Bishop, 1995].

PERCEPTRON MULTICAPA

El *Perceptron Multicapa* surgió como una versión mejorada del Perceptron Simple, debido a las limitaciones para resolver problemas de separabilidad no lineal. La principal diferencia estructural se basa en que el Perceptrón Multicapa tiene varias capas de neuronas artificiales, que le permiten realizar clasificaciones en más de dos grupos. La propagación (o transmisión) de los patrones de entrada inicia en la *capa de entrada*, se propaga a la o las *capas ocultas* donde se realiza un procesamiento no lineal de los datos para, finalmente, llegar a la *capa de salida* que proporciona al exterior la respuesta de la red para cada patrón de entrada. La propagación puede ser hacia delante o hacia atrás (conocido como retropropagación o *back propagation* en inglés) [Bishop, 1995].

MAPAS AUTO ORGANIZADOS

Los Mapas Auto Organizados (*Self-Organizing Map)* más conocidos por su sigla en inglés *SOM*, son algoritmos de agrupamiento (*clustering*). Se entrenan con un enfoque de aprendizaje no supervisado para producir una representación discreta del espacio (*mapa*) de las muestras de entrada, preservando las propiedades topológicas de ese espacio [Miljković, 2017].

2.6.3. BOSQUES ALEATORIOS

Los Bosques Aleatorios (*Random Forest*s) fueron propuestos por Breiman en 2001 [Breiman, 2001; 2017], son uno de los algoritmos de mayor uso en la comunidad. Sirven tanto para problemas de clasificación como de regresión y tienen buen rendimiento para datos de alta dimensionalidad. Un bosque aleatorio se compone de un conjunto de *árboles de decisión*. Cada árbol clasifica una nueva instancia, basado en atributos (o características). Es decir, el árbol *vota* por una clase y la clasificación que brinda el bosque, es aquella que obtuvo más votos.

ÁRBOLES DE DECISIÓN

Los Árboles de Decisión (*Decision Tree*) actualmente también son llamados Árboles de Regresión y Clasificación (*Classification and Regression Trees*, CART) son usados para resolver problemas de clasificación tanto de variables dependientes

categóricas como continuas. En base a las características más significativas, el árbol divide a la población en dos o más grupos tan distintos como sea posible [Timofeev, 2004, Brownlee, 2016]. En un Árbol de Decisión cada bifurcación divide en una variable predictiva y cada nodo final contiene una predicción para la variable resultado [Loh, 2011]. Existen varios algoritmos para construir árboles de decisión, como por ejemplo los basados en *Greedy Splitting* (división codiciosa) que buscan o prueban diferentes puntos de división basándose en una función de costo (se busca minimizar el costo). Se sigue un procedimiento de división binaria recursiva hasta que la cantidad de instancias de entrenamiento asignadas a cada nodo hoja sea menor a un mínimo, entonces la división no se acepta y el nodo se toma como un nodo hoja final [Brownlee, 2016].

2.6.4. COMITÉ ALEATORIO

El Comité Aleatorio (*Random Committee*) es una generalización de los Bosques Aleatorios, ya que se construye a partir de un conjunto de diferentes clasificadores de base aleatoria. La predicción del Comité Aleatorio, para una entrada nueva, se obtiene mediante una combinación de las predicciones individuales de cada miembro del comité (clasificadores diferentes). Funciona según el enfoque de *mezcla de expertos*, motivado en que la combinación de las predicciones de los distintos expertos puede ayudar a generar predicciones finales sin sesgo [Tresp, 2001].

2.6.5. OTROS MÉTODOS

Además de los métodos antes mencionados, en esta tesis se han empleado, aunque menos frecuentemente, los métodos que se listan a continuación, destacando alguna particularidad de estos cuando se lo considere relevante. Como se mencionó anteriormente, la implementación de todos estos métodos fueron provistas por WEKA [Witten *et al.*, 2016].

- Bayes Simple: Los modelos de Bayes Simples (*Naïve Bayes*) son los clasificadores probabilísticos más usados en Aprendizaje Maquinal. Se basan en el *teorema de Bayes*, y se llama *simple* o *ingenuo* porque asume que las características (o atributos) son condicionalmente independientes entre sí. Es un método simple y, particularmente, útil para grandes conjuntos de datos donde se verifican dichas suposiciones de independencia entre las variables [Brownlee, 2016]. En WEKA recibe el nombre de weka.classifiers.bayes.NaiveBayes y se basa en el trabajo de John y Langley [John & Langley, 1995].

- K Vecinos más cercanos: Los clasificadores de k vecinos más cercanos (*k-nearest neighbours, k-nn*), pueden seleccionar el valor apropiado para k siguiendo un proceso de validación cruzada. La k significa la cantidad de puntos vecinos que se tienen en cuenta en las cercanías para clasificar en los tantos grupos, que ya se conocen *a priori*, porque es un algoritmo supervisado. Entre los métodos para realizar ponderación más utilizados están aquellos que ponderan la distancia [Aha *et al.*, 1991]. En WEKA recibe el nombre de IBk y utilizada la distancia euclídea para determinar los vecinos.

- K estrella: el clasificador K* o *k-star* es un clasificar basado en instancias, es decir, la clase de una instancia de prueba se basa en la clase de las instancias de entrenamiento similares a ella, según lo determinado por una función de similitud. En WEKA recibe el nombre de weka.classifiers.lazy.KStar y a diferencia de otros clasificadores basados en instancias usa una función de distancia basada en entropía llamada entropicAutoBlend. [Cleary & Trigg, 1995].

- *LogitBoost*: es un clasificador para realizar regresión logística aditiva, realiza la clasificación utilizando un esquema de regresión como método de aprendizaje maquinal base, y puede manejar problemas de varias clases [Friedman *et al.*, 2000]. En WEKA recibe el nombre de weka.classifiers.meta.LogitBoost y utiliza por defecto Árboles de Decisión como método de aprendizaje.

- *Bagging*: también llamado *Bootstrap Aggregating*, es una técnica que consiste en crear diferentes modelos usando muestras aleatorias con reemplazo, y luego combinar los resultados. Dependiendo del método de aprendizaje utilizado puede ser implementado tanto para modelos de regresión como de clasificación. Su principal ventaja es la disminución de la varianza al realizar remuestreo con reemplazo. En WEKA recibe el nombre de weka.classifiers.meta.Bagging [Breiman, 1996].

- Árbol de Hoeffding: *Hoeffding tree* también conocido como VFDT (*Very Fast Decision Tree learner*) es un algoritmo de inducción de árbol de decisión incremental capaz de aprender de flujos de datos, suponiendo que la distribución que genera ejemplos o nuevas instancias no cambia con el tiempo. Los árboles Hoeffding explotan el hecho de que una pequeña muestra a menudo puede ser suficiente para elegir una característica o atributo de división óptimo (límite de Hoeffding) [Hulten *et al.*, 2001]. En WEKA, recibe el nombre de weka.classifiers.trees.HoeffdingTree, y utiliza *Naïve Bayes* como predictor del

umbral, es decir, para identificar el número de instancias (peso) que debe observar una hoja antes de permitir que el método haga predicciones.

2.7. REDUCCIÓN DE LA DIMENSIONALIDAD

En los últimos años, el aumento en la captura de datos es exponencial y debido a ese aumento la *Reducción de la Dimensionalidad* ha cobrado mayor importancia. Los datos provienen de todas las etapas de los procesos, y cada vez son más detallados. Sin embargo, no todos pueden ser utilizados para todas las tareas que se quieran analizar. Entonces surgen algunas preguntas: ¿es necesario explorar todas y cada una de las características o atributos?, ¿Son todas igualmente importantes para determinada tarea? Los algoritmos de Aprendizaje Maquinal pueden identificar las variables más significativas automáticamente, mediante un proceso de filtrado o reducción de la dimensionalidad, para obtener las variables que sean altamente significativas para esa tarea dentro de ciento de miles de variables disponibles. Los métodos más conocidos de este campo son la Selección de Características y Aprendizaje de Características (Figura 2.12).

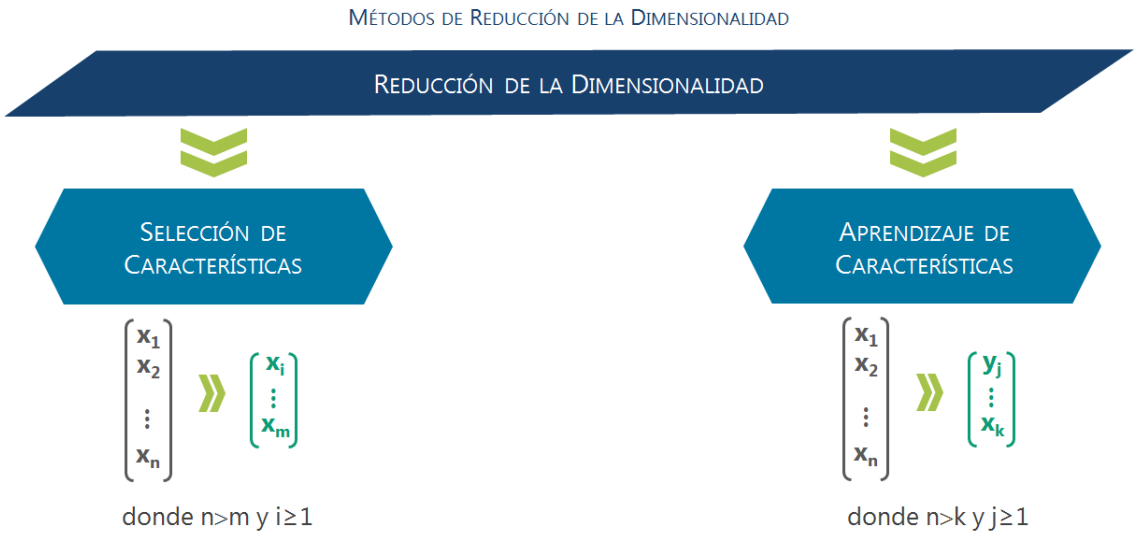


FIGURA 2. 12 MÉTODOS DE REDUCCIÓN DE LA DIMENSIONALIDAD MÁS CONOCIDOS.

2.7.1. SELECCIÓN DE CARACTERÍSTICAS

Los métodos de Selección de Características (*Feature Selection*) forman parte de los métodos de reducción de la dimensionalidad para sistemas de reconocimiento de patrones [Li *et al.*, 2018]. Dentro del campo del Aprendizaje Maquinal, el objetivo de estos métodos es obtener, del conjunto completo de características (variables), un subconjunto relevante para la construcción de un modelo (Figura 2.13). Una variable completamente inútil por sí misma, puede

generar una mejora en el desempeño del modelo si es tomada junto con otras [Cai *et al.,* 2018.]. Se asume que muchas de las variables pueden ser irrelevantes o redundantes para el modelo a diseñar. Las variables irrelevantes son aquellas que no aportan información útil para el armado del modelo. Por otro lado, las variables redundantes son aquellas que no proveen información adicional respecto de las restantes características seleccionadas.

FIGURA 2. 13 REPRESENTACIÓN GRÁFICA DEL PROCESO DE SELECCIÓN DE CARACTERÍSTICAS.

Los métodos de selección pueden ser tanto óptimos como subóptimos. Buscar el subconjunto óptimo es costoso computacionalmente, cuando no imposible, por lo que generalmente el proceso de selección evalúa el costo de agregar o quitar una característica, hasta cumplir con las restricciones impuestas previamente. La navaja de Ockham o ley de parsimonia es un principio metodológico según el cual, en igualdad de condiciones, la explicación más sencilla suele ser la más probable. Es decir, cuanto menos complejo sea un modelo de Aprendizaje Maquinal, más probable será que un resultado empírico correcto no se deba simplemente a las peculiaridades de la muestra. Es un principio muy utilizado para la reducción de la dimencionalidad en multiples áreas de aplicación [Ebrahimpour *et al.,* 2017]. Sin embargo, existe evidencia empírica que muestra su falla como práctica heurística (*la simplicidad conduce a una mayor precisión*), pero si es recomendable su uso para preferir u optar por modelos más comprensibles (*la simplicidad es una meta en sí misma*) [Domingos, 1999], ya que la baja cardinalidad de un modelo, es decir, el bajo número de variables presentes puede colaborar o ayudar a su interpretación [Khaire & Dhanalakshmi, 2019].

2.7.2. APRENDIZAJE DE CARACTERÍSTICAS

El Aprendizaje de Características (*Feature Learning*), también conocido como Extracción de Características (*Feature Extraction*), realiza una transformación del espacio del conjunto inicial de variables construyendo nuevos valores derivados o transformados (Figura 2.14) Estos valores deben ser relevantes y no redundantes, para el proceso de aprendizaje en el que serán luego utilizados. Deben contener la

misma información útil que los datos de entrada, para que la transformación pueda utilizarse en lugar de esos datos iniciales, sin pérdida de información [Hira & Gillies, 2015].

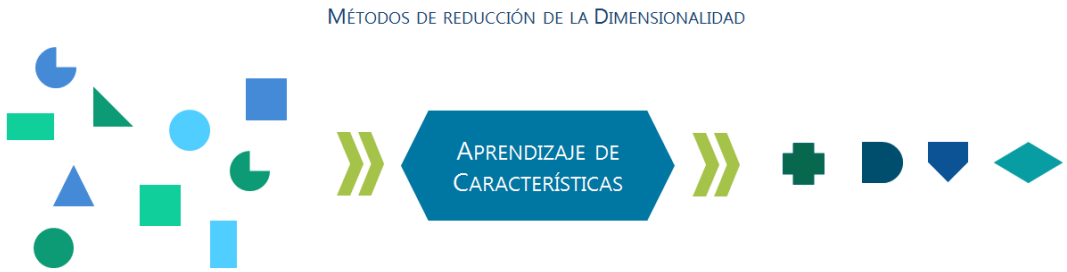


FIGURA 2. 14 REPRESENTACIÓN GRÁFICA DEL PROCESO DE APRENDIZAJE DE CARACTERÍSTICAS.

2.8. APRENDIZAJE PROFUNDO

Las investigaciones en redes neuronales sencillas o perceptrones, han evolucionado continuamente hasta llegar a contar con redes neuronales de numerosas capas ocultas, para dar origen al llamado Aprendizaje Profundo (*Deep Learning*) (Figura 2.15) [LeCun *et al.*, 2015; Goodfellow *et al.*, 2016; Voulodimos *et al.,* 2018]. Las redes neuronales generan modelos cada vez más grandes, que requieren mayor poder de cómputo para entrenarlos. Al emplear grandes cantidades de datos, los modelos tienden a ser mejores que los logrados con redes neuronales clásicas, pero lo hace a expensas de su interpretabilidad. Interpretabilidad que, en algunas disciplinas, es de vital importancia al menos hasta que el Aprendizaje Profundo sea aceptado por esa comunidad en particular y finalmente pueda ser adoptado [Kalousis *et al.*, 2007; Khaire & Dhanalakshmi, 2019].

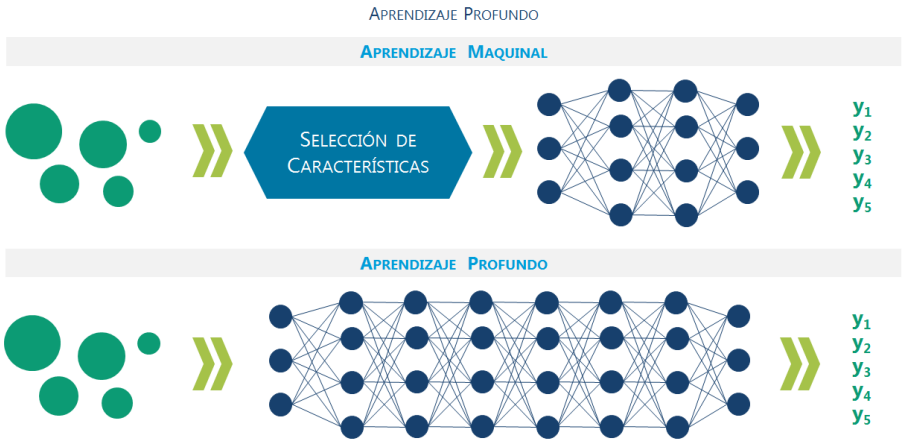


FIGURA 2. 15 REPRESENTACIÓN GRÁFICA DE LA DIFERENCIA EN LA CANTIDAD DE CAPAS OCULTAS EN LAS REDES UTILIZADAS EN APRENDIZAJE PROFUNDO FRENTE A LAS UTILIZADAS POR UNA RED NEURONAL CLÁSICA EN EL APRENDIZAJE MAQUINAL.

INFORMÁTICA MOLECULAR

CAPÍTULO 3

En el presente capítulo se introducirán los conceptos básicos referidos a la Informática Molecular, con un especial acento en discutir aspectos centrales referidos a la representación computacional de moléculas y cuestiones vinculadas a los desafíos algorítmicos que requieren ser abordados cuando la inferencia de modelos QSAR/QSPR se realiza mediante técnicas de Aprendizaje Maquinal. Asimismo, se detallarán algunos experimentos referidos al uso de Analítica Visual en el diseño de modelos QSAR/QSPR. Estas contribuciones son de alcance general para la Informática Molecular y han servido como etapa de entrenamiento para el desarrollo de esta tesis.

3.1. CONCEPTOS DE INFORMÁTICA MOLECULAR

Hoy en día estamos inmersos en la era de los datos. Ya desde el comienzo de este milenio era fácilmente posible acceder a publicaciones de aproximadamente 50 millones de sustancias químicas, 6 millones de reactivos, 7 millones de reacciones químicas, 16000 estructuras de proteínas por cristalografía de rayos X y 250000 estructuras de moléculas pequeñas [Glen & Aldridge, 2002]. Pero recién en los últimos años, esta recopilación masiva de datos fue demandando nuevas tecnologías de ciencias de los datos para analizarlos de manera sistemática. La Informática Molecular (*Molecular Informatics*) es un campo multidisciplinario que, mediante el uso de diferentes Tecnologías de la Información y la Computación, examina datos químicos variados con el objetivo de extraer información a partir de ellos.

La Informática se piensa, desde su origen, que uno de los aspectos sobre los cuales puede tener mayor impacto es la construcción de modelos matemáticos para predecir las interacciones de las moléculas [Baumann *et al.*, 2011]. Cuando la Informática y el Modelado Molecular comenzaron a desarrollar un papel central en las investigaciones químicas y biológicas, surgieron disciplinas como la Bioinformática (*Bioinformatics*) o la Quimioinformática (*Cheminformatics*), y posteriormente el concepto de Informática Molecular comenzó a definirse. En términos generales, la Informática Molecular es la ciencia que explora datos químicos y/o biológicos tanto a nivel molecular como sistémico [Baumann *et al.*, 2011]. Uno de sus grandes desafíos es, además de administrar, usar de manera más

creativa las grandes bases de datos de compuestos conjuntamente con la información asociada y sus estructuras [Glen & Aldridge, 2002].

El campo de acción de la Informática Molecular es la recopilación, la transformación y la visualización de datos químicos y/o biológicos para extraer un conocimiento más profundo de las propiedades subyacentes en ellos, por medio de la construcción de modelos basados en datos. Por lo tanto, el Aprendizaje Maquinal desempeña un papel central en la Informática Molecular. Este campo tuvo su origen en el diseño y el descubrimiento de drogas donde las bases de datos de moléculas pequeñas o dianas farmacológicas se analizan mediante diversas técnicas y enfoques tanto computacionales como matemáticos y estadísticos. El objetivo de esto es descubrir patrones útiles capaces de construir modelos matemáticos para predecir el comportamiento o respuesta biológica de nuevas moléculas, dando lugar al estudio de la Relación Cuantitativa Estructura-Actividad [Baumann *et al.*, 2011] (Figura 3.1). Existe una fuerte correlación entre la Informática Molecular y el modelado de la Relación Cuantitativa Estructura-Actividad/Propiedad. Este modelado es generalmente nombrado por sus siglas en inglés QSAR/QSPR que corresponden a *Quantitative Structure Activity/Property Relationship*. Aunque se originaron casi conjuntamente, la Informática Molecular abarca al Modelado QSAR/QSPR.

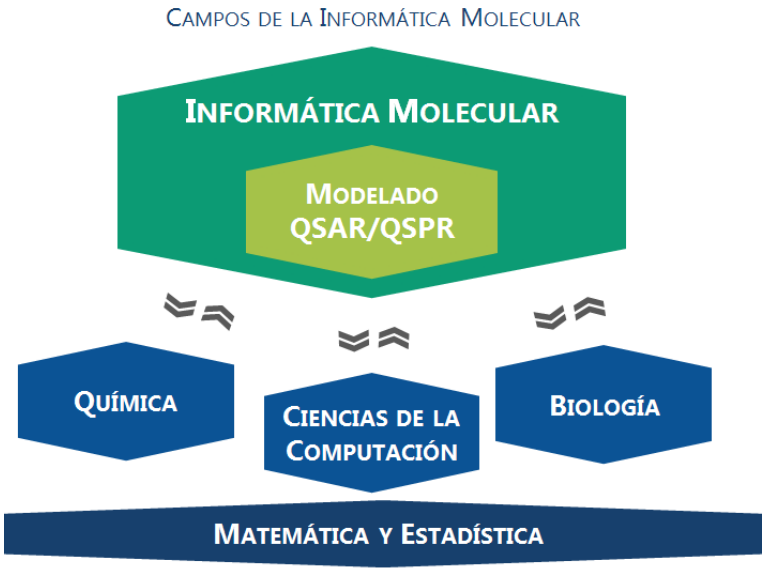


FIGURA 3. 1. DENTRO DEL CAMPO DE LA INFORMÁTICA MOLECULAR, SE ENCUENTRA EL MODELADO QSAR O QSPR, EN ESTRECHA RELACIÓN CON LOS CAMPOS DE LA QUÍMICA, LA BIOLOGÍA Y LAS CIENCIAS DE LA COMPUTACIÓN, SOPORTADOS POR LA MATEMÁTICA Y LA ESTADÍSTICA.

Entre los conceptos más utilizados en Informática Molecular y Quimioinformática se encuentra el de similaridad estructural o similitud molecular,

sobre todo para el diseño o descubrimiento de fármacos asistido por computadoras (CADD por sus siglas en inglés para *Computer-Aided Drug Design*) [Bender & Glen, 2004]. La similaridad estructural hace referencia a la semejanza en cuanto a características estructurales o funcionales entre dos o más moléculas en relación al efecto que tengan sobre elementos reactivos [Johnson & Maggiora, 1990]. Este concepto desempeña un papel importante en el diseño de productos químicos con un conjunto predefinido de propiedades, permitiendo la predicción de estas a través de la optimización de compuestos mediante estructuras moleculares alternativas que mantienen las propiedades requeridas mientras mejoran u optimizan, por ejemplo, los perfiles farmacocinéticos.

Los medicamentos, que existen desde las más antiguas civilizaciones, son uno de los ejemplos de varios de los aspectos de la vida cotidiana que han mejorado, basándose en el aumento sobre la comprensión de las propiedades de los átomos y las moléculas [Rosi, 2010]. Para diseñar productos químicos que se adapten a las demandas del mercado actual es necesario contar con grandes bases de datos que contengan la representación computacional de moléculas, para la identificación de patrones entre las estructuras químicas y las propiedades deseadas [Johnson & Maggiora, 1990; Bender & Glen, 2004]. El Cribado Virtual (*Virtual Screening*) es una de las técnicas empleadas para esta tarea. Se refiere al filtrado computacional, o *in silico,* capaz de seleccionar potenciales moléculas candidatas para cumplir con cierta propiedad objetivo o *target*, con el fin de reducir el tamaño del espacio químico (variaciones teóricas estructurales) con el que se realizaran las etapas más costosas del proceso (síntesis, pruebas toxicológicas, etc.) [Walters *et al.*, 1998; Rester, 2008].

Un punto fundamental en esta problemática es lograr representar computacionalmente a las moléculas en forma efectiva para dilucidar posibles patrones existentes en las bases de datos que las contienen. Mediante la representación computacional de moléculas es posible la manipulación de las mismas para generar, por un lado, propuestas moleculares diferentes, y por otro, probar su comportamiento en diferentes escenarios, generando alternativas y mejoras que pueden traducirse en aumento del confort de vida (por ejemplo: nuevos medicamentos con menos efectos secundarios o materiales con propiedades específicas). Diferentes propuestas de representación computacional de moléculas se han ensayado a lo largo de la historia, hasta llegar al desarrollo de los paquetes actuales de análisis y visualización de las mismas para su caracterización a través de Descriptores Moleculares (DMs).

3.2. REPRESENTACIÓN COMPUTACIONAL DE MOLÉCULAS

Históricamente, la mayoría de los químicos han modelado la estructura de las moléculas utilizando una representación altamente idealizada, donde los átomos se representan como vértices y enlaces [Adams, 2010]. El concepto moderno de moléculas surge de los filósofos griegos quienes entendían que el universo está compuesto por átomos y huecos. Cuando se analizan estas estructuras hay tres aspectos fundamentales a considerar: cómo se conectan los átomos, la distancia interatómica y la distribución espacial [Rosi, 2010]. En tal sentido, desde hace siglos, los científicos trabajan en la representación de las moléculas, el progreso de esto va desde la creación de modelos mecánicos hasta los modelos moleculares computacionales actuales que resolvieron todos los tres aspectos antes nombrados (conexión, distancia y distribución).

3.2.1. RESEÑA HISTÓRICA

En 1865, August Wilhelm von Hofmann presentó por primera vez (Figura 3.2 A) en las conferencias públicas de química que impartía, un modelo mecánico para representar moléculas conocido como "modelo de barras y esferas" (*ball and stick model*). Von Hofmann empleó barras de acero que funcionaban como enlaces para unir los átomos representados por pelotas de croquet [Rosi, 2010]. Tal fue la potencia didáctica de esta representación que surgieron varias empresas dedicadas a la fabricación y comercialización de kits para armar modelos moleculares (Figura 3.2 B), los cuales aún se usan en varios colegios y universidades.

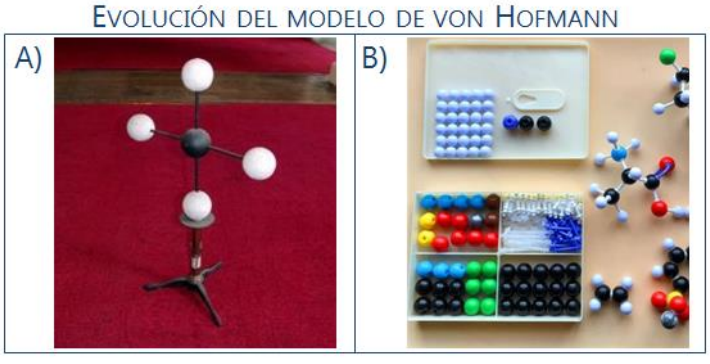


FIGURA 3.2 A) FOTO HISTÓRICA DEL MODELO DE VON HOFMANN. **B)** EJEMPLO DE KIT ACTUAL DE ARMADO DE MODELOS MOLECULARES.

Las principales deficiencias del modelo de von Hofmann fueron básicamente geométricas, sobre todo, la proporción de tamaños de los átomos: el hidrógeno era más grande que el carbono, lo cual puede verse en la Figura 3.2 A. En 1910, el físico teórico y termodinámico, Johannes Diderik van der Waals recibió el Premio Nobel

de Física por su estudio de los gases a través de la popularmente conocida "ecuación de estado de Van der Waals". En ese trabajo describió las consecuencias físicas de las dimensiones atómicas, es decir, cómo afecta el tamaño de los átomos y la distancia que existe entre los átomos que forman una molécula [Tabor & Winterton, 1969]. Esto permitió corregir los trabajos de von Hofmann.

En 1916 en su artículo *The atom and the molecule*, Gilbert Newton Lewis presentó "La estructura de Lewis", una representación gráfica que muestra los pares de electrones de enlaces entre los átomos de una molécula y los pares de electrones solitarios que pueden existir (Figura 3.3). Son representaciones adecuadas y sencillas de iones y compuestos, que facilitan el recuento exacto de electrones. Permiten mostrar los enlaces químicos que existen dentro de la molécula, indicando la posición de los átomos en el espacio. Las estructuras de Lewis muestran los diferentes átomos usando su símbolo químico, líneas que se trazan entre ellos indicando qué tipo de enlace los une y los electrones que no participan en los enlaces como puntos alrededor de los átomos a los que pertenecen [Rosi, 2010].

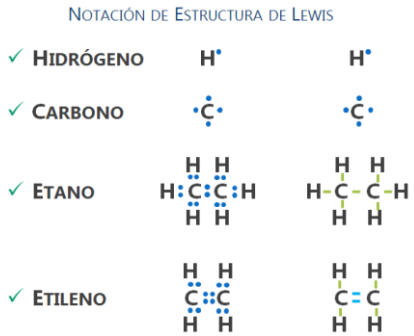


FIGURA 3.3. EJEMPLOS DE LA NOTACIÓN DE LEWIS PARA MOLÉCULAS SIMPLES

Desde mitad del siglo XX existe el concepto de "campo de fuerza" para la espectroscopía vibracional que considera las fuerzas que actúan entre pares de átomos en una molécula. Sin embargo, este concepto no cobró relevancia entre la comunidad física y química hasta 1946 cuando se generó por primera vez un enfoque de modelo molecular de manera cuantitativa (antes sólo se trataba de modelos de mecánica newtoniana para el alargamiento de enlaces, ajuste de ángulos, etc.) [Aines *et al.*, 2006]. En 1950 Derek Harold Richard Barton publicó un estudio sobre cómo la conformación de los esteroides afecta sus propiedades químicas, revelando la importancia de contar con estructuras tridimensionales para el estudio de la estructura, la estabilidad, la conformación y la reactividad de las moléculas [Barton, 1950].

En 1953, James Dewey Watson y Francis Harry Compton Crick, basándose en el trabajo de varios colegas, entre ellos Erwin Chargaff, Maurice Hugh Frederick Wilkins y Rosalind Elsie Franklin, presentaron la estructura del ADN (ácido desoxirribonucleico), constituyendo uno de los hitos en el descubrimiento científico posicionando a la estructura de la doble hélice del ADN como el modelo molecular más famoso (Figura 3.4). El descubrimiento de la estructura del ADN no fue casual. Durante décadas se trabajó en dilucidar la estructura de esta molécula. El ADN se aisló por primera vez en 1869, en 1930 se demostró cuáles componentes lo conforman y en 1944 se comprobó recién cuál era su función biológica. Conocer su estructura era importante porque se suponía que permitiría los grandes avances que hoy existen en Biología Molecular, Genética y Medicina [Aines *et al.*, 2006].

FIGURA 3.4. PRIMERA FOTO DE JAMES WHATSON Y FRANCIS CRICK CON EL MODELO DE LA ESTRUCTURA DE DOBLE HÉLICE DE ADN EN MAYO DE 1953.

En 1961, James Hendrickson demostró que las computadoras podían usarse para calcular la energía molecular y en 1966 Cyrus Levinthal expandió su uso a la Química cuando publicó un estudio que detallaba el empleo de gráficos moleculares y simulación por computadora para estudiar las estructuras de las proteínas y los ácidos nucleicos. La compañía *MDL Information Systems Inc.* fue la primera proveedora de productos informáticos para almacenar y recuperar moléculas como estructuras gráficas y para administrar bases de datos de reacciones químicas y otros datos relacionados. El primer programa fue presentado en 1979 y se llamó MACCS (*Molecular ACCess System*). Estos sistemas revolucionaron la forma en que los científicos accedieron y manejaron la información química desde los años ochenta [Richon, 2008].

Junto con otras compañías, *MDL* continuó a la vanguardia del campo ya conocido por entonces como Quimioinformática [Richon, 2008]. La Quimioinformática ha adoptado con éxito la representación con vértices y enlaces para moléculas. Muchas de sus técnicas se basan en lo que se conoce como "tabla de conexión", es decir, una lista de todos los átomos y enlaces que aparecen en la

molécula [Adams, 2010]. Existen diferentes tipos de archivos en los que se codifica la estructura de una molécula, muchos de los cuales fueron creados *ad hoc* por las compañías de software que necesitaban utilizarlos en sus herramientas, y otros fueron desarrollos académicos en respuesta a distintas demandas de la comunidad científica.

3.2.2. TIPOS DE FORMATOS DE ARCHIVOS

Un formato de archivo es un estándar que define la forma en que la información se codifica en un archivo informático. Cada formato de archivo tiene características diferentes que le permite almacenar determinados tipos de datos conforme a las necesidades que se tengan. Los diferentes programas informáticos que existen permiten crear moléculas que nunca antes han sido sintetizadas o descriptas, también dibujar moléculas en dos dimensiones para luego generar un modelo en tres dimensiones. Además, se pueden hacer mediciones entre los átomos del modelo o entender las moléculas a un nivel estructural superior representándolas con un detalle menor en cuanto al número de átomos, por ejemplo, mediante láminas betas, entre otras funciones. Estos programas trabajan con ciertas extensiones de archivos particulares. En las próximas subsecciones, se describirán los dos formatos más conocidos para representar estructuras químicas [Faulon & Bender, 2010].

3.2.2.a. BASADOS EN TABLA DE CONEXIONES

Los archivos basados en *tablas de conexiones* existen desde el algoritmo de Morgan propuesto en 1965. Actualmente han mejorado e incluido otro tipo de información adicional relevante [Horvath, 1992]. Un archivo de *tabla de conexión* contiene una lista de todos los átomos sin estructura, junto con la información de los enlaces que describen exactamente cómo se conectan cada uno de los átomos [Wang, 2006]. Es decir, es una lista o matriz con información sobre los átomos (nodos) y los enlaces (aristas) entre los átomos que forman una molécula. La principal ventaja de este tipo de representación es la simplicidad y una de sus desventajas, el incremento del tamaño de la matriz a medida que aumenta el número de átomos.

Tal vez la extensión más conocida de este tipo de archivos sea *MDL Molfile* (*.mol). El archivo MOL es un archivo de texto plano que contiene hasta tres líneas en el encabezado como información adicional, una lista de los átomos, una lista con los enlaces y coordenadas tridimensionales (marco de referencia espacial) e información de conectividad (Figura 3.5). Otra extensión ampliamente utilizada es

HIN (*.hin) del archivo *HyperChem Chemical Modeller Input*. El archivo HIN brinda información similar a los archivos MOL, se usa para almacenar uno o más bloques moleculares o atómicos, contiene información adicional y almacena información sobre el sistema de coordenadas del visor, por lo que al abrir el archivo se restaura la vista anterior que se tuvo del sistema molecular (Figura 3.6).

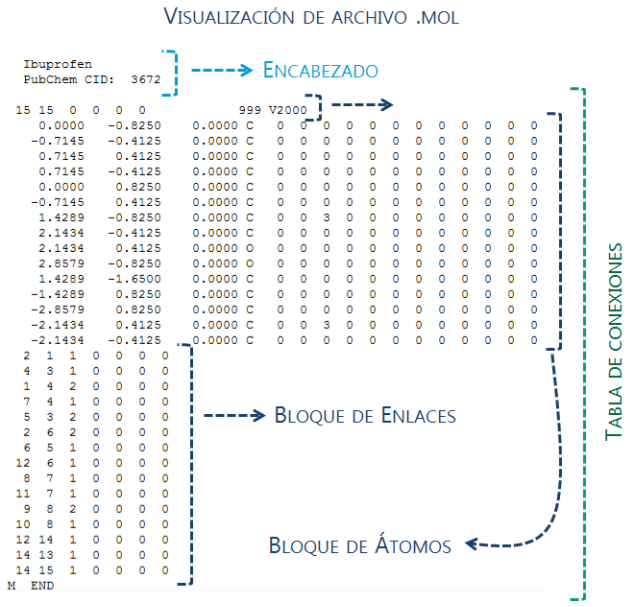


FIGURA 3.5. INFORMACION ESTRUCTURAL CONTENIDA EN UN ARCHIVO DE FORMATO MOL PARA EL IBUPROFENO.

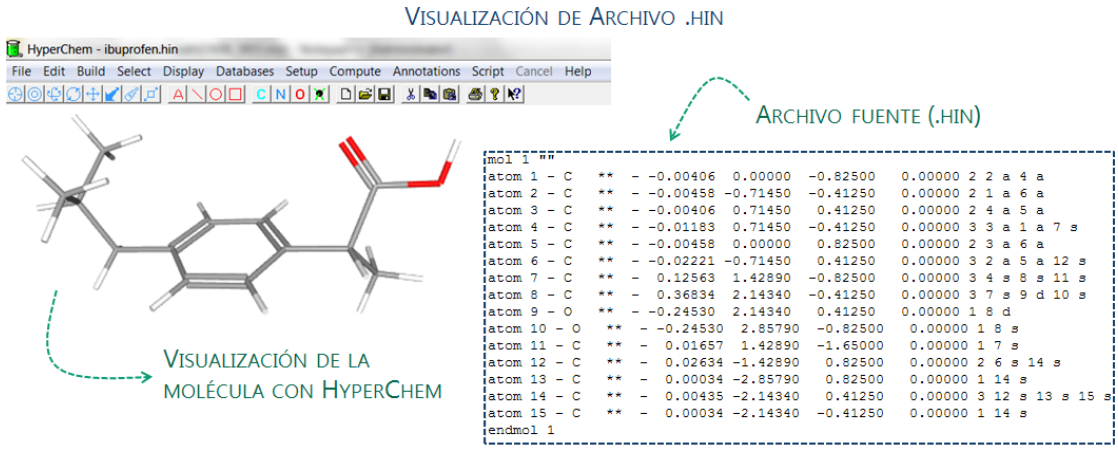


FIGURA 3.6. INFORMACION ESTRUCTURAL CONTENIDA EN UN ARCHIVO DE FORMATO HIN PARA EL IBUPROFENO.

3.2.2.b. BASADOS EN TEXTO

Los archivos basados en texto, también conocidos como notación de línea (*line notation*), representan una molécula como una cadena de caracteres de una sola línea. Son particularmente útiles para las aplicaciones de base de datos relacionales, ya que las notaciones de las moléculas pueden incluirse como "texto" haciendo

mucho más fácil, por ejemplo, realizar búsquedas que con las tablas de conexión que deben almacenarse como "BLOBs" (*Binary Large Objects*). La Notación de Línea Wisswesser (WLN), desarrollada en la década de 1950, fue la primera capaz de representar moléculas complejas de manera correcta y compacta. La complejidad de WLN para adoptar la forma canónica (forma "correcta") evitó su adopción generalizada, ya que la codificación de las reglas de canonización resultaba computacionalmente intratable [Craig, 2009].

Actualmente, el más conocido de estos archivos es el SMILES o código SMILES (*Simplified Molecular Input Line Entry Specification*), creado a finales de los 80 por David Weininger, como una versión más simple y más accesible para los humanos que WLN [Weininger, 1988]. SMILES está ideado para que las estructuras químicas sean fácilmente legibles. La forma en que hoy se almacenan grandes cantidades de datos y se realizan búsquedas más flexibles y rápidas, se basa en el trabajo pionero de Weininger que revolucionó el campo de la Quimioinformática. Al igual que WLN, SMILES tiene una forma canónica. Convertir un SMILES no canónico a uno canónico es función del programa informático y no del químico (humano). Esto simplificó el proceso y resultó clave para la popularización de su uso [Craig, 2009].

Hace algunos años el estándar InChI fue propuesto para la representación de estructuras moleculares por la IUPAC (*International Union of Pure and Applied Chemistry*). Sin embargo, SMILES no solo es más comprensible a la lectura humana, sino que el soporte a nivel de software es mayor, por lo que continúa usándoselo mayoritariamente. En la Figura 3.7 se puede observar la representación bidimensional de la molécula de ibuprofeno sin átomos de hidrógeno y debajo su código SMILES. En este código, cada átomo se representa con el símbolo que le corresponde en la tabla periódica, se asume que por defecto se completa el octeto de cada átomo con la cantidad necesaria de hidrógenos por lo que estos no forman parte del código SMILES. Se indica un doble enlace con el signo igual (=) y un enlace triple con el signo numeral (#). La quiralidad de un elemento se señala con el signo arroba (@) [Rosi, 2010].

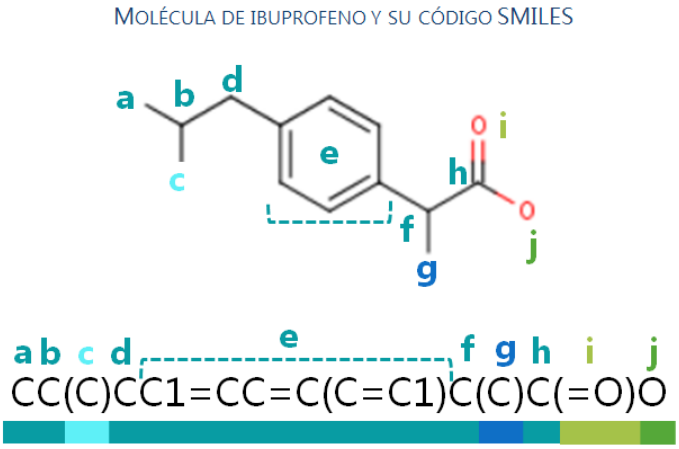


FIGURA 3.7 ESTRUCTURA BIDIMENSIONAL DEL IBUPROFENO CON HIDROGENOS Y SU RESPECTIVO CODIGO SMILES CANONICO

En el caso de moléculas que presentan anillos, es necesario *cortar* el anillo para obtener cadenas lineales ramificadas en la cadena principal que, se asume, se unirán por enlaces simples a ella. Cada rama, que aparece unida a la cadena principal, se representa dentro de un par de paréntesis, es decir, como cadenas anidadas [Rosi, 2010]. En la Figura 3.8 se presenta la molécula de estireno (izquierda) y la Unidad Repetitiva Estructural (URE) del poliestireno (derecha), con las cadenas coloreadas para hacerlas identificables con respecto a su código SMILES. Este ejemplo tiene la particularidad de presentar el símbolo asterisco (*) en su código SMILES que puede utilizarse como un carácter especial capaz de indicar la cabeza (*head*) y la cola (*tail*) de una URE. Identificar estas zonas es deseable cuando se trabaja en Informática de Polímeros, ya que permite realizar de manera simple, por ejemplo, una unión del tipo cabeza-cola.

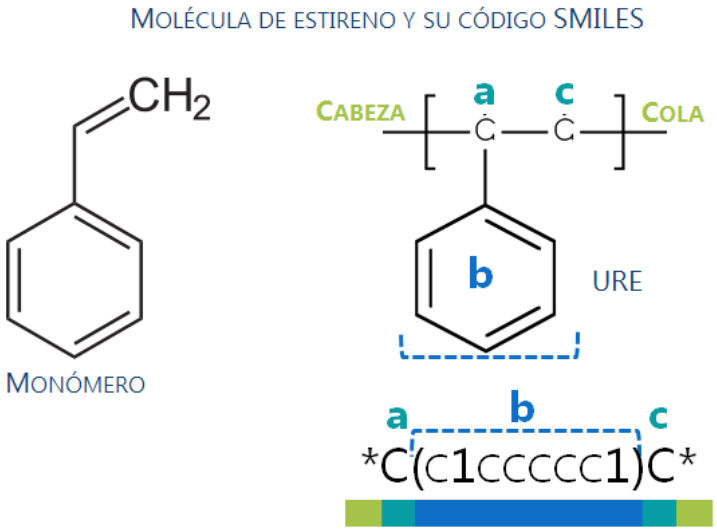


FIGURA 3.8 MOLECULA DE ESTIRENO MONÓMERO (IZQUIERDA) Y CORRELACION ENTRE LA URE DEL POLIESTIRENO Y SU CODIGO SMILES (DERECHA) IDENTIFICANDO CABEZA Y COLA (*).

3.3. CARACTERIZACIÓN DE MOLÉCULAS MEDIANTE DESCRIPTORES MOLECULARES

Se sabe que la estructura de una molécula se correlaciona con sus propiedades físico-químicas. La representación computacional de la estructura de las moléculas permite caracterizarlas a través del cálculo de sus DMs [Adams, 2010]. Estos descriptores captan características moleculares específicas y afectan directamente, en términos de resultados, rendimiento y aplicabilidad, a los modelos que relacionan la estructura de una molécula con una propiedad específica [Grisoni *et al.*, 2018]. Se utilizan para modelar propiedades de diversas disciplinas desde la Química Analítica, la Química Física y la Química Ambiental, hasta la Farmacéutica o la Medicina [Todeschini *et al.*, 2009].

Los DMs codifican una amplia variedad de información molecular y se han convertido en el soporte de muchas aplicaciones quimioinformáticas y bioinformáticas [Grisoni *et al.*, 2018]. Una de las razones del amplio uso de los DMs se debe a las propiedades de invariancia de los mismos, es decir su cálculo se basa en modelos determinísticos. Estas propiedades hacen referencia a la capacidad del algoritmo para proporcionar siempre el mismo valor de descriptor para la misma molécula, independientemente de las características particulares de la representación molecular, como por ejemplo la numeración o etiquetado de átomos, el marco de referencia espacial o las conformaciones moleculares [Kier & Hall, 1999; Block, 2000; Karelson, 2000].

3.3.1. DESCRIPTORES MOLECULARES

Los DMs son índices numéricos que codifican alguna información relacionada con la estructura molecular. La representación numérica de las estructuras moleculares que codifican la información química responsable de una propiedad molecular determinada (o actividad), se lleva a cabo mediante el cálculo de dichos descriptores, los cuales pueden realizarse en muchas formas, desde conteos de átomos simples hasta la cuantificación de características moleculares complejas [Kunal *et al.*, 2015]. Pueden ser propiedades fisicoquímicas experimentales de las moléculas o índices teóricos calculados mediante fórmulas matemáticas o algoritmos. Según Roberto Todeschini y Viviana Consonni *"los descriptores Moleculares son el resultado final de un procedimiento lógico-matemático que transforma la información química codificada dentro de una representación de una molécula en un número útil o el resultado de algún experimento estandarizado"* [Todeschini & Consonni, 2008].

3.3.1.a. DESCRIPTORES MOLECULARES CLÁSICOS

Los Descriptores Moleculares Clásicos (DMs Clásicos) se pueden clasificar de múltiples maneras. Diferentes tipos de descriptores codifican información química diferente. Usualmente, se dividen en dos categorías generales: mediciones experimentales, que incluyen, por ejemplo, Polarizabilidad, y en descriptores moleculares teóricos. Estos últimos derivan de la representación de la molécula y se clasifican adicionalmente de acuerdo al tipo de representación molecular. Por lo general, en la comunidad de la Informática Molecular abocada al estudio QSAR/QSPR se prefiere clasificarlos de forma que se vean reflejados los diversos niveles de representación de la estructura química (Figura 3.9): a) 0D, b) 1D, c) 2D, d) 3D. Además, existen algunos otros menos comunes como 4D, 5D, 6D y hasta 7D que tienen en cuenta la dependencia del tiempo, la dinámica de las moléculas, etc [Cherkasov *et al.*, 2014; Kunal *et al.*, 2015]. A continuación, se presenta información más detallada sobre los descriptores moleculares clasificados por su dimensión según varios autores [Xue & Bajorath, 2000; Consonni & Todeschini, 2001, 2010; Consonni *et al.*, 2002a, 2002b; García-Domenech *et al.*, 2008; Todeschini & Consonni, 2008, 2009; Todeschini *et al.*, 2009; O'Boyle *et al.*, 2011; Devinyak *et al.*, 2014; Kunal *et al.*, 2015].

a) Descriptores 0D: no necesitan información sobre la estructura molecular y las conectividades atómicas, no requieren optimización de la estructura molecular y son independientes de cualquier problema de conformación. Son descriptores de conteo de átomos y enlaces, que pueden interpretarse de forma natural. Generalmente, su contenido de información es bajo y muestran una degeneración muy alta, es decir, tienen valores iguales para varias moléculas, como los isómeros. Sin embargo, pueden jugar un papel importante en el modelado de varias propiedades físico-químicas o participar en modelos más complejos.

b) Descriptores 1D: representan listas de fragmentos estructurales (*fingerprints*). Los *fingerprints* son una forma de codificar la estructura de una molécula, como una serie de dígitos binarios (bits) que representan la presencia o ausencia de subestructuras particulares en la molécula. La comparación de *fingerprints* permite, por ejemplo, encontrar coincidencias con una subestructura y puede complejizarse tanto hasta necesitar miles de posiciones de bits.

c) Descriptores 2D: incluyen parámetros topológicos o estructurales. La representación de la estructura de la molécula depende de su topología, que indica la posición de los átomos individuales y las conexiones unidas entre ellos. Estos descriptores moleculares se calculan en función de la representación gráfica de las moléculas. Con el fortalecimiento de la Teoría de Grafos estos descriptores

moleculares topológicos fueron pioneros en la representación de estructuras moleculares en términos de descriptores cuantitativos.

d) Descriptores 3D: incluyen parámetros geométricos y la mayoría requiere un marco de referencia espacial. Entre los más conocidos se encuentran: 3D-MoRSE, WHIM, GETAWAY.

- Los descriptores 3D-MoRSE varían cuando se usan diferentes geometrías de inicio y aunque brindan información sobre la estructura de la molécula completa, se derivan principalmente de pares atómicos de corta distancia.
- Los descriptores WHIM (*Weighted Holistic Invariant Molecular*) son descriptores geométricos basados en índices estadísticos calculados sobre las proyecciones de los átomos a lo largo de los ejes principales. Capturan información 3D molecular relevante relacionada con el tamaño molecular, la forma, la simetría y la distribución de átomos con respecto a los marcos de referencia invariantes (marco de referencia único).
- Los descriptores GETAWAY (*GEometry, Topology, and Atom Weights AssemblY*) intentan hacer coincidir la geometría molecular 3D, proporcionada por las coordenadas atómicas centradas de Matriz de Influencia Molecular (MIM) y la relación con el átomo por topología, con información química utilizando diferentes esquemas de ponderación atómica.

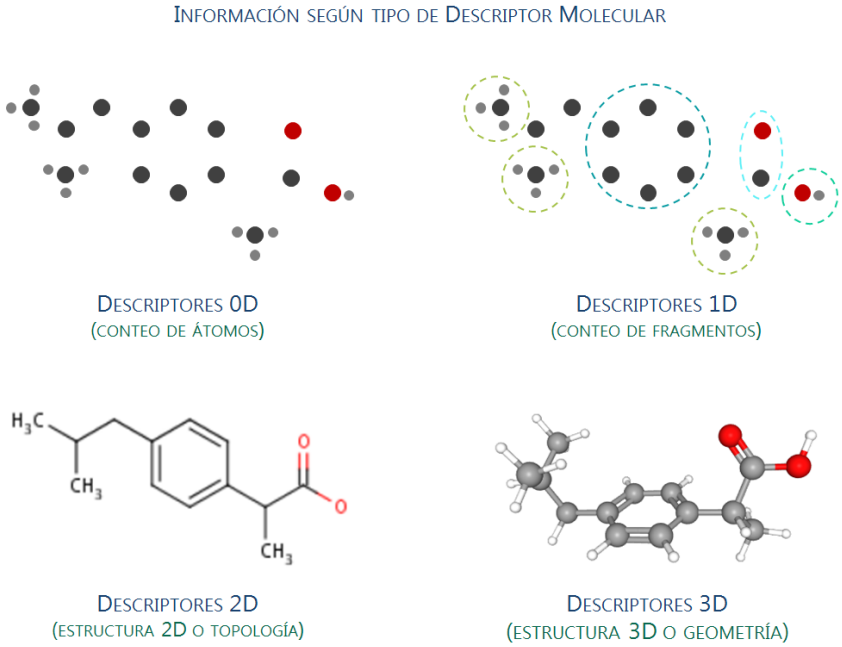


FIGURA 3.9 RESUMEN GRAFICO DE LA INFORMACION QUE BRINDA CADA TIPO DE DESCRIPTOR, TOMANDO COMO EJEMPLO LA MOLÉCULA DE IBUPROFENO.

3.3.1.b. DESCRIPTORES MOLECULARES DE VISIÓN MACRO

Aunque la gran cantidad de descriptores que se han desarrollados hasta inicios del milenio debieron haber sido suficientes para el estudio QSAR en el diseño de todo tipo de nuevos fármacos, esto no sucedió. La falta de ajuste en los modelos QSAR puede deberse a múltiples cuestiones, sin embargo, si una de ellas es la falta de descriptores adecuados es necesario que la comunidad invierta esfuerzo en generar nuevos descriptores capaces de describir fenómenos que los existentes no pueden [Carrasco-Velar, 2003]. Con las sofisticadas herramientas computacionales disponibles y el avance de la representación estructural y la explotación de estructuras químicas, es inevitable la proliferación de nuevos descriptores moleculares, aunque la calidad de los mismos no sea siempre la óptima [Kunal *et al.*, 2015; Sahoo *et al.*, 2016].

Son varias las disciplinas que se han sumado al uso de DMs para basar sus investigaciones desde los inicios del modelado QSAR en Quimioinformática. Una de ellas es el Diseño de Materiales Asistido por Computadoras en el que varios autores han propuesto nuevos descriptores según el tipo de materiales con los que trabajan o las propiedades que estudian. Para Informática de Polímeros existen varios [Cao & Lin, 2003; Palomba *et al.*, 2012b; Wu *et al.*, 2016], ya que admite la generación de *Descriptores Moleculares de Visión Macro* (DMs de visión Macro) porque sus moléculas son más grandes y más complejas debido a su formación de tipo cadena. En particular en esta tesis interesan los descriptores propuestos por Palomba [Palomba *et al.*, 2012b], los cuales se describen brevemente a continuación, para mayor detalle puede consultarse su trabajo de tesis doctoral [Palomba, 2014].

Palomba propone descriptores moleculares de Visión Macro especialmente ideados para tratar el problema de baja caracterización que podría tener la representación de un material polimérico a partir de su monómero[1]. Al respecto, Palomba formula descriptores *ad hoc* que capturan información de la estructura del polímero basándose en la unidad central del trímero[2], centrando la atención en dos partes de la cadena: la principal y la lateral (Figura 3.10). El ejemplo presentado corresponde a *Poly{hydroquinone-alt-[bis(4-fluorophenyl)methylphosphine oxide]}*, que en la base de datos de polímeros que se presentará en el siguiente capítulo corresponde al ID 35. Estos descriptores de Visión Macro se dividen en dos tipos: los simples y los normalizados. Estos últimos no varían dependiendo del modelo

[1] Polímero: muchas (poli) partes (mero). Monómero: una (mono) mero (parte).
[2] Trímero: es la unión de tres monómeros o tres unidades repetitivas estructurales.

molecular o representación utilizada, porque se normalizan con el número de átomos de la porción del polímero considerada para su cálculo. Además, en la categoría DMs de visión Macro también se incluyen los parámetros del ensayo bajo los cuales se realizó el experimento que midió las propiedades a predecir, lo que se conoce como procesamiento del material e historia [Palomba, 2014].

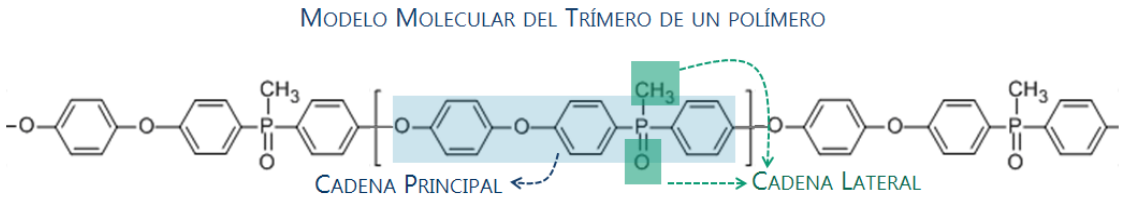


FIGURA 3.10 MODELO MOLECULAR DEL TRÍMERO DE UN MATERIAL POLIMÉRICO Y LA IDENTIFICACIÓN DE LOS FRAGMENTOS QUE CORRESPONDEN A LA CADENA PRINCIPAL Y LATERAL DE LA URE CENTRAL.

3.3.2. CÁLCULO DE DESCRIPTORES MOLECULARES

Existen diferentes herramientas libres y comerciales para el cálculo de múltiples descriptores moleculares, otros descriptores simples pueden ser calculados a través de scripts sencillos. Entre los software de generación o cálculo de descriptores moleculares más conocidos y más usados en el modelado QSAR/QSPR abocado al diseño de fármacos se encuentran: Dragon (3214 descriptores) [Dragon, 2007]; PaDEL (15387 descriptores) [Yap, 2011] y CDK [Guha, 2007]. A principios de 2018 se publicó Mordred, una aplicación capaz de calcular más de 1800 descriptores para moléculas grandes, lo que no siempre puede lograrse con otros software [Moriwaki *et al.*, 2018]. Sin embargo, al momento de la escritura de esta tesis, no fue posible realizar el cálculo con Mordred para nuestra base de datos de polímeros, aunque sí se pudo con la librería CDK calcular 57 descriptores para moléculas con peso promedio Mw de hasta 2.2 $x10^5$ [g/mol], lo que se verá en detalle en el capítulo 4. Cabe destacar que no existen programas especialmente ideados para el cálculo de descriptores moleculares para materiales poliméricos, es decir, que calculen descriptores específicos de polímeros, o que puedan manejar las enormes moléculas que representan a estos materiales.

3.4. MODELADO QSAR/QSPR

El Modelado QSAR (*Quantitative Structure-Activity Relationship*) relaciona de manera cuantitativa parámetros específicos de la estructura de la molécula (a través de descriptores moleculares) con la *actividad objetivo* (generalmente biológica) que se estudia. El estudio QSAR surgió para entender cómo las velocidades de reacción diferencial de las reacciones químicas dependen de las diferencias en la estructura

molecular [Livingstone, 2000]. Con el tiempo han surgido variantes al modelado QSAR, la que nos interesa en esta tesis en particular es el modelado QSPR (*Quantitative Structure-Property Relationship*). El Modelado QSPR relaciona los descriptores moleculares, de manera cuantitativa, con *una propiedad objetivo o target*. Los modelos basados en relaciones cuantitativas estructura-actividad pueden describirse como la aplicación de métodos estadísticos al problema de encontrar relaciones empíricas del tipo $P_i = k'(d_1, d_2, \dots, d_n)$, donde P_i es la propiedad de interés, k' es una transformación matemática (típicamente lineal) y los d_i representan el cálculo o la medición de las propiedades estructurales [Hansch & Fujita, 1964]. En términos matemáticos, un modelo QSAR es una función $P = f(D)$, donde $D = (d_1, d_2, \dots, d_n)$ es un compuesto químico representado como un vector de descriptores d_i, y P es una propiedad experimental de D [Yousefinejad & Hemmateenejad, 2015].

Una explicación *grosso modo* válida para modelado QSAR/QSPR mediante aprendizaje maquinal consiste en una base de datos de compuestos químicos, a los que se les calculó una serie de descriptores y para los cuales, se tiene información experimental de una propiedad fisicoquímica o actividad biológica de interés. A partir de esta base de datos (conjunto de entrenamiento), se construye la función f. Una vez obtenida f se la puede aplicar a compuestos sobre los cuales no se tiene información experimental de la propiedad a modelar [Yousefinejad & Hemmateenejad, 2015]. La relación cuantitativa no se da directamente entre la estructura del compuesto y la actividad/propiedad, sino que la estructura se traduce en descriptores moleculares que permiten un tratamiento matemático de la información fisicoquímica para la generación del modelo que explique la propiedad (ver Figura 3.11).

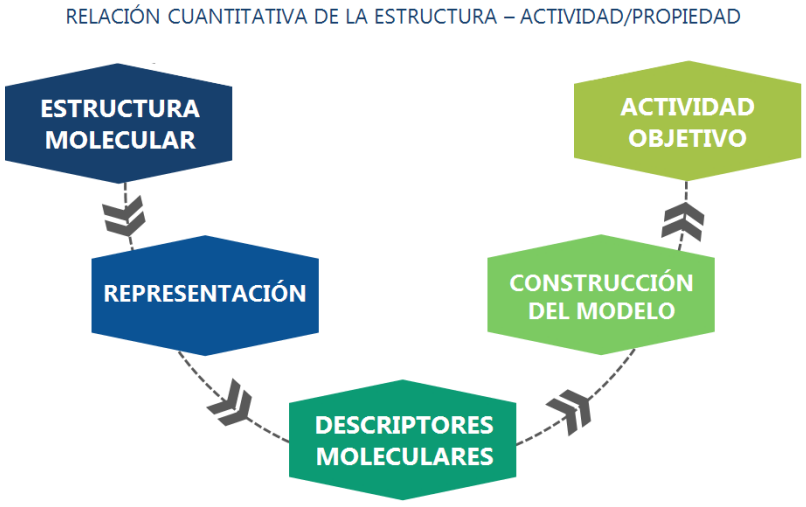


FIGURA 3.11 PASOS DEL MODELADO QSAR/QSPR.

El modelado cuantitativo de la relación estructura-actividad es una de las principales herramientas computacionales empleadas en Informática Molecular. Es ampliamente utilizado dentro de las áreas de investigación y regulación ambiental para el uso de alimentos, cosméticos y productos químicos industriales donde las regulaciones a menudo son limitadas, y los datos de propiedades y toxicidad son escasos o no están disponibles [Cherkasov *et al.*, 2014]. El interés en modelos predictivos capaces de dar estimaciones confiables efectivas aumentó en gran medida en los últimos años, ya que son herramientas consideradas cada vez más útiles y más seguras para predecir datos sobre productos químicos [Todeschini *et al.*, 2009].

Los métodos QSAR se utilizan en varias agencias gubernamentales, para la identificación, detección y priorización de peligros potenciales para la salud, [Hughes *et al.*, 2009; Arvidson *et al.*, 2010]. Existen reglamentos, de la Organización para la Cooperación y el Desarrollo Económicos (OCDE por las siglas en inglés para *Organisation for Economic Co-operation and Development*) para la validación de modelos QSAR con propósitos regulatorios. Son conocidos como "Principios de la OCDE para la validación, con fines regulatorios, de los modelos QSAR" [OECD Principles, 2013]. Entre estos principios destacamos la necesidad de contar con un algoritmo determinístico, con un dominio de aplicabilidad definido y también, si es posible, con una interpretación mecanicista.

Los modelos QSAR se aplican para evaluar impactos potenciales de químicos y materiales en sistemas ecológicos y de salud humana. El supuesto bajo el cual operan, los métodos de Quiminformática es: productos químicos similares tienen una actividad o propiedad similar. Unir el espacio químico y el biológico es la clave para el descubrimiento y desarrollo de nuevas moléculas. Mediante la búsqueda de similitudes en bibliotecas de compuestos con actividades conocidas es posible, idealmente, predecir la función biológica de un fármaco dada solo su estructura química. Además, es posible predecir los efectos ADMET (Absorción, Distribución, Metabolismo, Excreción y Toxicidad) sobre la salud humana y la toxicidad ambiental de los productos químicos o Compuestos Orgánicos Volátiles (COVs o *VOCs* por las siglas en inglés para *Volatic Organic Compounds*-) [Nettles *et al.*, 2006].

El proceso de modelado QSAR/QSPR, generalmente, además de propósitos predictivos también tiene propósitos explicativos [Todeschini *et al.*, 2009]. Los estudios QSAR/QSPR permiten una explicación del comportamiento de los productos químicos y esta posible explicación brinda la oportunidad de ajustar el comportamiento a uno deseado. Encontrar una relación f permite utilizarla para estimar la actividad de nuevos compuestos (que pueden aún no estar sintetizados),

es decir, predecir *in silico* el valor de una actividad o propiedad, a partir del análisis de información de otras moléculas. El modelado QSAR/QSPR es un método alternativo a la experimentación en laboratorio que no pretende reemplazarla, por el contrario, se nutre de ella, pero si permite la reducción de la selección inicial de moléculas en investigaciones costosas [Kunal *et al.*, 2015].

3.4.1. EVOLUCIÓN DEL MODELADO QSAR

En 1962 Corwin Herman Hansch, conocido como "'padre del diseño de moléculas asistidas por computadora", con su trabajo titulado *Correlation of Biological Activity of Phenoxyacetic Acids with Hammett Substituent Constants and Partition Coefficients* [Hansch *et al.*, 1962], fundó el campo QSAR y este ha evolucionado desde entonces. Inicialmente, su aplicación en el Diseño Racional de Drogas estuvo limitada a pequeñas series de compuestos congéneres y se utilizaban métodos de regresión relativamente simples. Gradualmente fue evolucionando hasta aplicarse una extensa variedad de métodos de Aprendizaje Maquinal para analizar conjuntos de datos masivos de estructuras moleculares con amplia diversidad. Esta evolución demando casi 60 años de avances y desarrollos interdisciplinarios de la comunidad científica para que el modelado QSAR/QSPR se convirtiera en uno de los métodos más empleados en el modelado de propiedades físicas y químicas [Cherkasov *et al.*, 2014].

Hansch y Unger utilizaron el análisis de conglomerados para agrupar posibles sustituyentes aromáticos [Hansch *et al.*, 1973]. Hansch y Fujita trabajaron con los coeficientes de partición octanol-agua (LogP) [Leo *et al.*, 1975]. La propiedad LogP es aditiva, y la contribución parcial de un sustituyente al LogP de una molécula, a menudo, es la misma que la contribución de ese sustituyente al LogP de otra molécula. Con el término π se denominó el sustituyente en la hidrofobicidad. Las tablas de valores π, generadas a partir de cuidadosas mediciones de la relación octanol-agua, fueron útiles para calcular el registro relativo de LogP de los miembros de una serie. Pero para comparar el valor óptimo de LogP se requería el de las moléculas de origen, entonces, Hansch y Leo usaron una base de datos de valores experimentales de LogP para cientos de diversas moléculas primarias para parametrizar un enfoque de fragmento aditivo para predecir LogP de forma automática [Cherkasov *et al.*, 2014].

A lo largo de los años, muchos otros métodos para calcular el LogP se han ideado. La relación parabólica entre la potencia y el registro P no se ajustó a todos los conjuntos de datos. Por lo tanto, Kubinyi sugirió una ecuación bilineal que permite diferentes pendientes en valores LogP altos y bajos [Kubinyi, 1977]. A

medida que los conjuntos de datos crecían y eran estructuralmente cada vez más diversos, los descriptores diseñados para ser aplicados dentro de un marco de mecanismo de reacción común ya no eran suficientes. La solución fue generar descriptores moleculares que pudieran capturar determinantes más generales de las variaciones de la actividad a lo largo y dentro de las series, basándose en el principio de que toda la estructura química de una molécula dicta sus propiedades fisicoquimicas [Cherkasov *et al.*, 2014].

En 1984 se desarrolló un método, por un lado, capaz de dividir una molécula en fragmentos 2D constituyentes y, por el otro, capaz de autogenerar estos fragmentos para grandes bases de datos de moléculas. Además, el método era capaz de correlacionar la frecuencia de cada uno de los fragmentos con una actividad biológica [Klopman, 1984]. Este trabajo constituyó un gran avance, ya que creó un método computacional eficiente para representar y correlacionar características estructurales fácilmente interpretables para una gran cantidad de productos químicos; dando origen a los modelos QSAR globales para la predicción de la actividad biológica [Cherkasov *et al.*, 2014]. Con el tiempo, y el progreso en el desarrollo de los métodos computacionales, el modelado QSAR/QSPR fue ganando terreno y se aplicó en gran diversidad de áreas.

3.4.2. ANALÍTICA VISUAL

El modelado QSAR/QSPR implica dar solución a varios problemas y uno de ellos es la selección de un conjunto relevante de descriptores para la actividad o propiedad a modelar. Los DMs pueden afectar la relación que se construya entre la estructura y la actividad/propiedad, por esto, la selección de los DMs es catalogada como el paso más importante en este tipo de modelado. La mayoría de los métodos de Selección de Características (*Feature Selection*) se enfocan en ajustar estadísticamente las relaciones entre los descriptores y la propiedad objetivo, es decir, sin tener en cuenta aspectos asociados con el conocimiento experto del área de la propiedad a modelar. Una estrategia para incorporar este conocimiento, en el proceso de selección de características y mejorar la confianza del usuario en los modelos, es la utilización de la Analítica Visual [Martínez *et al.*, 2014a; Martínez *et al.,* 2015].

El término Analítica Visual fue presentado por primera vez en 2004 [Wong & Thomas, 2004]. Es un campo multidisciplinario que combina el razonamiento humano con las capacidades analíticas de los ordenadores. Está enfocado en la investigación de la visualización de la información. Permite realizar análisis interactivos de datos, que combinados al conocimiento que posee el usuario en su

área de trabajo (experto en dicha área), dio origen al desarrollo del campo de la exploración visual de datos [Keim, 2001; Keim *et al.*, 2010]. Es particularmente útil en el proceso de selección de características ya que al combinar técnicas de análisis automatizados con visualizaciones interactivas permite una mayor comprensión, razonamiento y toma de decisiones efectivas sobre conjuntos de datos masivos y complejos [Cook & Thomas, 2005; Martínez *et al.*, 2015].

En 2015, la Dra. María Jimena Martínez desarrolló la herramienta de análisis visual de descriptores llamada con el acrónimo VIDEAN (*Visual and Interactive DEscriptor ANalysis*) [Martínez *et al.*, 2015]. Esta herramienta permite explorar los datos de manera visual e interactiva, proporcionando retroalimentación de los expertos en la selección de descriptores haciendo que este proceso no resulte una caja negra totalmente automatizada. De esta manera se logran modelos más informativos al momento de realizar la interpretación (por ejemplo, fisicoquímica) del conjunto de variables que forman a dicho modelo. VIDEAN muestra diferentes tipos de representaciones visuales coordinadas que capturan las relaciones e interacciones entre descriptores y descriptor-propiedad:

- **Grafos no dirigidos**: representan asociaciones entre pares de descriptores. Permiten evitar conjuntos de descriptores redundantes y comparar entre subconjuntos de descriptores alternativos.
- **Grafo bipartito**: representa la relación entre los subconjuntos de descriptores candidatos o alternativos y los descriptores individuales. Permite analizar la coexistencia de un descriptor en los diferentes conjuntos.
- **Área de trazado interactivo**: muestra diferentes relaciones entre cada descriptor y la propiedad, a través de gráficas de dispersión de puntos e histogramas.

En las próximas subsecciones se presentarán distintos escenarios y aplicaciones, en el ámbito de la Quimioinformática, donde se exploró el uso de VIDEAN. Esto sirvió de entrenamiento en el uso de técnicas de analítica visual para modelado QSAR/QSPR.

3.4.2.a. USO DE VIDEAN PARA EL ANÁLISIS DE CONJUNTOS DE DESCRIPTORES: SIN INTERVENCIÓN EN LA COMPOSICIÓN DE LOS SUBCONJUNTOS

Una de las posibilidades que brinda VIDEAN es analizar un conjunto de descriptores seleccionado, es decir, un subconjunto de la totalidad inicial de descriptores en términos de las relaciones existentes entre los descriptores tomados de a pares, con el objetivo de asegurar la ausencia de redundancias. En

estos escenarios, el rol del análisis visual apunta a reconfirmar la salida de un proceso de Selección de Características ya finalizado. A continuación, se presenta detalladamente un caso de estudio que ilustra el uso de VIDEAN post-proceso de selección de características. Luego se menciona brevemente un segundo caso, a modo de ejemplo.

CASO I: Predicción del IC50 para la proteína BACE I

Este es un caso aplicado al estudio de inhibidores de la proteína BACE I publicado recientemente donde se discuten múltiples aspectos referidos a cómo predecir inhibidores de BACE 1 mediante modelado QSAR [Ponzoni *et al.*, 2019]. Esta subsección se focaliza en resumir los aspectos metodológicos generales seguidos, a fin de comprender el rol del uso de VIDEAN. La metodología utilizada emplea, como rasgo distintivo, tanto la combinación de subconjuntos de descriptores, como la eliminación hacia atrás (*backward elimination*) de descriptores. El subconjunto resultante de este proceso es analizado con VIDEAN, para asegurar su calidad. BACE I, o β-secretasa I, es una enzima de segmentación de proteína precursora de amiloide (APP) del sitio beta I. Los péptidos β-amiloides largos requieren dos escisiones secuenciales para evitar su acumulación en el cerebro. Esta acumulación provoca alteraciones neuronales que causan la muerte neuronal y puede derivar en la enfermedad de Alzheimer (EA). Evitar la actividad de BACE I, evitaría la acumulación de los β-amiloide retardando o desacelerando la progresión a largo plazo de la EA. En la Figura 3.12 se esquematiza la secuencia de escisiones necesarias.

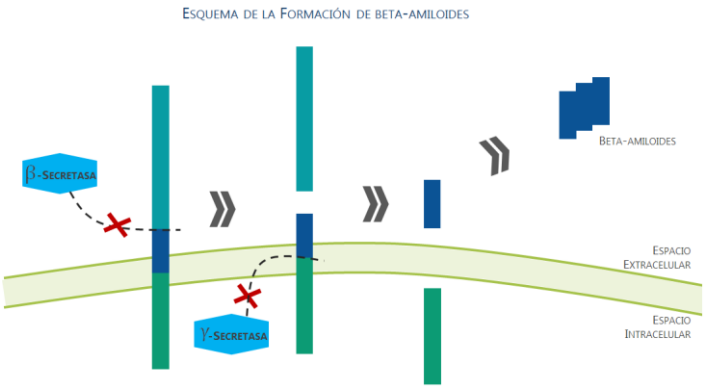


FIGURA 3.12 ESQUEMA DE LA FORMACIÓN DE LOS BETA-AMILOIDES FAVORECIDOS POR LA PRESENCIA DE LA ENZIMA DE SEGMENTACIÓN DE PROTEÍNA PRECURSORA DE AMILOIDE (APP) DEL SITIO BETA I.

El diseño de la metodología propuesta requiere inferir un modelo QSAR capaz de identificar posibles inhibidores de BACE I, es decir, medicamentos para bloquear la β-secretesa. Se utilizó la propiedad concentración inhibitoria máxima media (CI50), en otras palabras, se evaluó a qué nivel inhibía cada uno de los compuestos

a BACE I. Se trabajó con una base de datos de 215 compuestos químicamente diversos, con diferentes *scaffolds*, que reflejan un amplio espacio químico en el rango de medicamentos. Para categorizar los compuestos se definió un umbral para la discretización de los valores de IC50: Actividad alta (HA del inglés *High Activity*) para IC50 ≤ 1000 [nM] y Actividad baja/nula (LA del inglés *Low Activity*) para IC50> 1000 [nM]. Finalmente, 126 compuestos forman parte de HA y los restantes 89 compuestos de LA. Cada uno de los 215 compuestos de la base de datos fue caracterizado por 1867 descriptores moleculares calculados con la herramienta Dragon [Dragon, 2007]. Se realizó la Selección de Características con dos herramientas alternativas, por un lado, con DELPHOS [Soto *et al.*, 2009a, 2010] un subconjunto (Subconjunto A), y por otro, otros dos (Subconjunto B y C) extraídos con WEKA [Hall et al., 2009]. Además, se trabajó con un cuarto subconjunto de características extraído de la literatura (Subconjunto D) reportado años atrás por Gupta, quien propuso un modelo QSAR para modelar IC50 para BACE I [Gupta, 2014]. Luego, cada subconjunto es entrenado con el 75% de la base de datos (conjunto de entrenamiento) con la herramienta WEKA a través de tres métodos de aprendizaje maquinal: Redes Neuronales, Bosques Aleatorios y Comité Aleatorio. Los mejores modelos son seleccionados para continuar con la experimentación. Esta parte de la metodología puede verse en la Figura 3.13.

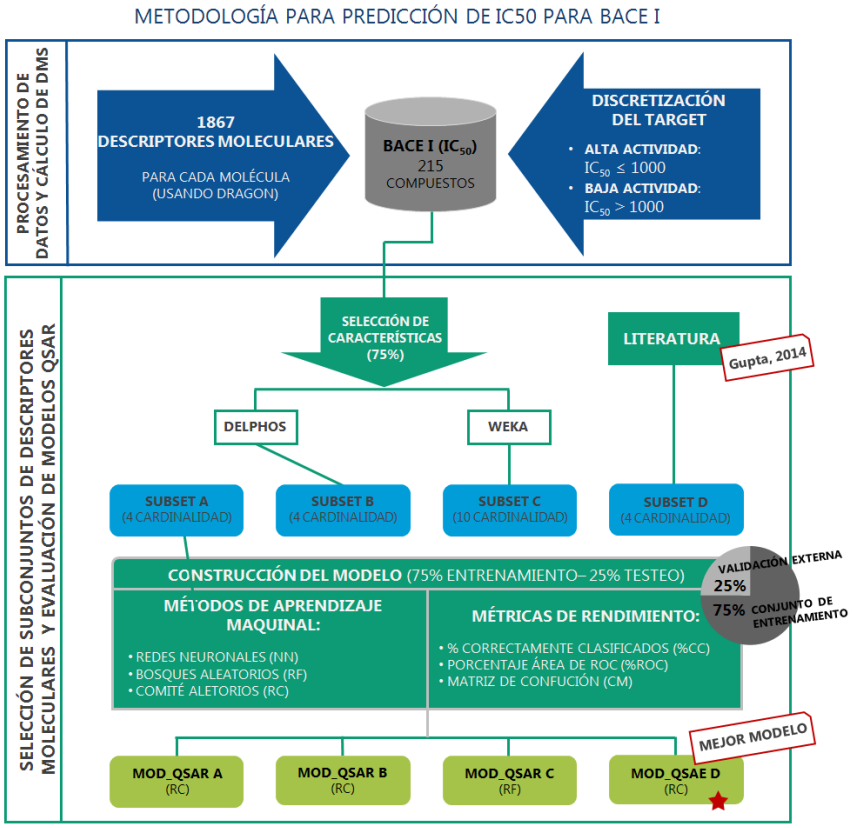


FIGURA 3.13 PASOS METODOLÓGICOS PARA LA CONSTRUCCIÓN DE MODELOS PREDICTIVOS PARA IC50.

El aspecto más distintivo de la estrategia QSAR propuesta, es la combinación de la hibridación de los subconjuntos con la eliminación hacia atrás de los descriptores, lo que contribuye a mejorar la calidad del modelo QSAR final. La primera hibridación corresponde a la combinación del subconjunto A con el subconjunto D, y es llamado HS1. El segundo, HS2, corresponde a la combinación del Subconjunto B con D y de la misma manera HS3 corresponde a la hibridación del Subconjunto C y D. El subconjunto HS4 corresponde a la unión global de los 4 conjuntos previos (A, B, C y D). Estos subconjuntos fueron entrenados con los mismos métodos que en la etapa anterior, por lo que cada modelo recibe el mismo nombre que su subconjunto precedido por Mod_QSAR. El mejor rendimiento fue conseguido por el Mod_QSAR HS1 entrenado con Comité Aleatorio (Figura 3.14).

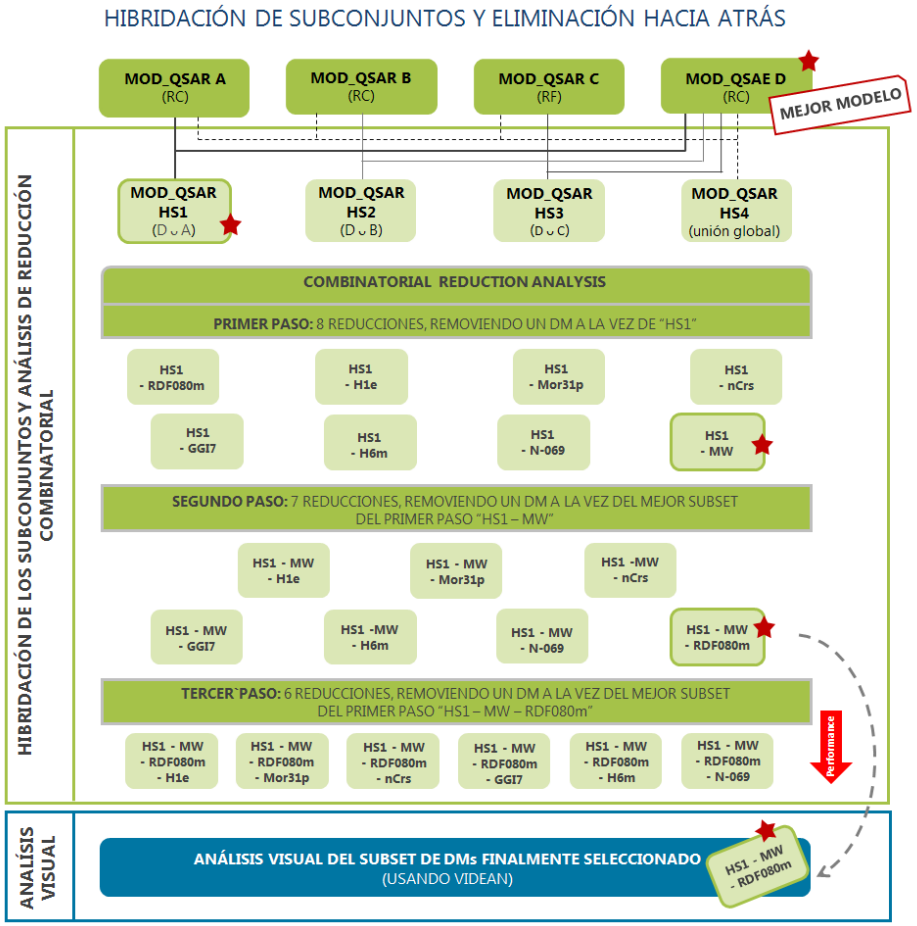


FIGURA 3.14 PASOS METODOLÓGICOS DE LA INFERENCIA DEL SUBCONJUNTO FINAL DE DESCRIPTORES QUE INGRESAN AL ANÁLISIS VISUAL.

El subconjunto HS1 tiene ocho descriptores, así que mediante el proceso de eliminación hacia atrás se obtienen ocho subconjuntos nuevos. Estos subconjuntos son analizados en términos estadísticos. Si alguno mejora, se elige el de mayor rendimiento para avanzar en el siguiente paso. Ahora, el mejor subconjunto tiene

siete descriptores. Nuevamente, mediante eliminación hacia atrás se obtienen otros siete nuevos subconjuntos y se elige el mejor para continuar. Al eliminar un nuevo descriptor, ninguno de los modelos mejora, por lo que este último paso se descarta (Figura 3.14). Se selecciona como mejor al Mod_QSPR HS1–MW–RDF080m para continuar con otro paso relevante en la metodología, el de la analítica visual aplicada al subconjunto de descriptores moleculares resultante. Esto permite garantizar la ausencia de redundancia de información en el modelo QSPR final.

Como última fase de la metodología propuesta, se analizó la correlación entre pares, de los seis, descriptores del modelo que resultó elegido. Para realizar este análisis estadístico en forma visual e interactiva se utilizó VIDEAN. La Figura 3.15 muestra la relación entre los descriptores en el modo de correlación de Kendall. En esta representación, los tonos anaranjados claro y azules claros de las aristas (correlación) entre los nodos (descriptores) muestran un bajo nivel de correlación. Puede verse que la mayoría son amarillos anaranjados, es decir la menor categoría de correlación. Este resultado es el esperado, lo que confirma que cada descriptor está aportando información única al modelo, lo que implica que el objetivo principal de esta estrategia se cumplió obteniendo un modelo generalizable y de baja cardinalidad.

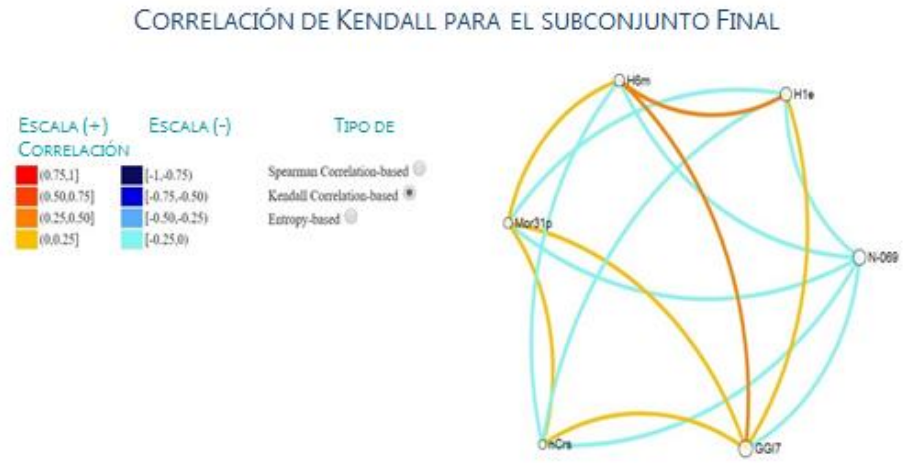


FIGURA 3.15 INTERFAZ DE VIDEAN PARA EL ANÁLISIS DE CORRELACIÓN DE KENDALL ENTRE LOS DESCRIPTORES DEL SUBCONJUNTO FINAL.

Concluyendo, con la metodología propuesta se infirió un modelo robusto QSAR preservando la baja cardinalidad y que ha mejorado, en términos de las dos métricas de rendimiento que se tuvieron en cuenta (porcentaje de moléculas correctamente clasificadas y porcentaje de área de ROC), aproximadamente un 8% con respecto al modelo de Gupta. Además, el riesgo de correlación aleatoria en el modelo QSAR propuesto fue descartado mediante la ejecución y el análisis de la

selección aleatoria de características y los experimentos de aleatorización. Debido a la amplia diversidad química de la base de datos utilizada, en comparación con estudios anteriores, la aplicabilidad del modelo propuesto incrementa. Por lo tanto, los resultados obtenidos por esta nueva estrategia muestran que este enfoque contribuyó a lograr un modelo QSAR que puede ser un método de detección virtual útil para la predicción de los inhibidores de BACE1.

CASO II: Predicción de la propiedad $LogP_{liver}$

Una de las posibilidades que brinda VIDEAN es analizar un grupo de cuatro subconjuntos alternativos de descriptores (obtenidos a partir de un conjunto inicial de 1391 de DMs con DELPHOS) con capacidades predictivas similares y colaborar en la elección de uno de ellos. Al aplicarlo a la predicción de la propiedad $LogP_{liver}$ (coeficientes de partición sangre-hígado) para COVs (compuestos orgánicos volátiles), los resultados mostraron la ventaja de utilizar esta herramienta interactiva para seleccionar conjuntos de descriptores con características deseables (baja cardinalidad, alta interpretabilidad, redundancia baja y alto rendimiento estadístico) de forma exploratoria y versátil [Martínez *et al.*, 2014b; Martínez *et al.*, 2015].

3.4.2.b. USO DE VIDEAN PARA EL ANÁLISIS DE CONJUNTOS DE DESCRIPTORES: CON INTERVENCIÓN EN LA COMPOSICIÓN DE LOS SUBCONJUNTOS

Otra de las posibilidades que brinda VIDEAN es emplear el análisis visual como estrategia para llevar adelante la selección de descriptores mediante el aprovechamiento del conocimiento sobre la propiedad en estudio aportado por un experto en el área. En particular, aquí se presenta un caso de estudio relacionado con la predicción de la propiedad $LogP_{liver}$ (coeficiente de partición sangre-hígado) para COVs (compuestos orgánicos volátiles), donde un experto en química fue interviniendo modelos iniciales, agregando y quitando descriptores estratégicamente con criterio fisicoquímico, hasta llegar a un modelo final con buena performance e interpretabilidad. En esto reside la principal diferencia en la estratégica metodológica del caso II de la subsección anterior.

Por un lado, se atravesó un proceso de selección de características tradicional utilizando WEKA [Hall *et al.*, 2009] y por el otro, se aplicó la analítica visual usando VIDEAN [Martínez *et al.*, 2015], para construir un subconjunto alternativo formado por aquellos descriptores que no presentan información mutua y completan juntos el espacio de información en el que se trabaja. De este modo, en esta subsección se intenta mostrar la efectividad del análisis visual como técnica de selección *per se*. A

continuación, se detallan los resultados obtenidos y las discusiones para el siguiente caso de estudio.

CASO III: Categorización de COVs

El aire interior típicamente contiene muchos COVs, que son químicos orgánicos que contienen carbono y presentan alta presión de vapor y baja solubilidad en agua. La exposición a los COVs puede acarrear con frecuencia, una variedad de síntomas asociados con enfermedades respiratorias y con alergias. Un ejemplo es el Síndrome del Edificio Enfermo, que es un conjunto de síntomas relacionados con concentraciones de ciertos COVs específicos presentes en los edificios donde las personas permanecen períodos prolongados de tiempo.

Los COVs que ingresan al organismo por vía inhalatoria se distribuyen según afinidad en los diferentes tejidos. Una de las propiedades más relevantes estudiadas al respecto es P_{liver}, un coeficiente de partición sangre-hígado determinado *in vitro* que resulta de relacionar las concentración de los COVs en los medios: aire:sangre y aire:hígado [Dashtbozorgi & Golmohammadi, 2010]. En la literatura se han propuesto algunos modelos computacionales para predecir $LogP_{liver}$ utilizando enfoques QSAR de regresión [Abraham *et al.*, 2007; Palomba *et al.*, 2012a], aunque ninguno de ellos se centró en la estimación de la toxicidad con enfoques QSAR de clasificación. En el presente caso de estudio abordamos este último enfoque para predecir $LogP_{liver}$ y sumado al uso de VIDEAN, el conjunto de la metodología resultó en modelos QSAR más interpretables en términos fisicoquímicos [Cravero *et al.*, 2017a].

Se comenzó discretizando los valores de $LogP_{liver}$ de los COVs de la base de datos empleada. La base de datos utilizada fue propuesta por Abraham cuenta con 122 moléculas [Abraham *et al.*, 2007], a las cuales se les calcularon varios descriptores moleculares utilizando DRAGON [Dragon, 2007]. Por un lado, están los compuestos que presentan afinidad por la sangre (medio acuoso), por el otro, aquellos con afinidad por el hígado (medio graso) y, finalmente, un tercer grupo que no presenta una preferencia entre estos medios (zona gris). En la Figura 3.16 puede verse el rango de $LogP_{liver}$ que se tuvo en cuenta para definir cada una de las categorías y cuántas moléculas pertenecen a cada una de ellas.

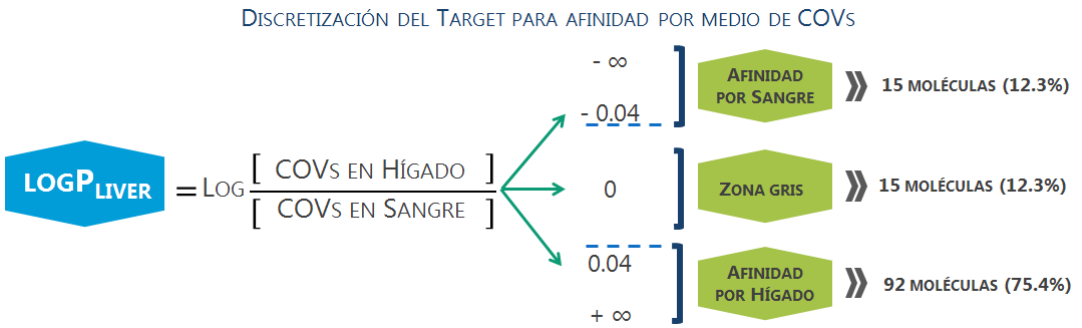


FIGURA 3.16 CRITERIOS DE DISCRETIZACION DE LA PROPIEDAD $LogP_{LIVER}$ EN TRES CLASES.

En la Figura 3.17 se resume la metodología seguida para este estudio. A partir de los descriptores moleculares derivados de la metodología DRAGON-DELPHOS, donde se seleccionaron los mejores 10 subconjuntos integrados por un total de 17 descriptores. Con DRAGON se calcularon los descriptores y con DELPHOS se redujo a 17 la cantidad de los mismos. Se siguieron dos caminos para seleccionar los mejores subconjuntos de descriptores a partir de los cuales entrenar los modelos QSAR, por un lado, se atravesó un proceso de Selección de Características tradicional utilizando WEKA [Hall *et al.*, 2009]. Por el otro se utilizó la analítica visual, a través de VIDEAN [Martínez *et al.*, 2015], para construir un subconjunto alternativo teniendo aquellos descriptores que no presentan información mutua y completan juntos el espacio de información en el que se trabaja. Finalmente, se cuenta con tres subconjuntos: uno derivado de esta última alternativa compuesto por 5 descriptores llamado VA_{COV} y dos, seleccionados por WEKA, uno de ellos de cardinalidad 4 ($FS4_{COV}$) y otro, con 5 descriptores ($FS5_{COV}$).

Se entrenaron los modelos QSAR con tres métodos disponibles en la herramienta Weka [Hall *et al.*, 2009]: Redes Neuronales, Arboles Aleatorios y Comité Aleatorio. Se usaron cuatro alternativas de partición entrenamiento/prueba para la base de datos: 75/25, 66/34, 50/50 y validación cruzada de 10 pliegues. En el desarrollo de este modelo QSAR fue importante identificar a cuál clase pertenece cada COVs. Por este motivo se prioriza la información que brinda la matriz de confusión por sobre las métricas globales de desempeño. Finalmente, el subconjunto VA_{COV}, el subconjunto obtenido aplicando analítica visual, usando una partición 50/50, y entrenado con Comités Aleatorios, es seleccionado como el mejor modelo. Alcanza un 72.13% de COVs Correctamente Clasificados (%CC) y tiene un área bajo de curva de ROC igual a 0.83.

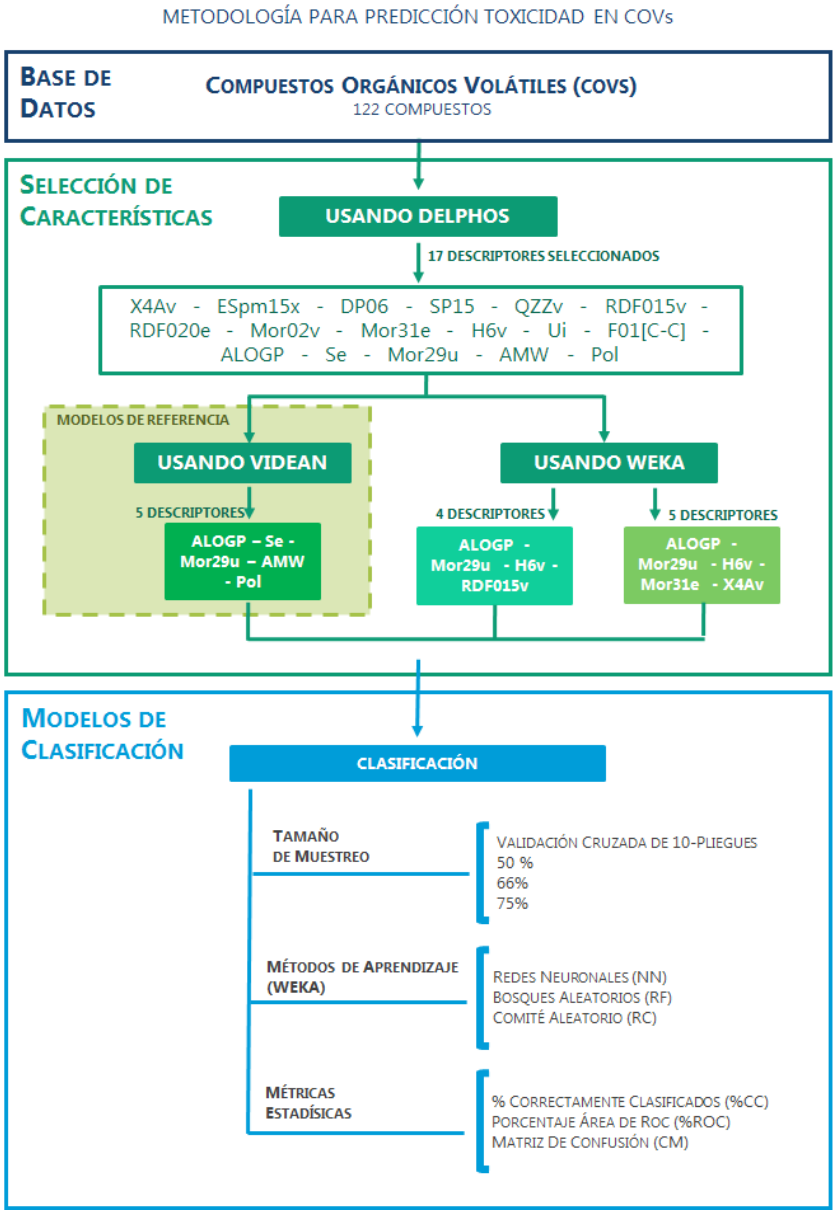


FIGURA 3.17 METODOLOGÍA PARA LA CONSTRUCCIÓN DE LOS MODELOS QSAR DE CLASIFICACIÓN.

Concluyendo, mientras que un modelo de regresión sólo brinda una salida numérica, que representa el valor de la propiedad ($LogP_{liver}$), el modelo QSAR de clasificación de esta experimentación puede predecir la afinidad de un COV por un medio (acuoso/graso), lo que puede ayudar en el desarrollo de modelos computacionales farmacocinéticos de base fisiológica indispensables en el área de Informática de la Salud. La analítica Visual en este caso demostró ser una estrategia adecuada para la selección de descriptores moleculares, resultado de la intervención de un experto en el tema, generando un modelo QSAR con rendimiento competitivo frente a los métodos tradicionales.

3.4.2.c. CONCLUSIONES SOBRE EL IMPACTO DE LA UTILIZACIÓN DE LA ANALÍTICA VISUAL

Usar Analítica Visual es una alternativa viable en la construcción de un subconjunto candidato en el modelado QSAR, ya que permite la intervención de un experto que orienta la búsqueda de descriptores según una interpretación basada en conocimiento fisicoquímico de la relación estructura propiedad. Sin embargo, es necesario notar que primero se debe atravesar un proceso tradicional de Selección de Características para reducir el universo de posibles descriptores y recién entonces poder abordarlos mediante Analítica Visual. Por otro lado, resulta de mucha utilidad aplicarla para elegir entre posibles conjuntos de descriptores (alternativos) brindando información oportuna para optar por uno de ellos evitando así tener que entrenar modelos con conjuntos de descriptores que *a priori* no son buenos candidatos.

3.4.3. METODOLOGÍA HÍBRIDA PARA LA OBTENCIÓN DE DESCRIPTORES

Para la construcción de modelos QSPR/QSAR, se requieren múltiples DMs que como fue definido anteriormente se refiere a la información química codificada por números a partir de una representación simbólica de una molécula. Entre estos descriptores solo se eligen aquellos que mejor se ajusten a la propiedad que se quiere predecir, y para esto se necesita aplicar un método de Selección de Características o uno de Aprendizaje de Características. En los métodos de Selección de Descriptores las variables (descriptores) se introducen en el modelo de manera algorítmica y una función de aptitud (o criterios de selección) determina qué variable debe retenerse o eliminarse del modelo. En el caso de los métodos de Aprendizaje de Descriptores, el conjunto original de variables se proyecta en nuevas variables en un espacio dimensional reducido, sin pérdidas de información. Ambas técnicas son reportadas en la literatura, como alternativas y/o competitivas.

Con el propósito de utilizar de manera combinada las fortalezas de ambas técnicas exploramos lo que nosotros denominamos Metodología Híbrida, la cual se empleó para predecir propiedades ADMET [Ponzoni *et al.*, 2017]. La metodología Híbrida permite la inferencia de modelos QSPR/QSAR de predicción combinando descriptores moleculares provenientes de técnicas de Selección y Aprendizaje de Características, con los objetivos de reducir el esfuerzo computacional de los modelos mediante mapeos entre dominios, identificando así los aspectos relevantes en la generación de un modelo y los aportes provenientes de las

distintas variables que inciden sobre un modelo, para así mejorar la generalización de los mismos mediante la reducción del sobreajuste.

La Figura 3.18 presenta un esquema de la Metodología Híbrida propuesta. La transformación de la información química contenida en una base de datos sigue dos ramas metodológicas hasta unirse en el modelado QSPR/QSAR. Por un lado, está la Selección de Características en la izquierda del esquema, donde el primer paso es el cálculo de descriptores moleculares. Por el otro lado, en la derecha del esquema, está el proceso de Aprendizaje de Características y su primer paso es la extracción de la información codificada en la estructura química de cada una de las moléculas en la base de datos. El último paso de la metodología, que incluye un módulo de Analítica Visual, y representa a la integración o hibridación de los descriptores de ambos procesos (ramas) para la confección de modelos QSPR/QSAR.

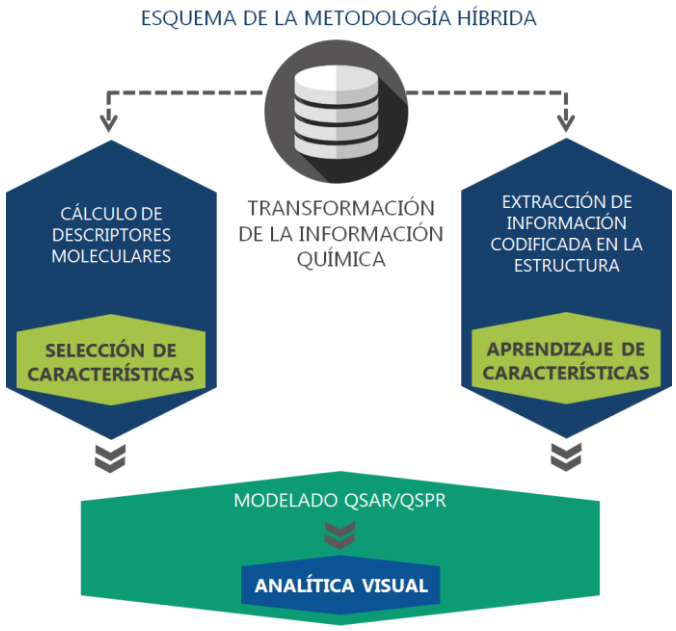


FIGURA 3.18 ESQUEMA DE LA METODOLOGÍA HÍBRIDA PARA LA OBTENCIÓN DE DESCRIPTORES Y POSTERIOR INFERENCIA DE MODELOS QSPR/QSAR.

La metodología propuesta se aplicó al desarrollo de modelos QSAR para la predicción de tres actividades biológicas ligadas al diseño racional de fármacos: 1) Barrera Hematoencefálica (*Blood-Brain Barrier* –BBB-), 2) Absorción Intestinal Humana (*Human Intestinal Absorption* –HIA-) y 3) Exceso Enantiomérico (*Enantiomeric Excess* -EE-) [Ponzoni *et al.*, 2017]. Este tipo de modelos QSAR permiten la selección virtual de medicamentos antes de la síntesis de nuevos diseños, y juegan un papel central en el estudio de propiedades fisicoquímicas de fármacos. En particular, las dos grandes ramas de la Metodología Híbrida fueron

realizadas por la dupla de herramientas DRAGON-DELPHOS para la Selección de Características y por CODES-TSAR para el Aprendizaje de Características [Dorronsoro *et al.*, 2004]. La fase de entrenamiento de los modelos QSAR se realizó con la herramienta Weka [Hall *et al.*, 2009], variando el porcentaje utilizado como conjunto de entrenamiento (1:2, 2:3 y 3:4). Luego, se utilizaron cinco métodos de entrenamiento diferentes: Regresión Lineal (RL), Árboles de Decisión (AD), Redes Neuronales (RN), Bosques Aleatorios (BA) y Comité Aleatorio (CA). Varias métricas estadísticas fueron tenidas en cuenta para evaluar el desempeño de los modelos de acuerdo a si eran modelos de regresión o clasificación. En la Figura 3.19 puede verse un resumen gráfico de la experimentación realizada.

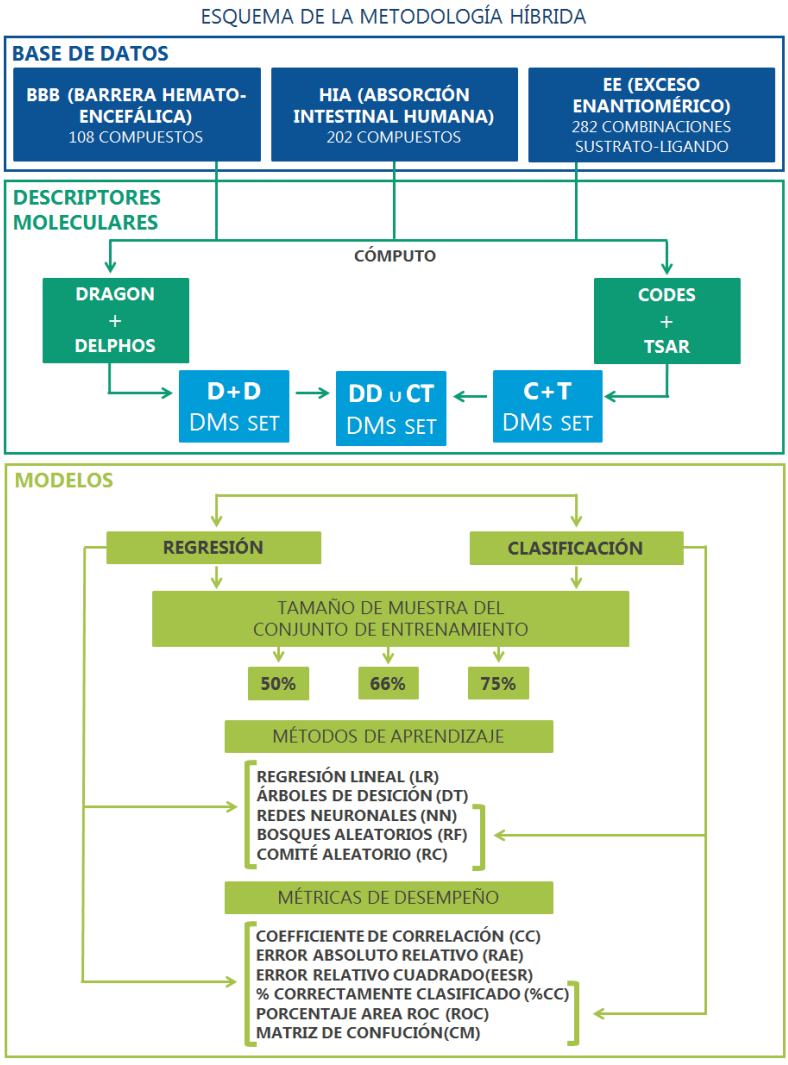


FIGURA 3.19 ESQUEMA DE LA METODOLOGÍA HIBRIDA APLICADA A LA PREDICCIÓN DE TRES PROPIEDADES DE INTERÉS FARMACOLÓGICO.

El objetivo de esta sección es presentar las ventajas de utilizar la Metodología Híbrida y por lo tanto sólo se focalizará en los modelos QSAR de regresión para predecir la Absorción Intestinal Humana para ejemplificar su uso y potencialidad. La

discusión sobre los restantes modelos puede ser consultada en el trabajo original [Ponzoni *et al.*, 2017]. La Absorción Intestinal Humana es el proceso mediante el cual los medicamentos administrados por vía oral se absorben desde el intestino hacia el torrente sanguíneo. La mayoría de los medicamentos disponibles en el mercado se administran por vía oral y para que sean eficaces es necesario que puedan ser absorbidos en el torrente sanguíneo. Para la base de datos de HIA se recopiló de la literatura [Zhao *et al.*, 2001] un conjunto de datos de 202 compuestos con los valores experimentales del porcentaje de absorción intestinal humana (% HIA). Es decir, el porcentaje del fármaco administrado por vía oral que llega a la vena porta hepática. El porcentaje de HIA varía entre 0 y 100 para los modelos de regresión.

Se trabajó con los siguientes 5 subconjuntos: los dos que presentaron valores RAE (error absoluto relativo) más bajos seleccionados por DELPHOS ($M5_{HIA}$ y $M9_{HIA}$), el subconjunto obtenido por CODES-TSAR (CT_{HIA}) y los dos subconjuntos construidos mediante la unión de los anteriores ($M5_{HIA} \cup CT_{HIA}$ y $M9_{HIA} \cup CT_{HIA}$). Usando estos subconjuntos de descriptores, se infirieron varios modelos QSAR de regresión y clasificación aplicando diferentes métodos de aprendizaje automático. El mejor modelo de regresión QSAR se obtuvo utilizando la partición 3:4 de la base de datos, entrenando con Redes Neuronales el subconjunto de descriptores $M5_{HIA} \cup CT_{HIA}$, uno de los propuestos por la metodología híbrida. Posee una cardinalidad igual a siete, donde cuatro de los descriptores corresponden a la metodología DRAGON-DELPHOS y los tres restantes, a CODES-TSAR.

El subconjunto $M5_{HIA}$ está integrado por descriptores moleculares que corresponden a diferentes familias. La selección del Peso Molecular Promedio conocido como AMW por sus siglas en inglés para *Average Molecular Weight,* puede deberse a la importancia del peso molecular para la HIA, en concordancia con la regla de Lipinski que establece que un MW<500 [Da] para obtener compuestos activos por vía oral. Otro de los descriptores seleccionado para $M5_{HIA}$ es el área de superficie polar topológica (TPSA) y puede explicarse debido a la bien conocida correlación entre el área superficial polar de las moléculas y su capacidad para someterse a la absorción intestinal humana. También hay otros dos descriptores: MATS7m (autocorrelación de Moran de retraso 7 ponderada por la masa) y ESpm01d (momento espectral 01 de la matriz de adyacencia de borde ponderada por los momentos dipolares). Este último tiene una interpretación directa en términos de fragmentos estructurales de las moléculas que tienen cierta semejanza con los esquemas aditivos de Free-Wilson y Fujita-Ban [Ponzoni *et al.*, 2017].

ANÁLISIS VISUAL PARA MODELOS QSPR DE REGRESIÓN DE HIA:

Se utilizó VIDEAN para analizar la relación entre los descriptores proporcionados por $M5_{HIA} \cup CT_{HIA}$. En este caso la información mutua es medida por el modo basado en Entropía, donde es deseable una baja entropía (baja información mutua) entre descriptores. La Figura 3.20 muestra los lazos de los descriptores CODES-TSAR en color púrpura, lo que significa alta información mutua, pero cada uno de ellos (CODES-T1, CODES-T2 y CODES-T3) presenta una baja entropía con los demás descriptores de $M5_{HIA}$ (rosado y rosado claro). Este hecho demuestra que la información proporcionada por ambos subconjuntos, $M5_{HIA}$ y CT_{HIA}, es complementaria soportando la idea de utilizar la metodología híbrida.

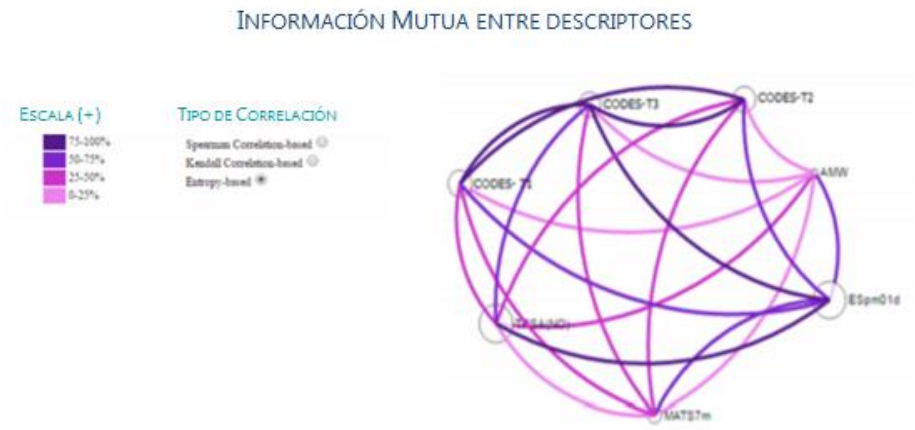


FIGURA 3.20. INTERFAZ DE VIDEAN PARA EL ANÁLISIS DE INFORMACIÓN MUTUA ENTRE LOS DESCRIPTORES DEL CONJUNTO $M5_{HIA} \cup CT_{HIA}$.

3.4.3.a. CONCLUSIONES SOBRE EL IMPACTO DE LA UTILIZACIÓN DE LA METODOLOGÍA HÍBRIDA

La estrategia Híbrida propuesta permite la confección de modelos QSAR de predicción, combinando descriptores provenientes de métodos de Selección de Características y de Aprendizaje de Características. Al analizar los diferentes escenarios planteados se pudieron evaluar los beneficios potenciales relacionados con la hibridación en el modelado QSAR. Se pudo concluir que la hibridación de ambas estrategias puede ser útil tanto si se usan enfoques tanto de regresión como de clasificación, aunque encontramos que el rendimiento de los modelos QSAR inferidos dependía, entre otros factores, de las características del conjunto de datos. Por lo tanto, cada caso deberá ser evaluado según su particularidad.

Síntesis y Conclusiones del Capítulo 3

En este capítulo se introdujeron los conceptos básicos referidos a la Informática Molecular, haciendo un recorrido histórico por los diferentes modos de representación computacional de moléculas, desde el modelo de barras y esferas de von Hoffman, hasta los modelos computacionales en 3D o 4D conseguidos en la actualidad. Además, se presentaron las ventajas y desventajas de los distintos tipos de formatos computacionales para representación de moléculas a través de su caracterización mediante diferentes tipos de descriptores moleculares. Esto es un punto fundamental para dilucidar posibles patrones existentes en las bases de datos que contienen, por ejemplo, potenciales fármacos de interés. Además, se realizó un resumen de la evolución del modelado QSAR/QSPR (Relación Cuantitativa de Estructura-Actividad/Propiedad) desde su aplicación a pequeños conjuntos de moléculas estructuralmente similares con técnicas estadísticas básicas, hasta su aplicación a grandes volúmenes de datos con sofisticados métodos de Aprendizaje Maquinal, es decir, hasta convertirse en un método guiado por datos y las cuestiones vinculadas a resolver los desafíos algorítmicos que fueron requeridos.

Concretamente, en cuanto contribuciones de esta tesis en el área, se exploró el uso de la Analítica Visual tanto en la construcción de nuevos subconjuntos de descriptores moleculares como en el soporte para la elección entre subconjuntos alternativos o candidatos a ser entrenados para inferir un modelo QSAR/QSPR. Otro aporte fue la propuesta de la Metodología Híbrida que busca combinar las potencialidades de las técnicas de Selección de Características y el Aprendizaje de características. La hibridación de ambas técnicas permite la confección de modelos QSAR/QSPR de predicción combinando descriptores provenientes de ambas, que puede ser útil tanto para enfoques de regresión como de clasificación.

INFORMÁTICA DE POLÍMEROS

CAPÍTULO 4

Este capítulo, luego de describir las particularidades de los materiales poliméricos, hace un recorrido bibliográfico sobre los trabajos realizados en el contexto de la Informática de Polímeros. A continuación, se enfoca en la predicción a través de modelos QSPR de propiedades mecánicas para polímeros de alto peso molecular siguiendo tres metodologías diferentes. Finaliza planteando y respondiendo lo que se considera la primera pregunta de investigación de esta tesis. Cabe destacar que en este capítulo no se aborda la complejidad más desafiante que presentan los materiales poliméricos, la polidispersión. Se trabaja en modelado QSPR sobre modelos moleculares simples.

4.1. CONCEPTOS DE INFORMÁTICA DE POLÍMEROS

La cotidianidad se ve invadida por imágenes y propagandas donde se presentan a los plásticos[1] como uno de los protagonistas del deterioro del medio ambiente, y se crearon muchas campañas para sustituirlos, reducirlos y reciclarlos. Las matrices ambientales, como el agua y el suelo, son las más afectadas, porque el tiempo de degradación de estos materiales oscila entre 150 y 1000 años, dependiendo del tipo de plástico y condiciones del medio. El consumo de productos plásticos ha aumentado exponencialmente en las últimas décadas, debido a las bondades, ventajas y bajo costo que poseen. Pero, ¿es posible imaginar hoy una vida sin polímeros sintéticos? Deshacerse completamente de los plásticos es algo muy poco probable, más aún, es innecesario y poco razonable. Por mencionar solo algunas razones de esta afirmación, estos materiales han promovido cambios sustanciales en el cuidado de la salud, conteniendo y evitando el contagio de infecciones o enfermedades mediante jeringas y cánulas descartables, etc. También han ayudado a mejorar la calidad de vida de las personas, por ejemplo, con el desarrollo de válvulas cardíacas o prótesis ortopédicas vía impresión 3D entre otras aplicaciones. Sin materiales poliméricos, no habría internet (fibra óptica), ni agua potable en ciudades alejadas de las principales fuentes (caños de PVC), ni viajes de exploración espacial. Las industrias como la automotriz, la electrónica, la aeronáutica, la aeroespacial, de la

[1]Plásticos, es un nombre inapropiado (ya que solo refiere a un tipo de material polimérico) y muy difundido para los polímeros en general, que son moléculas sintéticas de cadena larga.

construcción, farmacéutica, médica, del empaque, de los juguetes, de la informática, la información, etc., han evolucionado moviendo las fronteras de los desarrollos como nunca antes en la historia de la humanidad gracias a los polímeros sintéticos. Pareciera que estos materiales llegaron para quedarse, por lo que se vuelve una necesidad estudiarlos en profundidad [Kutz, 2002].

Cualquier actividad ingenieril depende, en gran medida, de una cuidadosa e inteligente selección de materiales, que a menudo debe hacerse con el fin de satisfacer requisitos de rendimiento y/o costo. Así, surgen constantes demandas del mercado por nuevos materiales con propiedades específicas que atiendan una problemática particular de alguna de las industrias [Kutz, 2002]. En el desarrollo de materiales poliméricos, tanto en la industria como en la academia, la etapa de diseño tiende a ser guiada por algoritmos basados en inteligencia artificial, lo que deriva en ciclos de innovación cada vez más cortos, desde el diseño hasta la aplicación. La creciente interdisciplinariedad en este ámbito resulta en la Informática de Polímeros, como una ciencia cada vez más basada en datos, como lo son las estructuras moleculares asociadas a propiedades medidas y todos los parámetros incluidos en los ensayos [Adams & Murray, 2008; Audus & de Pablo, 2017].

Los polímeros sintéticos son, posiblemente, la clase más importante de materiales de esta era. La razón de su éxito puede deberse a la combinación de varias razones, por ejemplo, los polímeros *commodities* son de muy bajo costo y fáciles de procesar, la mayoría no son tóxicos, por lo general son estables y relativamente resistentes y, además, son muy versátiles en cuanto a procesamientos posteriores y aditivación [Adams, 2010]. Sin embargo, no siempre es posible desarrollar nuevos materiales por modificación de su formulación o procesamiento posterior y por lo tanto se vuelve imprescindible diseñarlos desde la estructura molecular. Aunque el enfoque típico en el diseño de nuevos materiales ha sido empírico (formulación, montaje, síntesis, procesamiento y testeo), en la actualidad se avanzó mucho en el conocimiento de las relaciones entre la estructura molecular de un material y sus propiedades [Van Krevelen, 2009; Hill *et al.*, 2016]. Estos avances condujeron a mejorar la capacidad de predecir las propiedades del material previo a su síntesis, que a su vez se traduce en enormes ahorros de recursos y tiempo. Esto ha hecho que actualmente se considere al modelado computacional tan relevante como la síntesis y la caracterización de moléculas, durante los procesos de descubrimiento de nuevos compuestos químicos [Adams, 2010; Nosengo, 2016]. Sin embargo, cuando se trata de materiales poliméricos, no es fácil conseguir estas predicciones ya que las variables que intervienen son muy

complejas desde un punto de vista cuantitativo y cualitativo. De este modo, obtener nuevos materiales poliméricos con propiedades específicas y novedosas ha resultado en uno de los campos más dinámicos de la ciencia moderna. El aprendizaje maquinal es un aliado en la reducción de las barreras entre el diseño, la síntesis, la caracterización y el modelado de químicos y materiales [Audus & de Pablo, 2017].

Desarrollar un polímero específico con las propiedades deseadas puede ser una tarea desafiante debido a la gran cantidad de variaciones posibles que se le pueden hacer a la estructura molecular y a la distribución de pesos moleculares. Las herramientas computacionales que colaboran en las decisiones de diseño son, por un lado, las bibliotecas de polímeros y, por otro, las técnicas de cribado virtual. Estas últimas asisten en la evaluación de relaciones cuantitativas de estructura/propiedad (QSPR) permitiendo la identificación de un "*match*" entre una cierta propiedad medida y la estructura molecular de los materiales del conjunto analizado [Meier & Webster, 2009]. En Informática de Polímeros, generalmente las bases de datos tienen decenas o cientos de muestras, en lugar de miles o decenas de miles como pasa en las bases de datos destinadas al diseño racional de drogas. No obstante, estos datos deben administrarse, manejarse y almacenarse, relacionándolos con sus metadatos como las condiciones de medición y las normas seguidas durante el ensayo.

La Informática de Polímeros necesita, para alcanzar su máximo potencial, el compromiso a largo plazo de la industria y la academia para generar y publicar repositorios de datos fidedignos y completos [Adams, 2010]. Persiguiendo este fin, surgió la Iniciativa Genoma de Materiales (*Materials Genome Initiative*) que busca potenciar la combinación sinérgica de experimento, teoría e informática para acelerar el ritmo del descubrimiento y diseño de nuevos materiales. La disponibilidad, o no, de materiales inteligentes y/o estratégicos afecta de formas impensadas cómo interactuamos con el mundo que nos rodea [de Pablo, *et al.*, 2019].

4.2. QUÍMICA DE POLÍMEROS

La Química de Polímeros es una ciencia multidisciplinaria que abarca desde la síntesis al estudio de las propiedades químicas de polímeros. Los polímeros fueron definidos por Hermann Staudinger, en 1920, como macromoléculas, trabajo por el cual recibió el 56º Nobel de Química en 1953 [Seymour & Carraher, 1998]. En términos sencillos, los polímeros se definen como macromoléculas compuestas por

cientos de miles, y más, de Unidades Repetitivas Estructurales (UREs) que se unen a lo largo de las cadenas que los forman. Como consecuencia, son macromoléculas significativamente más complejas que las estudiadas en el modelado QSAR de fármacos. Por lo tanto, las soluciones informáticas efectivas para el diseño de fármacos, no lo son para la Informática de Polímeros, donde se requiere de enfoques creativos a medida que soporten los fenómenos que subyacen en el comportamiento de estos materiales [Adams, 2010].

Como se dijo antes, los polímeros son macromoléculas formadas por la repetición de unidades químicas, y entonces estas podrían pensarse como la parte básica de la estructura. Para denominar a estos "bloques" se utilizan frecuentemente los términos: unidad repetitiva estructural (URE) y monómero, aunque químicamente hablando no son lo mismo. El monómero tiene un grupo funcional reactivo que dará lugar a la reacción de polimerización, y por lo tanto este grupo cambiará en el producto final, es decir, la URE no tendrá esa reactividad[2]. Por ejemplo, el monómero del poliestireno es el estireno que presenta un doble enlace reactivo, y la URE del estireno es la unidad que se repite n veces (grado de polimerización) a lo largo de toda la cadena, la cual ya no presenta el doble enlace (Figura 4.1). En la mayoría de los estudios publicados sobre modelado QSPR para materiales se utiliza como modelo molecular la URE con átomos de hidrógeno como átomos de terminación [Wu, *et al.*, 2016]. Esta estructura difiere de la URE básica ya que completa con hidrógenos terminales todos los enlaces posibles. En la Figura 4.1 se presenta un esquema de cada uno de los modelos moleculares para una mejor comprensión de lo antes descripto.

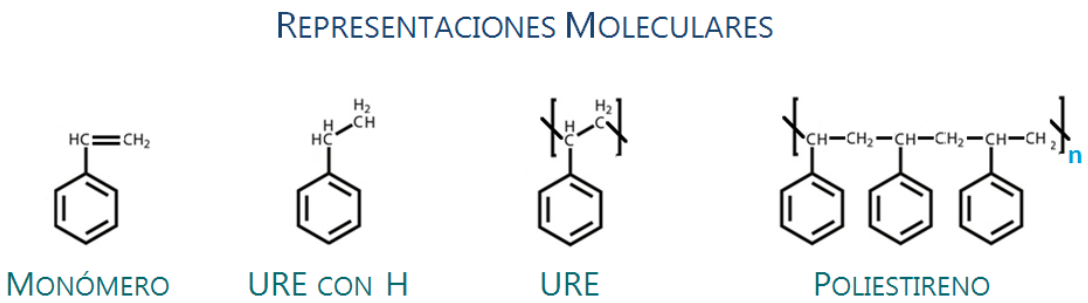


FIGURA 4. 1 REPRESENTACIONES MOLECULARES SINTÉTICAS, TOMANDO COMO EJEMPLO AL POLIESTIRENO.

El proceso mediante el cual los monómeros, iguales o diferentes, reaccionan entre sí por adición o condensación y forman otras moléculas de mayor peso, se conoce como reacción de polimerización. En la polimerización por adición se da un proceso de agregación de monómeros, tal que, la molécula final resulta en

[2] En esta tesis siempre se utilizará el termino URE para referirse a la unidad básica.

sucesivas UREs y en la polimerización por condensación una parte del monómero se elimina cuando pasa a integrar de la cadena polimérica [Seymour & Carraher, 1998]. Como producto de estas reacciones se obtienen cadenas de diferentes pesos moleculares, es decir un material no tendrá un único peso sino una familia de largos o peso de cadenas (polidispersión). Esta característica se puede modificar con las variables del proceso de polimerización, en busca de propiedades específicas. Debido a esto, los materiales poliméricos tienen un comportamiento único, comparados con otros. Sumado a esto, las diferentes cadenas se entrelazan físicamente entre sí. Para comprender mejor este fenómeno, se suele utilizar la metáfora del plato de espaguetis (*spaghetti-like*) donde un solo espagueti cocido se enreda con otros espaguetis en un plato de pasta (Figura 4.2). Los materiales poliméricos constituyen un caso particular de sistemas complejos, ya que la naturaleza polidispersa de estas macromoléculas le confieren al material propiedades que emergen más allá de las que podrían conseguirse con cada uno de los pesos por separado.

FIGURA 4. 2 ESQUEMA DE LA METÁFORA DEL PLATO DE ESPAGUETIS QUE EXPLICA CÓMO SE ENTRELAZAN FÍSICAMENTE LAS MOLÉCULAS DE UN MATERIAL POLIMÉRICO.

4.2.1. COMPLEJIDAD ESTRUCTURAL

Desde el punto de vista del peso, el término macromolécula no significa lo mismo en el contexto de la química de polímeros que en el de la biología molecular, donde a menudo es sinónimo de proteína. Desde el punto de vista estructural intra e intermolecular, los polímeros son moléculas mucho más inefables y complejas que los fármacos. Por lo tanto, la Informática de Polímeros, aún necesita desarrollar tecnologías que permitan la correcta descripción de estos. A un polímero se lo describe químicamente, entre otros parámetros, a través de: grado de polimerización, variedad de los monómeros, distribución de peso molecular (polidispersión) y pesos moleculares promedios. De acuerdo a la variedad o repetición de UREs los polímeros se clasifican en (1) Homopolímeros: están formados por la misma URE a lo largo de todas sus cadenas y (2) Copolímeros:

están formados por al menos 2 UREs diferentes. Además, suelen tenerse en cuenta sus propiedades térmicas, especialmente la temperatura de transición vítrea (Tg), que determina la temperatura en la cual el polímero aumenta abruptamente su movilidad y, como se verá más adelante, afecta marcadamente las propiedades mecánicas. Lo que respecta a peso, tamaño o largo de cadena se explica, en detalle, en la siguiente sección.

4.2.1.a. MATERIALES POLIDISPERSOS

Los polímeros son compuestos estructuralmente heterogéneos, aunque solo sea en términos del grado de polimerización. Todos los polímeros sintéticos son materiales polidispersos porque contienen cadenas poliméricas de diferentes longitudes. Retomando la metáfora del plato de espaguetis, podríamos pensar que el largo de cada uno de los espaguetis no es el mismo, ya que muchos de ellos se parten dentro del paquete y no todos lo hacen en el mismo punto (Figura 4.2). Tal es así, que aquellas reacciones de polimerización que tienen como objetivo un producto monodisperso, alcanzan, a lo sumo, Índices de Polidispersión (IPD) de aproximadamente 1.03 (un $IPD = 1$ indica que no existe polidispersión) [Ma *et al.*, 2005, Adams, 2010].

CURVA DE DISTRIBUCIÓN DE PESOS MOLECULARES

En el ámbito experimental existen diferentes técnicas para determinar la distribución de pesos moleculares de un polímero. Una de ellas es la cromatografía por permeación de geles (GPC), un tipo especial de cromatografía por exclusión de tamaño (SEC). Esta técnica separa los analitos según su volumen hidrodinámico, determinando la distribución de pesos moleculares de las cadenas que conforman al material polimérico. Al graficar el número de cadenas contra su peso molecular, se obtiene la curva de distribución de pesos moleculares (Figura 4.3). A partir de esta, pueden calcularse los parámetros que describen la población de tamaños, es decir, los diferentes pesos moleculares promedios. La distribución de pesos moleculares afecta el perfil de comportamiento mecánico del material polimérico, entre otras propiedades fisicoquímicas, y de aquí su importancia.

PESOS MOLECULARES PROMEDIOS

Comprender muchas de las propiedades físicas de los polímeros requiere el conocimiento sobre las longitudes de sus cadenas en términos del peso molecular. Este se puede relacionar con el número de UREs en la cadena (grado de polimerización). A su vez, deben describirse los pesos moleculares promedios calculados a partir de los pesos moleculares de cada una de las cadenas que componen el material.

Los pesos promedios más reportados son dos: peso molecular promedio en número (Mn) y peso molecular promedio en peso (Mw); además, también suele aparecer su cociente que sería el IPD ($IPD = Mw/Mn$). Habitualmente estos tres parámetros caracterizan la curva de polidispersión (Figura 4.3). Mn hace referencia al peso molecular promedio aritmético de todas las cadenas de polímero en la muestra. Mw hace referencia al peso molecular promedio en peso. Estos pesos moleculares promedios se calculan a partir de las siguientes fórmulas donde N_i es el número de moléculas con masa molecular M_i. Para peso molecular promedio en número: $M_n = \frac{\sum_i N_i M_i}{\sum_i N_i}$ y para el peso molecular promedio en peso: $M_w = \frac{\sum_i N_i M_i^2}{\sum_i N_i M_i}$

En la Figura 4.3 se ilustra una representación a modo de ejemplo de la curva de distribución de pesos y los pesos promedios. El IPD cuantifica la amplitud de la curva y por lo tanto la heterogeneidad en el largo de cadenas. Si IPD tiende a 1, el material será monodisperso y al contrario será polidisperso. Cuanto mayor es IPD, más ancha es la curva y más polidisperso es el material, lo que significa que conviven muchos largos de cadena diferentes en la macro-estructura.

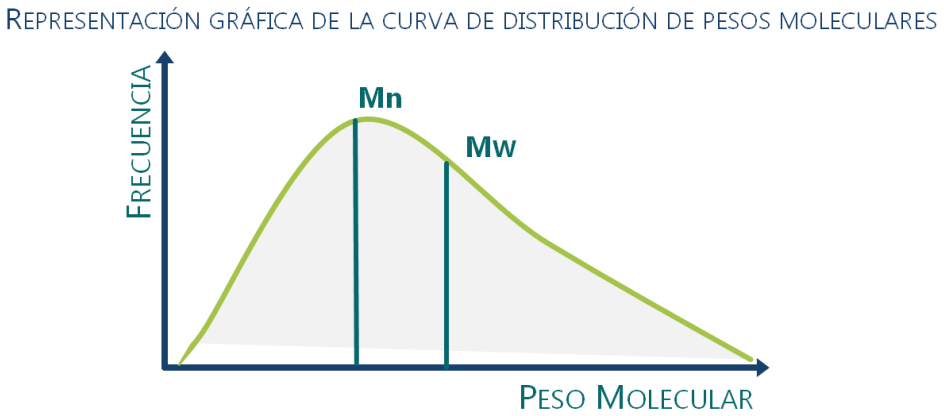


FIGURA 4. 3. REPRESENTACIÓN DE UNA CURVA DE DISTRIBUCIÓN DE PESOS MOLECULARES DE UN MATERIAL POLIMÉRICO Y SUS PESOS PROMEDIOS ASOCIADOS.

4.2.2. PROPIEDADES MECÁNICAS DE LOS POLÍMEROS

Las propiedades mecánicas de un material polimérico tienen mucha importancia desde un punto de vista técnico, ya que determinan su perfil de aplicación industrial. Estas propiedades hacen referencia al comportamiento del material al aplicarle una fuerza. El ensayo de tensión, es quizás el tipo más relevante de pruebas mecánicas que se pueden realizar en materiales poliméricos [Ward & Sweeney, 2004]. En esta tesis se trabajó con una base de datos de polímeros termoplásticos, por lo que todas las definiciones y consideraciones se enfocarán solo en ellos.

4.2.2.a. ENSAYO DE TENSIÓN

El ensayo de tensión o prueba de tracción es simple, relativamente económico y totalmente estandarizado, ya que la variabilidad de la temperatura y la velocidad a la que se realiza la prueba modifican los resultados. Estos ensayos proporcionan información sobre la rigidez, la resistencia, la tenacidad y la ductilidad de los polímeros. La Figura 4.4 (a) muestra una representación gráfica de un ensayo de tensión. Básicamente, el proceso consiste en someter a tensión axial, a velocidad constante, a una muestra con forma de hueso (*dog-bone shaped specimen*), hasta que presente fractura o falla. Esta velocidad recibe el nombre de CHS por las siglas en inglés para *cross-head speed*. Como resultado, se obtiene un perfil de tracción completo, es decir, una curva que describe el comportamiento del material respecto de las fuerzas aplicadas, en la Figura 4.4 (b) puede verse una representación gráfica de esta curva. Cada zona de la curva describe un tipo de comportamiento y se pueden obtener diferentes propiedades como: Módulo de Tensión, Elongación a la Rotura y Resistencia, entre otras.

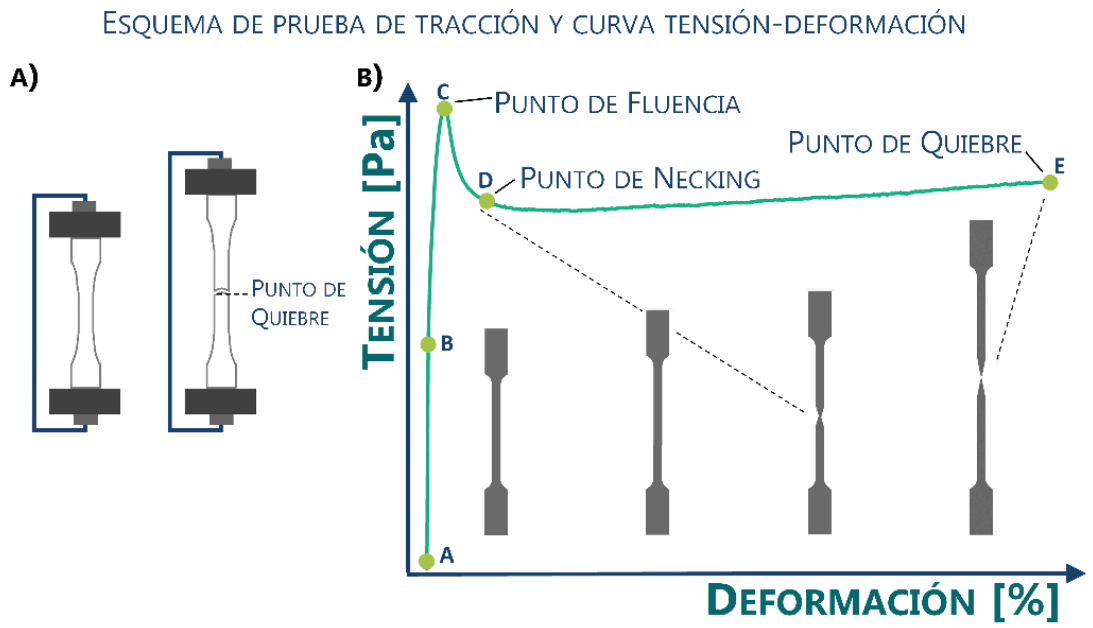


FIGURA 4. 4 A) ESQUEMA DE PRUEBA DE TRACCIÓN; **B)** CURVA TENSIÓN-DEFORMACIÓN DERIVADA DEL ENSAYO DE TRACCIÓN. ESTE ES UN EJEMPLO DE UNA POLIOLEFINA CON ALTA DUCTILIDAD.

MÓDULO DE TENSIÓN

El módulo de tensión, también conocido como Módulo Elástico o Módulo de Young es un indicador de la rigidez del material. En otras palabras, esta propiedad muestra la facilidad con que el polímero se deforma en condiciones elásticas, es decir, la deformación producida es recuperable o reversible. Se define como la relación entre el esfuerzo de tracción y la deformación en la zona de

comportamiento elástico que corresponde al primer tramo de la curva (puntos A y B en la Figura 4.4 B). Es decir, resulta de la pendiente de la curva (Figura 4.5) en la zona de comportamiento elástico donde al producirse una deformación en el material, este retoma la forma inicial [Seymour & Carraher, 1998; Ward & Sweeney, 2004].

REPRESENTACIÓN GRÁFICA DE LA MEDICIÓN DEL MÓDULO DE TENSIÓN

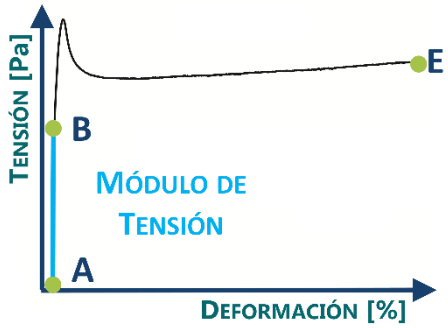


FIGURA 4. 5 REPRESENTACIÓN GRÁFICA DE LA MEDICIÓN DEL MÓDULO DE TENSIÓN DE UN MATERIAL POLIMÉRICO EN UNA CURVA DE TENSIÓN-DEFORMACIÓN.

ELONGACIÓN A LA ROTURA

La elongación a la rotura es un indicador del incremento de longitud que ha sufrido la probeta sometida a una fuerza de tracción hasta su falla. Está dada por la relación entre la variación de longitud inicial y la longitud después de la falla del material en el ensayo de tracción. Se mide entre dos puntos cuya posición está normalizada y se expresa en tanto por ciento. La elongación a la rotura para los materiales frágiles puede ser muy baja, tendiendo a cero; mientras que algunos termoplásticos tienden a presentar valores superiores al 100% indicando una mejor capacidad para manejar una carga sin fallas, pero con deformación (Figura 4.6) [Callister *et al.*, 2007].

REPRESENTACIÓN GRÁFICA DE LA MEDICIÓN DE LA ELONGACIÓN A LA ROTURA

FIGURA 4. 6 REPRESENTACIÓN GRÁFICA DE LA MEDICIÓN DE LA ELONGACIÓN A LA ROTURA DE UN MATERIAL POLIMÉRICO EN UNA CURVA DE TENSIÓN-DEFORMACIÓN.

Resistencia a la Rotura

La resistencia a la rotura se refiere al esfuerzo que soporta la probeta al momento de su falla (Figura 4.7). Cuando la tensión máxima se produce en el punto de fluencia, se conoce como resistencia a la tracción en la fluencia y cuando se produce la fractura, la tensión correspondiente se conoce como resistencia a la rotura o simplemente tensión a la rotura. Los materiales que tienen una baja resistencia a la rotura a menudo se denominan materiales débiles [Seymour & Carraher, 1998; Ward & Sweeney, 2004].

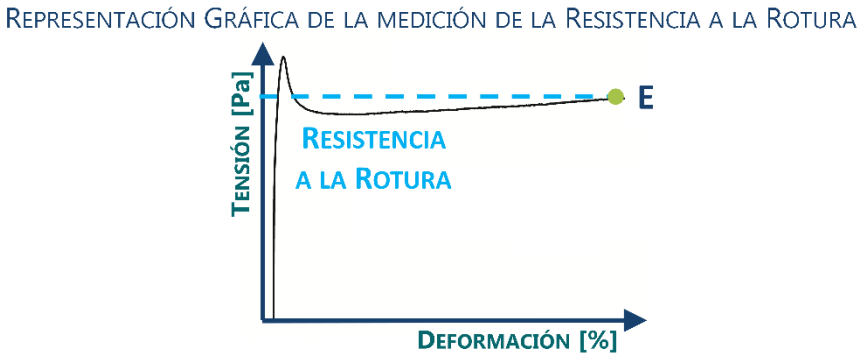


Figura 4. 7 Representación gráfica de la medición de la Resistencia a la Rotura de un material polimérico en una curva de tensión-deformación.

Comportamiento dúctil

El comportamiento dúctil hace referencia a la capacidad que presentan algunos materiales de deformarse plásticamente de manera sostenible sin romperse. A los materiales que presentan este comportamiento se los denomina dúctiles, y a aquellos que no, se los denomina frágiles o no dúctiles (Figura 4.8). No existe una medida numérica de ductilidad [Askeland *et al.*, 2011], no obstante, de esta tesis se desprende una propuesta para llegar a un valor tal que se pueda ubicar al material en una zona de comportamiento. Se propone relacionar los valores de Elongación a la Rotura y Módulo de Tensión (Comportamiento Dúctil = Elongación a la Rotura/Módulo de tensión). Para valores altos de esta relación se puede ubicar al material en una zona de comportamiento dúctil y viceversa con valores bajos.

FIGURA 4. 8 REPRESENTACIÓN GRÁFICA DEL COMPORTAMIENTO DÚCTIL DE UN MATERIAL POLIMÉRICO EN UNA CURVA DE TENSIÓN-DEFORMACIÓN.

4.3. BASE DE DATOS DE POLÍMEROS UTILIZADA

La curación de datos es un aspecto importante y, a menudo, descuidado al desarrollar colecciones de datos de polímeros. Varias propiedades de los polímeros dependen de factores que son independientes de la naturaleza química de las macromoléculas constituyentes, tales como los métodos y condiciones de medición del ensayo del que se obtienen las mediciones de esas propiedades [Adams & Murray-Rust, 2008]. La mayor parte de la información sobre polímeros está contenida en documentos no estructurados, por lo que obtener información útil es una tarea muy complicada y de difícil automatización. De aquí la complejidad en la recolección de información para armar una base de datos consistente y fidedigna [Adams, 2010].

Para esta tesis se utiliza la base de datos de 77 polímeros de alto peso molecular desarrollada por Palomba en su tesis doctoral en Ciencias y Tecnología de los Materiales defendida en marzo de 2014 [Palomba, 2014]. La construcción de esta base de datos se hizo mediante la recolección de datos de PoLyInfo[3], según estrictos parámetros y su curación posterior se hizo de forma manual.

La base de datos PoLyInfo proporciona sistemáticamente diversos datos necesarios para el diseño de materiales poliméricos. Su principal fuente es la literatura académica sobre polímeros. Contiene datos sobre homopolímeros, copolímeros, mezclas de polímeros y materiales compuestos. Incluye información sobre cerca de 100 propiedades (térmicas, eléctricas y mecánicas), estructuras químicas, métodos de procesamiento de muestras medidas, condiciones de medición, monómeros, métodos de polimerización y nombres IUPAC (*International Union of Pure and Applied Chemistry*) [Otsuka *et al.*, 2011].

[3] https://polymer.nims.go.jp/

La base de datos utilizada contiene 77 polímeros lineales de alto peso molecular. Se confeccionó especialmente para el estudio de las propiedades mecánicas derivadas de las pruebas de tracción (Módulo de Tensión, Elongación a la Rotura, Resistencia a la Rotura) para polímeros amorfos y termoestables. Las pruebas de tracción reportadas en la literatura se realizaron bajo los estándares ASTM D638, ASTM D882-83 y DIN 53504.53A. Además, todos los polímeros se ensayaron en su región vítrea (temperatura de prueba: 20-25 [°C]). Son 9 las familias químicas representadas: poliestirenos, polisulfuros, polivinilos, polióxidos/éteres/acetales, poliimidas/tioamidas poliamidas/tioamidas poliésteres/tioésteres, polisulfonas/sulfóxidos/sulfonatos/sulfoamidas y policetonas/tiocetonas [Palomba, 2014]. Los polímeros del conjunto de datos pueden pertenecer a más de una familia.

Las propiedades incluidas tienen los siguientes rangos de valores[4]: Mn = 4700-765000 [g/mol], Mw = 19500-2200000 [g/mol], IPD = 1.15-5.60, CHS = 1-100 [mm/min]. En cuanto al rango de valores de las propiedades que contempla para el Módulo de Tensión varían desde 0.1300 hasta 4.000 [GPa], y aunque pueda parecer pequeño desde un punto de vista matemático, los materiales exhiben comportamientos muy diferentes en los límites del mismo. Para la elongación a la rotura el rango es: 0.4 - 39.1 [%]; y para la resistencia a la rotura, es: 7.5 - 103 [MPa].

En la Tabla A4.1 del anexo se presentan, según datos extraídos de PoLyInfo, los nombres de los 77 polímeros, junto con el identificador numérico único (ID) que los nombrará de ahora en más, sus valores de Mn, Mw y CHS y las imágenes de estructura molecular 2D correspondientes a los monómeros que los conforman. Es importante notar que los primeros 7 polímeros, al igual que aquellos con ID 38 e ID 39 junto con ID 76 e ID 77 están compuestos por el mismo monómero respectivamente. Sin embargo, sus valores de pesos promedios varían. A continuación, en la Figura 4.9 se presenta el diseño de la tabla utilizando el polímero ID 10 como ejemplo. En la imagen del monómero se resalta en color verde las partes correspondientes a la cadena lateral del mismo y aquellas partes no resaltadas constituyen la cadena principal.

[4] En esta tesis se utiliza el punto como separador decimal y ningún grafema como separador de miles, según lo habilita la ISO 80000-1 del año 2009 para textos en español.

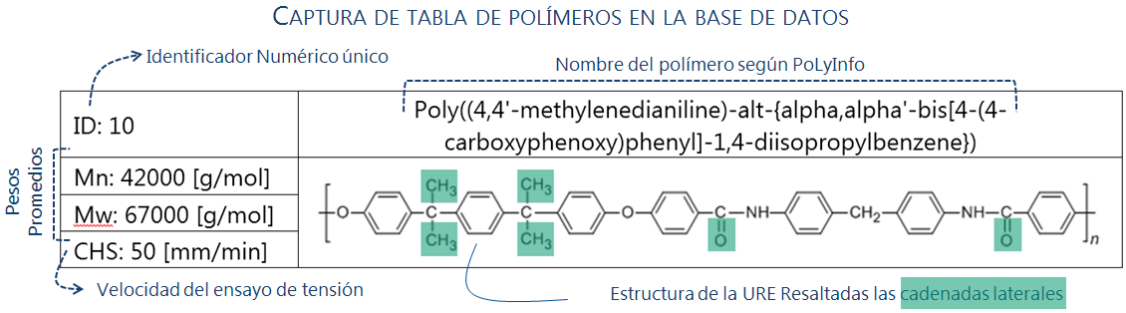


FIGURA 4. 9. CAPTURA DE IMAGEN DE LA TABLA DE LOS POLÍMEROS INCLUIDOS EN LA BASE DE DATOS ORIGINAL, TOMANDO COMO EJEMPLO EL POLIMERO CON ID 10. SE PRESENTAN DATOS DE: ID, NOMBRE, PESOS PROMEDIOS (Mn Y Mw), VELOCIDAD CHS Y UNA IMAGEN 2D DE LA URE, RESALTANDO EN LA ESTRUCTURA 2D LAS CADENAS LATERALES.

Se calcularon dos tipos de descriptores moleculares para las correspondientes 77 UREs: 1-Descriptores moleculares clásicos con DRAGON [Dragon, 2007] y 2-Descriptores Moleculares de visión Macro con PaDEL [Yap, 2011], utilizando la cadena principal y la cadena lateral de una URE central de la estructura trimérica. Una descripción más detallada de estos descriptores se incluye en el capítulo 3 (Informática Molecular). Resumiendo, la base de datos cuenta con 77 polímeros de alto peso molecular caracterizados por 998 descriptores moleculares clásicos y 51 descriptores moleculares de visión macro. A su vez, los polímeros están asociados a 4 propiedades mecánicas, tres de ellas se obtuvieron por medio de PoLyInfo, Elongación a la Rotura y Resistencia a la Rotura y la cuarta, es la propuesta surgida en esta tesis de comportamiento dúctil que se obtiene por la relación entre Elongación a la Rotura/Módulo de Tensión.

4.4. MODELOS QSPR CON DESCRIPTORES MOLECULARES VALUADOS EN LA URE

Existen dos tipos de modelos QSPR en la evolución de la predicción de propiedades de polímeros. Por un lado, están los métodos de contribución grupal [van Krevelen & Nijenhuis, 2009] que consideran las propiedades generales del polímero como la suma escalar de las propiedades de los grupos químicos contenidos en las moléculas que forman el polímero. Por el otro, se encuentran los basados en el uso de descriptores moleculares [Bicerano, 2002]. La mayoría de los estudios publicados utilizan modelos moleculares sintéticos, es decir, caracterizan el polímero a través de descriptores moleculares calculados sobre una única URE. La temperatura de transición vítrea (Tg) es una de las propiedades de polímeros amorfos más estudiadas, coincidentemente es para la que se cuenta mayor disponibilidad de datos para el desarrollo de modelos [Adams, 2010; Chen *et al.*, 2018].

Son varios los autores que han realizado modelos QSPR teóricos que utilizan descriptores moleculares basados en la URE o monómeros para predecir la Tg. Esta propiedad térmica está relacionada con el rendimiento mecánico y la procesabilidad del material. Katritzky aplicó el método CODESSA para predecir los valores de Tg para 22 homopolímeros y copolímeros lineales simples con poca diversidad estructural, y un conjunto de 238 descriptores moleculares [Katritzky *et al.*, 1996]. Luego, amplió su trabajo utilizando 88 homopolímeros lineales considerando el fragmento medio del trímero para el cálculo de descriptores. Una de las diferencias con el trabajo anterior es que el nuevo modelo está validado [Katritzky *et al.*, 1998].

Utilizando 84 polímeros de los 88 de la base de datos de Katritzky, García-Domenech y Julián-Ortiz desarrollaron un modelo QSPR de 10 parámetros generados a partir de índices teóricos de grafos basados en monómeros [García-Domenech & Julián-Ortiz, 2002]. Afantitis logró, para los mismos 84 polímeros y descriptores, un coeficiente de correlación mejorado mediante el uso de redes neuronales artificiales [Afantitis *et al.*, 2005]. Yu utilizó una red neuronal artificial de retropropagación (RPANN) para modelar los valores de Tg de tres tipos de polímeros de vinilo con cuatro descriptores obtenidos del monómero [Yu *et al.*, 2008]. Liu y Cao [Liu & Cao, 2009] utilizaron una red neuronal RPANN con cuatro descriptores cuánticos obtenidos a partir del monómero, para correlacionarlos con los valores de Tg de 113 poliacrilatos y poliestirenos.

Autores como los antes mencionados se dedicaron a probar que las redes neuronales muestran generalmente, para estos conjuntos, mejor desempeño y que los descriptores que codifican el tamaño de una cadena lateral son los más adecuados para la predicción de la Tg. Otros, en cambio, proponen nuevos descriptores para intentar capturar más eficientemente la información de la estructura molecular del polímero. Cao y Lin diseñaron cinco descriptores de la unidad repetitiva para expresar la rigidez de la cadena y las fuerzas intermoleculares de los polímeros, y los correlacionaron con los valores de Tg [Cao & Lin, 2003]. Schweizer y Curro desarrollaron descriptores para la unidad repetitiva fusionando descripciones de conectividad atómica, de valencia y de masa con un modelo de sitio de interacción [Schweizer & Curro, 1994]. Ulmer los utilizó para predecir la Tg en policarbonatos mediante redes neuronales y una estrategia de Validación Cruzada para su evaluación [Ulmer *et al.*, 1998].

Según Ulmer, estos descriptores de Schweizer y Curro que combinan las interacciones intermoleculares e intramoleculares describen el comportamiento tanto de la fase atomística como de la condensada de un polímero, convirtiéndolos

en mejores candidatos para construir modelos QSPR ya que los descriptores topológicos y teóricos tradicionales no proporcionan una forma óptima de representación de polímeros, y tienden a descuidar aspectos como el entrelazamiento físico [Ulmer *et al.*, 1998]. Por otro lado, Palomba, también al considerar insuficientes a los descriptores disponibles, presentó 26 nuevos descriptores para polímeros de alto peso molecular. Estos descriptores (en esta tesis llamados descriptores moleculares de visión macro) están basados en las características físicas, químicas, geométricas y electrónicas de las cadenas principales y laterales de los polímeros basados en la URE central de la estructura trimérica [Palomba *et al.*, 2012b].

Se encuentra en la literatura que se realizaron estudios QSPR de diversas propiedades de los polímeros como índice de refracción, temperatura de solución crítica más baja, viscosidad intrínseca, constante dieléctrica, factor de disipación dieléctrica, solubilidad, entre otras. Sin embargo, hay muy pocos trabajos relacionados con las propiedades mecánicas de los polímeros que han comenzado a estudiarse en los últimos años, a diferencia de la Tg que lleva más de tres décadas de investigación. Para los polímeros, estas propiedades tienen características particulares: dependen en gran medida de la temperatura del ensayo, la Tg y la escala de tiempo.

En las próximas subsecciones se presentan los primeros abordajes metodológicos para el modelado QSPR que se probaron en esta tesis, sin tratar la complejidad asociada a la polidispersión, aunque si fue considerado el IPD como descriptor adicional que comienza a dar información sobre este aspecto. Primero se presentan los modelos basados solo en URE que siguen una metodología denominada clásica y aquellos donde se incorpora la técnica de analítica visual. Finalmente se muestran los resultados obtenidos con una metodología denominada híbrida que no se corresponde con la combinación de las dos anteriores sino que resulta de una combinación de técnicas de Selección y Aprendizaje de características para la obtención de descriptores. La presentación de estas primeras experiencias se realiza en forma breve, pero en cada caso se citan las publicaciones donde está la información más detallada de estos enfoques.

4.4.1. METODOLOGÍA CLÁSICA

Hasta donde sabemos, el grupo de investigación en donde se desarrolló esta tesis ha sido uno de los pioneros en el modelado QSPR de propiedades mecánicas para polímeros de alto peso molecular [Audus & de Pablo, 2017; Peerless *et al.*, 2019]. Por lo tanto, hemos denominado metodología clásica a la seguida en el

trabajo que se publicó por primera vez para la predicción de la Elongación a la rotura [Palomba *et al.*, 2014a]. A continuación, se presentan dos trabajos de modelos QSPR para la predicción del Módulo de tensión y para la Resistencia a la rotura que desarrollamos siguiendo la metodología clásica, todos corresponden a modelos de regresión. Además, se presentan modelos de clasificación QSPR para la predicción del Comportamiento dúctil de un polímero. Cabe aclarar, que no todos los resultados que se presentan a continuación son producto de esta tesis, sino constituyen un antecedente directo de la misma que sirven para mostrar la evolución de las investigaciones, y esto será especificado en cada punto.

4.3.1.a. MODELOS QSPR DE REGRESIÓN

ELONGACIÓN A LA ROTURA

Como trabajo previo a la presente tesis, en nuestro grupo se desarrollaron modelos predictivos utilizando la base de datos de 77 polímeros amorfos de alto peso molecular especialmente desarrollada para el estudio de propiedades mecánicas (ver sección 4.3. Base de Datos Utilizada) donde Palomba infirió por primera vez un modelo QSPR entrenado por una red neuronal perceptrón multicapa (MLP) para la predicción de Elongación a la rotura [Palomba *et al.*, 2014a]. Los resultados se presentan en la Tabla 4.1. En este trabajo se caracterizó a la URE mediante los DMs provistos por Dragon. Además, se propusieron nuevos descriptores para representar mejor las características estructurales, y se incluyeron los parámetros experimentales. El modelo está formado por los siguientes tres DMs: velocidad del ensayo (CHS), cociente peso molecular promedio en número/área superficial de la cadena principal (Mn/SA_{MC}), y masa de la cadena principal normalizada (nM_{MC}). El modelo QSPR resultante tiene las ventajas de utilizar parámetros bien conocidos en el campo de los polímeros, así como de capturar las características estructurales de las cadenas principales y laterales.

RESISTENCIA A LA ROTURA

Siguiendo con los resultados previos a esta tesis, para esta propiedad se presentan resultados utilizando los mismos 77 polímeros de la base de datos. Toda la experimentación realizada en esta parte sirvió como entrenamiento en el uso de estas herramientas computacionales y como acercamiento al campo de la Informática de Polímeros. Entonces, siguiendo la misma metodología, se construyó el subconjunto de descriptores moleculares que fue entrenado con un perceptrón multicapa (MLP) 4-4-1 (capa de entrada, capa oculta y capa de salida) proporcionado por el software Statistica 8.0 [STATISTICA, 2007], con un esquema

de Validación Cruzada de 4 iteraciones (*4-fold cross-validation*). El conjunto final de DMs está formado por 2 DMs clásicos y 2 DMs de visión macro, el modelo QSPR inferido logró muy buen desempeño estadístico [Palomba *et al.*, 2014b]. Los DMs incluidos en el modelo son: peso molecular promedio en número (Mn), velocidad del ensayo (CHS), medida de contribución de átomos dadores de enlaces de hidrógeno (ETA_dEpsilon_D), cociente masa de la cadena principal/masa del grupo lateral (M_{MC}/M_{SC}). Los resultados se presentan en la Tabla 4.1.

MÓDULO DE TENSIÓN

Como parte del entrenamiento y trabajo precursor de los necesarios para cumplir los objetivos de esta tesis, se desarrolló también un modelo QSPR capaz de predecir el módulo de tensión y se realizó la selección de variables teniendo en cuenta todos los polímeros de la base de datos. En este caso los polímeros fueron 76, debido a que para uno (ID1) no se contaba con información sobre esta propiedad. El conjunto que finalmente fue elegido para entrenar el modelo QSPR está formado 5 descriptores moleculares: velocidad del ensayo (CHS), IDP (Mw/Mn), cociente número de enlaces rotacionales/número de enlaces múltiples (RBN/nBM), cociente área superficial de la cadena principal normalizada/área superficial del grupo lateral normalizada (nSA_{MC}/nSA_{SC}), y refractividad del grupo lateral (R_{SC}). Estos descriptores fueron elegidos aplicando la herramienta DELPHOS y además mediante la intervención de un experto se incorporaron manualmente DMs que por conocimientos fisicoquímicos previos se sabe que influyen en la propiedad. Luego, este conjunto de descriptores utilizando un esquema de Validación Cruzada de 4 iteraciones (*4-fold cross-validation*) se entrenaron con un perceptrón multicapa (MLP) 5-5-1 (capa de entrada, capa oculta y capa de salida) proporcionado por el software STATISTICA 8.0 [STATISTICA, 2007]. Finalmente, el modelo QSPR está constituido por 5 DMs (4 clásicos y 1 de visión macro) [Palomba *et al.*, 2014c]. Los resultados se presentan en la Tabla 4.1.

TABLA 4. 1. RESUMEN DE INDICADORES DE CALIDAD DE LOS MODELOS OBTENIDOS APLICANDO LA METODOLOGÍA CLÁSICA USANDO STATISTICA Y VALIDACIÓN CRUZADA DE 4 ITERACIONES.

Propiedad Objetivo	Cardinalidad	DMs	R^2	MAE
Elongación a la Rotura	3	CHS, Mn/SA_{MC} y nM_{MC}.	0.86	1.88
Resistencia a la Rotura	4	Mn, CHS, ETA_dEpsilon_D y M_{MC}/M_{SC}	0.85	8.4
Módulo de Tensión	5	CHS, IPD, RBN/nBM, R_{SC} y nSA_{MC}/nSA_{SC}	0.84	0.24

4.3.1.b. MODELOS QSPR DE CLASIFICACIÓN

COMPORTAMIENTO DÚCTIL

Si bien no hay una medida numérica de ductilidad, en 2017 propusimos como producción original de esta tesis, caracterizar el comportamiento dúctil de un polímero como la relación entre Elongación a la rotura y Módulo de Tensión [Martínez *et al.,* 2017]. En un extremo es común encontrar materiales con alto módulo y baja elongación ("frágiles") y en el otro extremo, los de bajo módulo y alta elongación ("dúctiles"). Basados en estos razonamientos definimos zonas de comportamiento como: "No-Dúctil" cuando la relación es menor a 2, "Dúctil" cuando es mayor a 5 y como "Indefinido" cuando se encuentra entre estos valores (Figura 4.10).

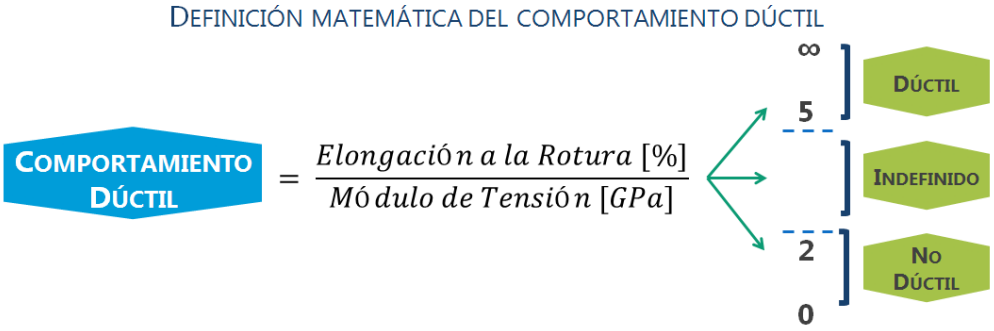


FIGURA 4. 10. DEFINICIÓN MATEMÁTICA DEL COMPORTAMIENTO DÚCTIL DE UN MATERIAL POLIMÉRICO, BASADO EN NUESTROS DATOS.

La base de datos de polímeros caracterizados a través del cálculo de sus descriptores moleculares con Dragon (DMs clásicos), fue fraccionada en tres conjuntos: conjunto de entrenamiento (50%), conjunto de validación interna (25%) y conjunto de validación externa (25%). Utilizando solo el conjunto de entrenamiento se realizó la Selección de características. De esta etapa se obtuvieron 5 subconjuntos diferentes de DMs: Perceptron Multicapa (NN), Naive Bayes (NB), *K-nearest neighbors* (KNN), *Random Committee* (RC) y *Random Forest* (RF). Se entrenaron 50 modelos QSPR con diferentes métodos de aprendizaje maquinal provistos por WEKA [Hall *et al.,* 2009]. A saber, Redes Neuronales: Mapas Auto-organizados (SOM) y Perceptron Multicapa (NN); Clasificadores Bayesianos: Naive Bayes (NB); Clasificadores Lazy: Kstar; Meta clasificadores: *Random Committee* (RC), *K-nearest neighbors* (KNN), LogitBoost y Bagging; Árboles de clasificación: *Random Forest* (RF) y HoeffdingTree (H-T) [Schustik *et al.*, 2019a]. Con el objetivo de obtener una mejora en el rendimiento, a aquellos modelos con más de un 80% de casos correctamente clasificados (%CC) y un error absoluto relativo (RAE) menor al 50%

para la Validación Interna, se les agregaron dos DMs de visión macro. Estos DMs son CHS (velocidad del ensayo de tensión) e IDP (índice de polidispersión), los cuales se agregaron de forma individual y conjunta en los modelos elegidos según el criterio descripto. Se pudo observar que ningún modelo mejoró al incluir IDP, pero al agregar CHS mejoraron algunos (Tabla 4.2). En general, los modelos del subconjunto original fueron los de mejor rendimiento global, por lo tanto, la inclusión manual de estos dos descriptores propuestos, no demostraron incrementar notoriamente el rendimiento.

TABLA 4. 2. RESUMEN DE LOS RESULTADOS PARA LOS MEJORES MODELOS EN LA VALIDACIÓN INTERNA, EN TÉRMINOS DE PRECISIÓN (%CC), ÍNDICE KAPPA QUE INDICA EL GRADO DE ACUERDO QUE EXISTE POR ENCIMA DEL ESPERADO POR AZAR (KAPPA) Y ERROR RELATIVO ABSOLUTO (RAE).

	Original			**+ CHS**			**+ IPD**			**+ CHS + IPD**		
Modelo	%CC	Kappa	RAE	%CC	Kappa	RAE	%CC	Kappa	RAE	%CC	Kappa	RAE
KNN-RF	80.7	0.67	45.7	80.7	0.67	43.2	77.2	0.61	49.9	80.7	0.67	49.5
NB-HT	**80.7**	**0.66**	**32.6**	80.7	0.66	31.7	80.7	0.66	34.3	80.7	0.66	33.6
NB-NN	80.7	0.65	44.1	78.9	0.61	46.6	75.4	0.57	51.4	77.2	0.60	46.4
NB-NB	**80.7**	**0.66**	**32.2**	**82.5**	**0.69**	**30.6**	78.9	0.63	34.1	**82.5**	**0.69**	**32.2**
NN-NN	**82.5**	**0.69**	**44.5**	71.9	0.52	52.2	75.4	0.57	47.6	68.4	0.46	59.6
NN-NB	80.7	0.66	46.3	78.9	0.64	43.4	75.4	0.57	43.0	78.9	0.64	44.8
RF-RF	80.7	0.67	45.7	80.7	0.67	43.2	77.2	0.61	49.9	80.7	0.67	49.5

Los modelos QSPR mejores desde el punto de vista estadístico fueron evaluados externamente. Se eligió el modelo NB-HT (Naive Bayes – Hoeffding Tree), el cual fue entrenado con *Hoeffding Tree* como método de aprendizaje maquinal y el subconjunto de DMs había sido previamente seleccionado por *Naive Bayes*. Este modelo es el de menor caída en la Validación Externa, sin presentar valores extraños. El modelo NB-NB también presenta métricas estadísticas esperables, que aseguran la no existencia de sobreajuste. Sin embargo, NN-NN tiene un comportamiento poco esperado ya que %CC (porcentaje de casos correctamente clasificados) es mayor en la Validación Externa (V.E.) que en la Validación Interna (V.I.). Los resultados de esta experimentación se presentan en la Tabla 4.3. Además, se realizaron experimentos de aleatorización (prueba de *Y-Randomization y FS-Randomization)* a al modelo NB-HT original. Este supera la prueba, es decir, el modelo tiene un rendimiento por encima del azar. El 99.6% de los experimentos (498/500) obtienen un rendimiento menor al 80.7 %CC del modelo elegido, que sólo es superado por el 0.40% (2/500) de los experimentos realizados.

TABLA 4. 3. RESUMEN DE LOS RESULTADOS PARA LOS MEJORES MODELOS EN LA VALIDACIÓN EXTERNA, EN TÉRMINOS DE PRECISIÓN (%CC), ÍNDICE KAPPA QUE INDICA EL GRADO DE ACUERDO QUE EXISTE POR ENCIMA DEL ESPERADO POR AZAR (KAPPA) Y ERROR RELATIVO ABSOLUTO (RAE).

	Original			**+ CHS**			**+ IPD**			**+ CHS + IPD**		
Modelo	%CC	Kappa	RAE	%CC	Kappa	RAE	%CC	Kappa	RAE	%CC	Kappa	RAE
NB-HT	**78.9**	**0.59**	**32.0**									
NB-NB	78.9	0.60	35.0	78.9	0.60					78.9	0.60	34.9
NN-NN	84.2	0.71	37.6									

Los modelos con respuestas categóricas resultan, a veces, más útiles que los modelos que relacionan DMs con valores específicos de una propiedad (regresión). Este trabajo presenta un modelo QSPR robusto que permite predecir el comportamiento dúctil para un nuevo polímero y clasificarlo como: Dúctil, Indefinido, No-Dúctil.

4.3.1.c. HERRAMIENTA POLYPP

La predicción *in silico,* o testeo virtual, de los materiales poliméricos en las etapas tempranas de diseño, tiene como ventaja obtener un perfil estimado de propiedades del material (existente o hipotético) sin haberlo sintetizado. PolyPP (*Polymer Property Predictor*) es una herramienta computacional pensada y desarrollada en esta tesis para operar en la web, capaz de predecir propiedades derivadas del ensayo de tensión (Módulo de Tensión, Elongación a la Rotura y Resistencia a la rotura) y además clasificar al material según su Comportamiento Dúctil [Cravero *et al.*, 2017b]. PolyPP permite al modelador de materiales descartar en las primeras etapas de su trabajo, candidatos poco viables o aquellos que no se ajusten al perfil de propiedades buscadas. El usuario debe ingresar como datos de entrada: la URE en código SMILES y los pesos moleculares promedio que desearía que tuviera su polímero. Además, debe elegir una velocidad de ensayo (CHS) a la cual virtualmente se testeará el material (baja, media o alta). En la Figura 4.11 puede verse una captura de imagen de PolyPP y el flujo de trabajo seguido por esta herramienta. Una vez completados todos los campos, se habilita el botón *predecir* y el usuario debe presionarlo, para ejecutar la predicción y posteriormente visualizar y descargar los resultados si así lo desea. Como salida, PolyPP brinda una tabla con los valores predichos de las propiedades y la clasificación correspondiente.

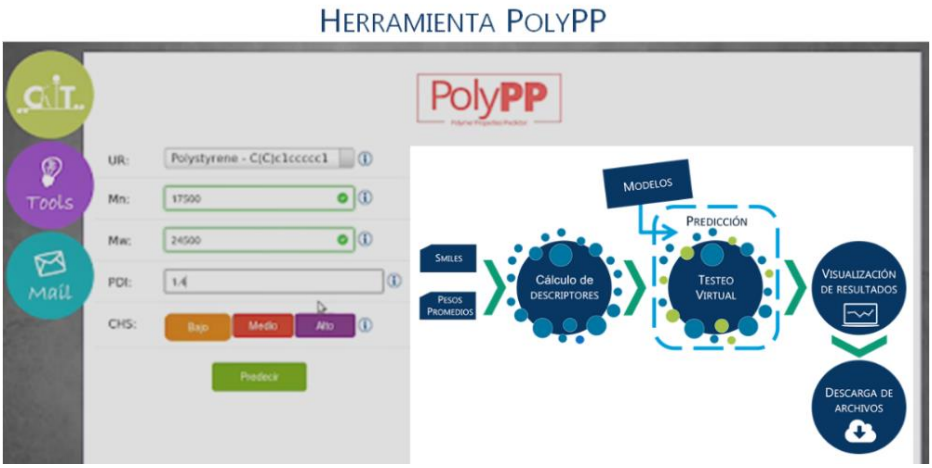


FIGURA 4. 11 CAPTURA DE PANTALLA DE LA HERRAMIENTA POLYPP Y FLUJO DE TRABAJO QUE SE SIGUE.

En PolyPP se engloban los modelos QSPR desarrollados bajo la metodología clásica, con algunas diferencias como un tratamiento de datos más riguroso, ya que en esta oportunidad se reservan 18 polímeros para ser usados en la Validación Externa. Los modelos QSPR fueron inferidos usando diferentes técnicas de aprendizaje maquinal provistas por Weka. En particular, para Elongación y Resistencia a la Rotura los rendimientos más altos fueron conseguidos con *Random Forest*, en cambio para Módulo de Tensión, *Linear Regression* con siguió el mejor rendimiento. En la Tabla 4.4 se presentan las métricas de rendimiento estadístico para la Validación Externa para cada una de estas tres propiedades, en términos de coeficiente de correlación (R^2) y Error Porcentual Absoluto Medio (MAPE). Las métricas no son presentadas para el Comportamiento Dúctil, ya que en este caso dicha predicción en la categorización del polímero testeado se hace utilizando la fórmula presentada en 4.3.1.b.

TABLA 4. 4 RESULTADOS OBTENIDOS PARA LA VALIDACIÓN EXTERNA DE LOS MODELOS QSPR DE POLYPP.

	Cardinalidad	**R^2**	**MAPE**
Elongación a la Rotura	4	0.72	31.66%
Resistencia a la Rotura	7	0.83	12.45%0
Módulo de Tensión	10	0.82	33.49%

4.3.1.d. CONCLUSIONES SOBRE LA METODOLOGÍA CLÁSICA

Concluyendo, la metodología clásica permitió comenzar a tratar los polímeros con las mismas técnicas QSPR desarrolladas para fármacos o moléculas pequeñas. Como adecuación se utilizaron descriptores moleculares de visión macro, producidos en la tesis de Palomba, especialmente diseñados para caracterizar las estructuras de las cadenas principales y laterales de los polímeros. Los modelos de

la sección 4.3.1.a., fueron entrenados con redes neuronales y seleccionado por la misma heurística. A estos DMs se les agregaron de forma manual DMs de visión macro por intervención de un especialista, logrando en general rendimientos estadísticos aceptables. Con esta metodología denominada Clásica se pudo constatar que el manejo de datos resultó muy laborioso y propenso a errores, lo que derivó en una inversión de tiempo y sucesivos controles que hemos considerado excesivo.

Adicionalmente, la herramienta PolyPP es única en su tipo, según nuestro conocimiento, siendo capaz de predecir las cuatro propiedades de polímeros estudiadas en esta tesis a través del modelado QSPR empleando una interfaz gráfica. Los modelos QSPR utilizados por PolyPP difieren de los anteriores, inferidos siguiendo la Metodología clásica propuesta por Palomba, por dos puntos clave: la división de los datos y los métodos utilizados. En cuanto a la rigurosa división de los datos, se reserva un subconjunto de polímeros para realizar la Validación Externa, y, en lo referido a los métodos de aprendizaje maquinal, durante el proceso de entrenamiento de los modelos QSPR se utiliza un esquema de Validación Cruzada Dejando Uno Fuera (*Leave One Out Cross Validation*) y diferentes métodos de aprendizaje maquinal provistos por Weka utilizando la configuración de parámetros por defecto. Estos cambios además de representar mejoras metodológicas básicas, disminuyen la posibilidad de sobreajuste. Cuestión que era muy evidente en los modelos anteriores.

Finalmente, la propuesta de un modelo QSPR de clasificación capaz de describir el Comportamiento Dúctil de un material polimérico en tres clases resultó un instrumento útil para el modelador de polímeros en el diseño de nuevos materiales con perfiles específicos. Muchas veces es más relevante poder categorizar el comportamiento de un material que relacionarlo a un valor numérico específico de propiedad. En este sentido, la propuesta de la fórmula que relaciona la Elongación a la Rotura con el Módulo de Tensión es innovadora y por ende, en el futuro podría pensarse en otros rangos de valores para la categorización cuando las bases de datos crezcan en número y tipo de familias de polímeros representadas.

4.4.2. METODOLOGÍA QUE INCLUYE ANALÍTICA VISUAL

La metodología clásica tratada en el inciso anterior, presentaba cierta dificultad en el análisis de los múltiples subconjuntos derivados de DELPHOS, ya que se deben ir tomando decisiones para llegar al mejor modelo. Como consecuencia se buscó una alternativa metodológica que incluyera la Analítica Visual, constituyendo un avance con respecto a la metodología clásica (Figura 4.12).

Específicamente, la diferencia reside en la facilidad para descubrir, de manera visual, relaciones entre descriptores de los subconjuntos, y por lo tanto el experto en materiales puede interactuar de manera más eficiente, evitando errores de traspaso de datos en forma manual, como se hacía antes [Palomba *et al.*, 2014d].

FIGURA 4. 12. METODOLOGÍA QUE INCLUYE ANALÍTICA VISUAL EN LA SELECCIÓN DE LOS DESCRIPTORES MOLECULARES PARA INFERIR MODELOS QSPR.

Dado que DELPHOS no prueba la combinatoria completa de posibles subconjuntos, es factible introducir algunos DMs de visión macro relevantes para la propiedad, y aun considerarse correcto. Esa introducción de descriptores al subconjunto se hace a través de VIDEAN [Martínez *et al.*, 2015], la herramienta desarrollada en nuestro grupo de investigación, basada en el análisis visual interactivo. VIDEAN, surgió como una necesidad en el modelado QSPR de polímeros, donde se tenían más de un tipo de descriptores y se buscaba llegar a modelos interpretables y de baja cardinalidad, lidiando con la imposibilidad de contrastar con la literatura por falta de datos [Cravero *et al.*, 2015c]. A continuación, debido a la enorme cantidad de tablas, gráficas y capturas de pantalla que debieran presentarse con todos los datos que derivan de este tema, se decidió mostrar un breve resumen de los resultados de la predicción de propiedades mecánicas del ensayo de tensión siguiendo la metodología con Analítica Visual para el desarrollo de modelos QSPR. De este modo, el lector podrá tener una idea general de toda la experimentación realizada y las conclusiones de manera rápida, con la posibilidad de ver todos los detalles en las publicaciones existentes.

4.4.2.a. MODELOS QSPR DE REGRESIÓN

ELONGACIÓN A LA ROTURA

Este es uno de los primeros trabajos donde se testearon las ventajas de usar VIDEAN en el modelado QSPR. Se trabajó con la base de datos de 77 polímeros de alto peso molecular, caracterizados por 998 DMs clásicos y 51 DMs de visión macro. El objetivo fue realizar un análisis visual del modelo de 3 descriptores obtenido usando la metodología clásica para predecir la Elongación a la rotura: velocidad del ensayo (CHS), cociente entre peso molecular promedio en número/área superficial de la cadena principal (Mn/SA_{MC}), y masa de la cadena principal normalizada (nM_{MC}) [Palomba *et al.*, 2014a]. Para este análisis se usó un esquema de Validación Cruzada de 10 iteraciones (*10-fold cross-validation*) y las tres técnicas de Aprendizaje Maquinal brindadas por VIDEAN a través de WEKA utilizando todos los hiper-parámetros por default, lo que permite la reproducibilidad: Regresión Lineal (LR), Árboles de Decisión (M5P) y Redes Neuronales (Perceptron Multicapa) [Martinez *et al.*, 2014a]. Como conclusión, se pudo constatar que la analítica visual resultó ser efectiva para una rápida evaluación de los subconjuntos de candidato. Además, el modelo QSPR logró un rendimiento aceptable de R^2=0.69 que, aunque no superó el de la metodología clásica R^2=0.88, se considera un modelo robusto, libre de sobreajuste, interpretable en términos fisicoquímicos, totalmente reproducible y de baja cardinalidad. En contraste con la metodología clásica, que siendo el primer desarrollo no cumplía con todas estas cualidades, en especial la ausencia de sobreajuste.

RESISTENCIA A LA ROTURA

Para la predicción de la Resistencia a la Rotura se decidió utilizar los 10 subconjuntos de DMs que retornó DELPHOS como los mejores candidatos para ser entrenados por el método de aprendizaje maquinal [Cravero *et al.*, 2015a] derivados de la misma base de datos de 77 polímeros, siendo este trabajo completamente de esta tesis. Además, con el objetivo de realizar una comparación, se incluyó el subconjunto de descriptores a partir del cual se infirió el mejor modelo QSPR con la metodología clásica (Sección 4.3.1.a) para el cual se había reportado un rendimiento de R^2=0.85 [Palomba *et al.*, 2014b]. Cada subconjunto fue intervenido por un experto agregando descriptores relevantes para la propiedad y sustrayendo aquellos que no aportaban información adicional, es decir, que estaban muy correlacionados entre sí. Para el entrenamiento de los modelos QSPR se usó un esquema de Validación Cruzada de 10 iteraciones (*10-fold cross-validation*) y las tres técnicas de Aprendizaje Maquinal brindadas por VIDEAN a

través de WEKA utilizando todos los hiper-parámetros por default, que como se mencionó anteriormente permiten la reproducibilidad del experimento: Regresión Lineal (LR), Árboles de Decisión (M5P) y Redes Neuronales (Perceptron Multicapa). Finalmente, el modelo elegido quedó conformado por 5 descriptores (CHS, PDI, M_{SC}, JGI5, y ETA_bBetaP) con un R^2=0.86. En este caso el rendimiento alcanzado por el mejor modelo QSPR de cada metodología, clásica y usando analítica visual, es similar y la cardinalidad en ambos es 5. Concluyendo, el uso de la analítica visual aceleró los tiempos de análisis de los subconjuntos de DMs candidatos, lo que permitió evaluar de manera metódica mayor cantidad de posibles subconjuntos de DMs, y además evitó errores por tratamiento manual de datos.

MÓDULO DE TENSIÓN

Siguiendo con el mismo tipo de experimentación, como parte de esta tesis se utilizaron los 10 mejores subconjuntos de DMs generados por DELPHOS con el objetivo de inferir modelos QSPR para la predicción del Módulo de Tensión [Cravero *et al.*, 2015b]. Además, se incluyó en el análisis con VIDEAN el subconjunto de DMs a partir del cual se infirió el mejor modelo QSPR siguiendo la metodología clásica (Sección 4.3.1.a) con un R^2=0.84 con cardinalidad 5. También se usó un esquema de Validación Cruzada de 10 iteraciones (*10-fold cross-validation*) y las tres técnicas de Aprendizaje Maquinal brindadas por VIDEAN a través de WEKA utilizando todos los hiper-parámetros por default: Regresión Lineal (LR), Árboles de Decisión (M5P) y Redes Neuronales (Perceptron Multicapa), para el entrenamiento de los modelos QSPR. El subconjunto de DMs que mejor rendimiento había obtenido con la metodología clásica según estos nuevos parámetros de experimentación decae desde R^2=0.84 hasta un R^2=0.61 lo cual denota el sobreajuste ocasionado por las redes neuronales que se habían aplicado. El mejor modelo QSPR obtenido por la metodología que utiliza analítica visual (cardinalidad 5) alcanza un R^2=0.69. La diferencia de rendimientos entre estos modelos genera la necesidad de una nueva y cuidadosa experimentación para lograr inferir modelos QSPR de predicción del Módulo de Tensión robustos, confiables y con mejores resultados en cuanto a precisión estadística.

Para lograr este último objetivo, se utilizó un criterio más riguroso de división de los datos, el cual reserva un subconjunto de polímeros para ser usados como Validación Externa. Se aplicaron diferentes procesos esta última etapa para comprobar la robustez de los modelos inferidos, para descartar modelos que *a priori* pueden parecer buenos candidatos. Además, del uso de DELPHOS para seleccionar 50 subconjuntos de DMs alternativos que fueron reducidos a 10 según

su MAE (se busca menor error), se empleó VIDEAN para descartar aquellos que aun teniendo bajo error no reunían los rasgos necesarios para ser buenos candidatos (baja información mutua entre DMs). Luego, los 5 subconjuntos resultantes fueron entrenados, usando un esquema de Validación Cruzada Dejando Uno Fuera (*Leave One Out Cross Validation),* para inferir los modelos QSPR con cuatro técnicas de aprendizaje maquinal provistas por WEKA, con parámetros por defecto,: *Linear Regression* (LR), *Neural Networks* (NN), *Random Forest* (RF) y *Random Committee* (RC). Los 20 modelos QSPR obtenidos en esta etapa fueron seleccionados solo aquellos con un $R^2 \geq 0.75$ con un $RAE > 0.5$ para ingresar a la etapa de Validación Externa. A los modelos con mejor desempeño se les aplicó el método de Dominio de Aplicabilidad propuesto por Soto [Soto *et al.,* 2009b] utilizando la partición de datos reservada para la Validación Externa. Un resumen de todos los pasos seguidos en la metodología y los tipos de métricas estadísticas computadas en cada paso se muestran en la Figura 4.13.

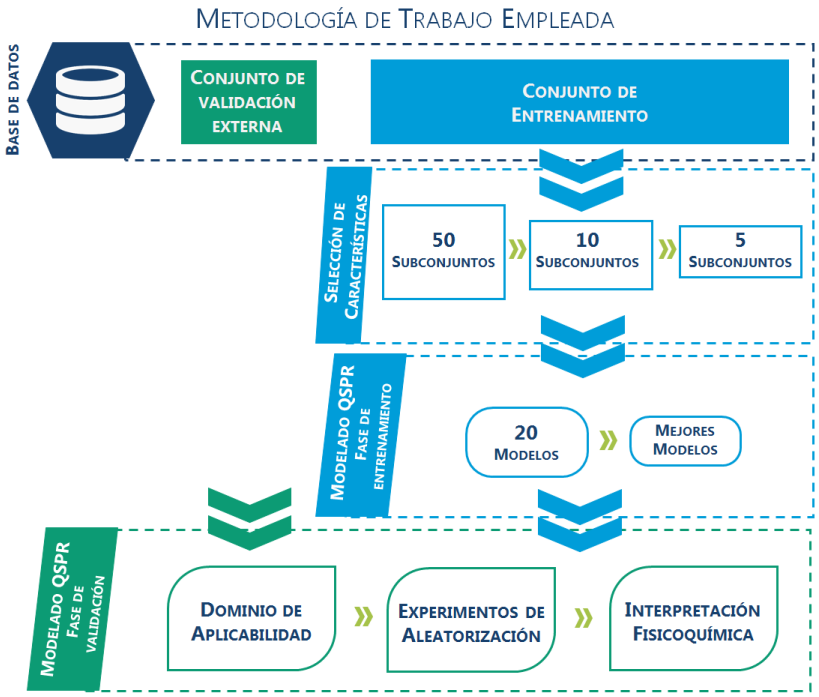


FIGURA 4. 13. RESUMEN DE LA METODOLOGÍA EMPLEADA EN EL MODELADO QSPR PARA PREDICCIÓN DEL MÓDULO DE TENSIÓN CON MEJORAS.

A los modelos QSPR con mejor desempeño en la Validación Externa (R^2 = 0.73 y R^2 = 0.75) se les aplicaron las técnicas de aleatorización, *Y-Randomization* y *FS-Randomization,* para descartar que las correlaciones obtenidas sean por azar (*by chance*) (Definiciones en Cap. 2). Ambos modelos superaron las dos pruebas, por lo que no se descarta ninguno ya que sus rendimientos estadísticos son muy similares y tampoco permiten tomar una decisión. Como resultados de este trabajo se

obtuvieron dos modelos QSPR para predecir el Módulo de Tensión siguiendo una metodología robusta y aunque no pudieron ser comparados trabajos ajenos en la bibliografía debido a la ausencia de antecedentes, frente a nuestros esfuerzos previos se logró un mejor rendimiento estadístico a la vez que se siguió una experimentación más rigurosa [Cravero *et al.*, 2019a].

4.4.2.b. MODELOS QSPR DE CLASIFICACIÓN

Como parte de esta tesis, a partir de la base de datos de 77 polímeros se obtuvieron tres subconjuntos alternativos de descriptores moleculares relevantes, uno de ellos obtenido por la herramienta Weka, otro construido con VIDEAN por un experto y el tercero construido a partir de los dos anteriores utilizando todos los análisis que VIDEAN permite realizar [Martínez *et al.,* 2017]. A partir de esta experimentación, concluimos que el DM de visión macro CHS que es un parámetro experimental de la prueba de tracción, juega un papel central en todos los modelos. Finalmente, los modelos QSPR de clasificación que permiten categorizar un nuevo material virtual como dúctil, no-dúctil o indefinido se infirieron mediante tres diferentes métodos de aprendizaje automático: *Neural Network* (NN), *Random Forest* (RF) y *Random Committee* (RC). El modelo con el rendimiento más alto clasifica correctamente el 88.46% (% CC) de polímeros y tiene un área de curva de ROC (Característica Operativa del Receptor) de ROC= 0.97.

4.4.2.c. CONCLUSIONES SOBRE LA METODOLOGÍA QUE INCLUYE ANALÍTICA VISUAL

Esta metodología constituyó una ayuda para eliminar el sobreajuste que presentaban las primeras experimentaciones, realizadas con Statistica, de la línea de investigación (Metodología Clásica). También, hubo un relevante aporte en el tratamiento de datos libre de error de manipulación, así como en la facilidad y rapidez para visualizar los patrones de las relaciones buscadas entre descriptores y el contenido de información de estos. Esto se debe a que VIDEAN tiene una interfaz que brinda diferentes tipos de grafos, con aristas de diferentes colores que se relacionan con grados de información mutua, correlación lineal y co-ocurrencia de descriptores entre modelos alternativos, por ejemplo. También brinda diferentes tamaños de nodos, gráficos de dispersión e histogramas que relacionan valor de propiedad con valor de descriptor. Finalmente, todas estas facilidades para la interpretación y análisis visual condujeron a una importante reducción de tiempo en la inferencia de modelos QSPR robustos, de baja cardinalidad e interpretables fisicoquímicamente.

4.4.3. METODOLOGÍA HÍBRIDA

Hasta el momento se presentaron resultados de metodologías donde se aplicaba Selección de Características, y en esta sección se muestran experimentos con una estrategia llamada Metodología Híbrida. Esta metodología fue presentada en detalle en el capítulo anterior (Capítulo 3: Sección 3.4.3. Metodología Híbrida para la obtención de descriptores), su esencia es la combinación de las variables obtenidas por procesos de Selección y Aprendizaje de Características para potenciar la información que brinda cada tipo de variable en la inferencia de modelos QSPR. Las dos ramas de la Metodología Híbrida son: DRAGON-DELPHOS (D-D) para la Selección de Características y CODES-TSAR (C-T) para el Aprendizaje de Características. DRAGON es una herramienta de cálculo de descriptores y DELPHOS una herramienta de selección de DMs, esta dupla fue ampliamente utilizada en los experimentos presentados hasta el momento. CODES procesa los SMILES de cada molécula de la base de datos y devuelve como resultado una matriz dinámica por cada una. La dimensión de estas matrices es n x m, donde n es el número de átomos de cada molécula sin hidrógenos y m es el número de iteraciones necesarias para lograr la convergencia en el proceso de entrenamiento. Cada matriz dinámica generada representa la descripción estructural de la molécula [Dorronsoro *et al.*, 2004]. A continuación, la metodología TSAR es aplicada con el fin de obtener un conjunto reducido de descriptores moleculares a partir de estas matrices. Este método utiliza un algoritmo auto-codificador (*auto-encoder*) [Livingstone *et al.*, 1991], con la premisa que si el patrón de entrada (capa de entrada) se reproduce en la capa de salida de la red neuronal utilizada, la capa intermedia 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. Cabe notar que estos descriptores no pueden ser interpretados fisicoquímicamente, a diferencia de los obtenidos usando DRAGON.

Esta Metodología Híbrida fue probada para bases de datos de compuestos químicos relacionados con estudios farmacológicos y se obtuvieron modelos QSAR con altos niveles de precisión [Ponzoni *et al.*, 2017]. Sin embargo, su rendimiento no había sido evaluado en la inferencia de modelos QSPR para diseñar polímeros sintéticos. De esta manera surge lo que denominamos Primera Pregunta de Investigación de esta tesis: *¿Puede el aprendizaje de características empleadas en enfoques QSAR resultar de utilidad en el contexto de Informática de Polímeros? ¿Qué sucede si combinamos las ventajas del aprendizaje de características con las de la selección de características para la inferencia de modelos QSPR en polímeros*

de alto peso molecular? Para investigar estas hipótesis se experimentó con la predicción de Resistencia a la Rotura, ya que había sido la propiedad con mejores rendimientos estadísticos empleando las metodologías anteriores.

4.4.3.a. MODELOS QSPR DE REGRESIÓN

Los descriptores derivados de un proceso de Aprendizaje de Características constituyen un nuevo espacio de variables. Particularmente, los derivados de C-T codifican información derivada de toda la estructura molecular de los compuestos. El objetivo de esta sección fue explorar las ventajas y los inconvenientes de este enfoque en la predicción de propiedades mecánicas relevantes para la Ciencia de los Materiales a través de modelos QSPR de regresión, específicamente la Resistencia a la Rotura [Cravero *et al.*, 2016a]. El punto de partida de esta metodología fue codificar los monómeros de los polímeros de la base de datos utilizando el sistema de código SMILES. Luego, utilizando la dupla de softwares C-T, con dos arquitecturas de redes diferentes de 2 y 3 neuronas en la capa intermedia, se obtuvieron 2 y 3 descriptores respectivamente. Se trabajó solo con 66 polímeros de la base de datos original, debido a la dificultad encontrada en el cálculo de la matriz dinámica que realiza CODES que impidió completar el cálculo en 11 moléculas [Cravero *et al.*, 2015d].

Como primer paso, solo los descriptores obtenidos por C-T fueron usados para inferir modelos QSPR para predecir Resistencia a la Rotura y los resultados mostraron que la información reunida por ellos no fue suficiente para describir la propiedad estudiada, alcanzando en las diferentes arquitecturas (número de neuronas en la capa intermedia) para la Validación Externa: R^2 = 0.58 (C-T N2) y R^2 = 0.44 (C-T N3). Esto motivó el uso de los DMs correspondientes al mejor modelo, modelo de referencia (MR), obtenido mediante la Metodología que incluye Analítica Visual (Sección 4.4.2.a) para el enriquecimiento de los modelos C-T al incorporarle los DMS correspondientes a MR (Tabla 4.1). Se plantean dos estrategias de composición para los nuevos subconjuntos de descriptores, por un lado está el enriquecimiento que incluye los descriptores CHS y Mn a los subconjuntos C-T N2 y C-TN3. Por el otro lado se encuentra la combinación de todos los descriptores de ambas ramas metodológicas, es decir, agregar todos los DMs del MR primero a C-T N2 y luego a C-TN3 (Figura 4.14).

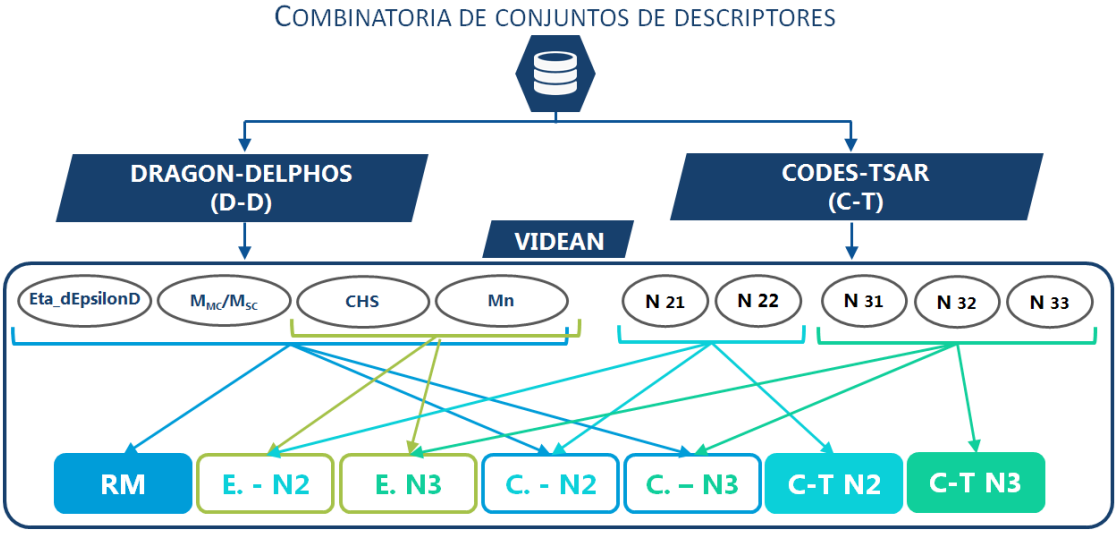


FIGURA 4. 14. COMBINATORIA DE CONJUNTOS DE DESCRIPTORES PROVENIENTES DE TÉCNICAS DE SELECCIÓN DE CARACTERÍSTICAS Y APRENDIZAJE DE CARACTERÍSTICAS.

Un resumen de los mejores resultados para la Validación Externa se presenta en la Tabla 4. 5, en término de R^2 y cardinalidad. A partir de estos experimentos, se concluyó que los descriptores extraídos por C-T proporcionan información complementaria al MR, aunque la cardinalidad final sube y el modelo se vuelve no interpretable en términos fisicoquímicos [Cravero *et al.*, 2016b].

TABLA 4. 5 RESUMEN DE LOS RESULTADOS OBTENIDOS POR EL MÉTODO C-T Y LA COMBINACIÓN DERIVADA DE LA METODOLOGÍA HIBRIDA

Modelos	Cardinalitdad	Validación
C-T N2	2	0.58
C-T N3	3	0.44
RM: Mn + CHS + Eta_dEpsilon_D + M_{MC}/M_{SC}	4	0.82
Enriquecido N2: C-T N2 + Mn + CHS	4	0.75
Enriquecido N3: C-T N3 + Mn + CHS	5	0.59
Combinado N2: C-T N2 + RM	6	**0.85**
Combinado N3: C-T N3 + RM	7	**0.85**

Frente a estos resultados, para llegar a una mejor compresión de los modelos QSPR generados por la Metodología Híbrida, se planteó una actualización metodológica definiendo un enfoque de trabajo integrador con herramientas de Analítica Visual [Cravero *et al.*, 2015e]. Se propone analizar los modelos QSPR obtenidos mediante el entrenamiento de los conjuntos de DMs provenientes de C-T y de D-D (MR) y las combinaciones surgidas entre los descriptores de ambas técnicas. Se muestra en la Figura 4.15 los grafos, obtenidos con VIDEAN, de las combinaciones surgidas entre los descriptores de ambas técnicas. Estos grafos representan la información mutua existente entre un par de descriptores (nodos) a

través del tono en que se grafica la arista. Los tonos púrpura (mayor tonalidad) de las aristas entre descriptores indican que ese par de descriptores comparten en mayor grado información, en cambio tonos rosados claros (menor tonalidad) indican que comparten en menor grado de información. Por lo tanto, se buscan grafos visualmente con tonos rosa claro, es decir, se pretende elegir subconjuntos de DMs con baja información mutua. Como conclusión de este exhaustivo análisis se observó nuevamente que la información química resumida por los descriptores C-T no es suficiente para describir la propiedad mecánica. Sin embargo, estos DMs proporcionan información potencialmente significativa y complementaria cuando se combinan con un modelo de referencia, como se puede ver en las Figura 4. 15 y Figura 4.16. Además, se probó la precisión de los modelos QSPR obtenidos mediante la combinación de todos los descriptores aprendidos por CODES-TSAR con todos los del MR, no tuvo un aumento significativo con respecto a MR [Cravero *et al.*, 2016a].

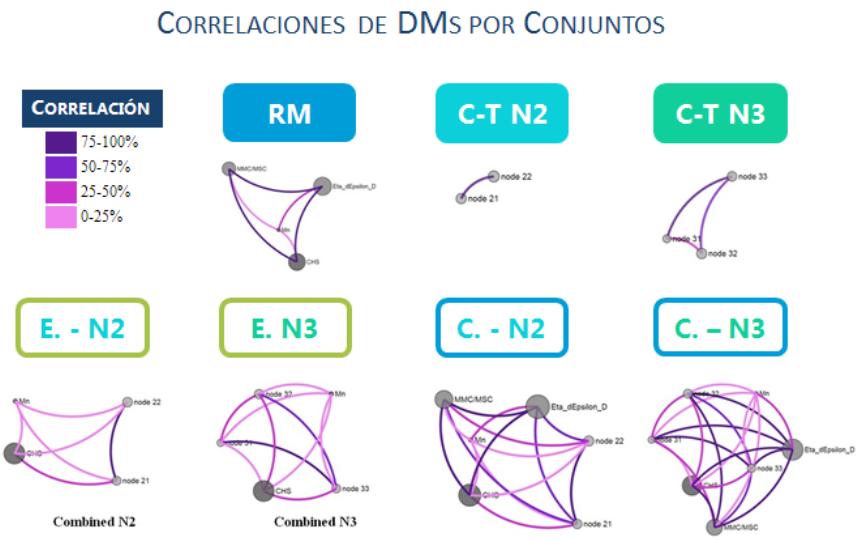


FIGURA 4. 15. VISUALIZACIÓN CON VIDEAN PARA LA CORRELACIÓN ENTRE DESCRIPTORES DE LOS MODELOS GENERADOS A PARTIR DE TÉCNICAS DE SELECCIÓN Y APRENDIZAJE DE CARACTERÍSTICAS.

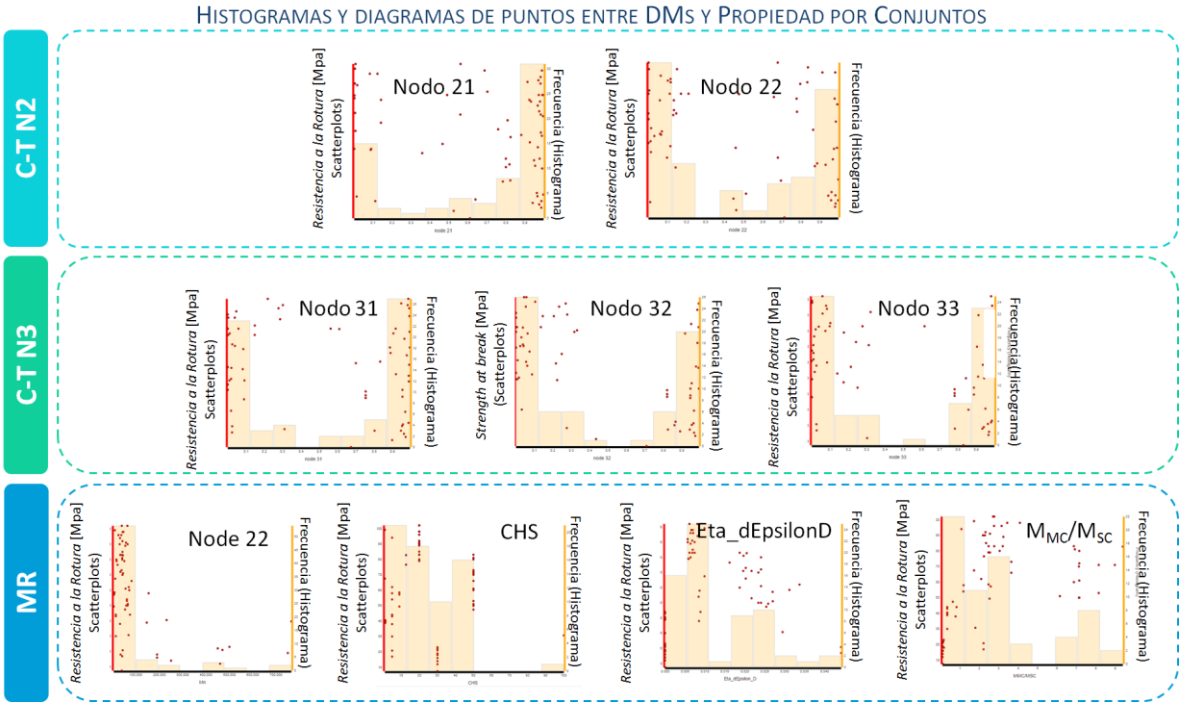


FIGURA 4. 16 VISUALIZACIÓN CON VIDEAN PARA LA DISPERSIÓN DE PUNTOS E HISTOGRAMAS ENTRE DESCRIPTORES Y PROPIEDAD DE LOS MODELOS GENERADOS A PARTIR DE TÉCNICAS DE SELECCIÓN Y APRENDIZAJE DE CARACTERÍSTICAS

4.4.3.b. CONCLUSIONES SOBRE LA METODOLOGÍA HÍBRIDA

Concluyendo, es posible responder a la parte inicial de la primera pregunta de investigación, ¿puede el aprendizaje de características empleadas en enfoques QSAR resultar de utilidad en el contexto de Informática de Polímeros?, que el uso exclusivo de aprendizaje de características mediante el método CODES-TSAR no es suficiente para modelar y predecir propiedades mecánicas empleando una base de datos de polímeros de alto peso molecular. Por otro lado, con respecto a la pregunta sobre ¿qué sucede si combinamos las ventajas del aprendizaje de características con las de la selección de características para la inferencia de modelos QSPR en polímeros de alto peso molecular?, puede responderse que la hibridación propuesta contribuye con información complementaria a los modelos QSPR, aunque la alta precisión predictiva lograda por los modelos QSAR de propiedades farmacocinéticas no se mantienen en el estudio de las propiedades mecánicas asociadas con los materiales poliméricos.

Síntesis y Conclusiones del Capítulo 4

El uso de la informática en el dominio de la ciencia de los polímeros es un desafío constante, y en pleno auge, con muchos problemas que siguen sin resolverse. Sin embargo, la informática es ineludible para ayudar a comprender el comportamiento físicoquímico y a diseñar *in silico* materiales. En este capítulo se hizo un recorrido por los conceptos más importantes de Informática y Química de Polímeros. Además, se presentaron las contribuciones más destacadas en la literatura en el modelado QSPR de polímeros que utilizan a la Unidad Repetitiva Estructural como modelo molecular. En particular para la representación molecular, se vuelve necesario el alejamiento del paradigma de la tabla de conexión, y en este sentido, en esta tesis, se avanzó en la representación de polímeros por código SMILES. Esto permitió el empleo de herramientas computacionales como la dupla metodológica CODES-TSAR, inicialmente creadas para el diseño y descubrimientos de fármacos, en Informática de Polímeros, aunque con resultados no alentadores.

Al respecto, se trabajó con tres metodologías que denominamos 1-Clásica, 2-con Análisis Visual y 3-Híbrida, que a su vez emplean diferentes métodos de Aprendizaje Maquinal. La Metodología Clásica fue propuesta por Palomba y sirvió en esta tesis como entrenamiento en el área, ya que a través de ella se infirieron por primera vez modelos QSPR para la predicción de propiedades mecánicas para polímeros de alto peso molecular. Sin embargo, esta metodología es de difícil y tediosa aplicación, ya que es necesario trabajar con múltiples herramientas computacionales de forma manual, es decir, que no permite la construcción de un flujo de trabajo automático. Además, tiene asociados ciertos problemas de sobreajuste en sus modelos resultantes. En este sentido se avanzó con la propuesta de la Metodología que incluye Analítica Visual, donde al hacer uso de VIDEAN se reduce la cantidad de softwares implicados en el análisis y/o construcción de los subconjuntos de descriptores que darán lugar a los modelos QSPR. Al emplear VIDEAN, se reduce el riesgo de error en la manipulación manual de datos y, además, se brinda la opción de usar tres tipos diferentes de métodos de aprendizaje maquinal para la inferencia de modelos QSPR. Como resultado, los modelos obtenidos tienen baja cardinalidad, alta interpretabilidad fisicoquímica y están libres de sobreajuste. En contraste, la Metodología Híbrida no funcionó como se esperaba en el modelado QSPR de polímeros de alto peso molecular, ya que no resultó lo suficientemente precisa como para lograr modelos predictivos QSPR robustos en términos estadísticos, de baja cardinalidad y fisicoquímicamente interpretables.

Una de las problemáticas en la Informática de Polímeros es la escasa disponibilidad de bases de datos. Algunos autores sugieren que esto se debe principalmente a un problema cultural, ya que no existe un fluido intercambio y reutilización de datos como en otras disciplinas. Al respecto, en este capítulo se describe la base de datos utilizada, sus ventajas y sus limitaciones. Otro problema no resuelto en el área es la representación molecular fidedigna de estos materiales complejos. Sobre esto se avanzó en la representación de polímeros mediante código SMILES, que resulta adecuado para abordar las preguntas de investigación planteadas en este capítulo, así como las demás de esta tesis.

MODELADO QSPR CON DMS UNIVALUADOS

CAPÍTULO 5

Este capítulo abarca la experimentación desarrollada buscando responder a la pregunta sobre la existencia de una representación univaluada, es decir, una única representación de peso capaz de caracterizar la complejidad estructural de los polímeros polidispersos. Parte de la hipótesis dice que el modelo molecular basado en URE, mayoritariamente utilizado en la literatura, es un modelo sobresimplificado.

5.1. MODELADO QSPR DE MACROMOLÉCULAS

En esta sección, se pretende mostrar las limitaciones más importantes respecto del modelado QSPR de polímeros de alto peso molecular. Para modelar polímeros, lo primero que debe definirse es cómo deben ser representados a nivel molecular (modelo molecular) [Wu *et al.*, 2016]. Los métodos tradicionales de representación molecular, en Quimioinformática, solo tienen una aplicabilidad limitada [Petrosyan *et al.*, 2019]. Una idea natural y común es usar la Unidad Repetitiva Estructural (URE) de los polímeros para caracterizar polímeros no ramificados, como es el caso de nuestra base de datos (ver Capítulo 4). También se han utilizado representaciones de oligómeros, aunque menos frecuentemente, como por ejemplo, en el trabajo de Cypcar *et al.*, en el que usan 20 UREs, para la generación de modelos QSPR para calcular la temperatura de transición vítrea (Tg), usando cálculos de minimización de energía y dinámica molecular [Cypcar *et al.*, 1996].

El cálculo directo de Descriptores Moleculares (DMs) en cadenas de polímeros de todas las longitudes realistas no es factible, porque las cadenas de polímeros son mucho más grandes que las moléculas pequeñas clásicas. No existen herramientas para la representación computacional ni para el cálculo de descriptores moleculares (DMs) especialmente diseñadas para trabajar con macromoléculas que alcancen el peso de las cadenas que componen a los polímeros sintéticos. Debido a esta brecha tecnológica es que la gran mayoría de los estudios publicados de modelado QSPR utilizan la URE como modelo molecular [Wu *et al.*, 2016].

Con respecto al problema del cálculo de descriptores, se pueden encontrar algunas alternativas en la literatura científica. Por un lado, está la estrategia de Katrizky que considera la extensión de los valores del descriptor para cadenas largas utilizando un tratamiento numérico [Katritzky *et al.*, 1996]. Por el otro, está la técnica llamada Descriptores de Cadena Infinita propuesta en 2016 [Wu *et al.*, 2016]. Estos últimos descriptores están libres de la influencia de los átomos extremos (*end-capped*) y fueron especialmente ideados para caracterizar polímeros orgánicos utilizando bibliotecas de fragmentos predefinidos. Sin embargo, el cálculo de DMs sobre cadenas realistas de polímeros, en cuanto al tamaño verdadero, aún representa un desafío. Es importante lograr avances en este sentido, ya que al utilizar, por ejemplo, el enfoque de Descriptores de Cadenas Infinitas no se tienen en cuenta los descriptores extensivos (varían con la extensión de la representación molecular, dependen del modelo molecular utilizado) que resultan muy útiles a la hora de la predicción de propiedades mecánicas [Cravero *et al.,* 2019a].

5.1.1. LIMITACIONES EN LA REPRESENTACIÓN COMPUTACIONAL

Para poder realizar una predicción *in silico* es necesario representar computacionalmente las moléculas implicadas. Esta tarea no resulta para nada sencilla en el caso de polímeros de alto peso molecular [Cravero *et al.,* 2016c]. Nuestra base de datos cuenta para cada polímero con sus valores de pesos promedios en número (47000-765000 [g/mol]), conocidos con el acrónimo en inglés Mn (*Number Average Molecular Weight*), y de pesos promedios en peso (19500-2200000 [g/mol]), conocidos con el acrónimo Mw (*Weight Average Molecular Weight*). El objetivo es describir los polímeros con modelos moleculares a escalas más realistas, por ejemplo, empleando las cadenas poliméricas cuyos pesos se correspondan con sus valores de Mn y Mw, para posteriormente poder calcular los DMs de dichas cadenas. Por ende, en las siguientes secciones del presente capítulo, cuando un polímero esté caracterizado por la molécula correspondiente a su URE, diremos que estamos empleando el ***modelo molecular URE*** para representar al polímero. De manera similar, cuando el polímero esté caracterizado por las cadenas poliméricas que conciernen a los pesos Mn o Mw de dicho polímero, diremos que estamos empleando como representación al ***modelo molecular Mn*** y al ***modelo molecular Mw*** respectivamente.

El primer intento para poder generar cadenas poliméricas que representen a los modelos moleculares Mn y Mw de nuestra base de datos se realizó utilizando el software HyperChem [Laxmi & Priyadarshy, 2002], el cual cuenta con una opción

para polimerizar desde un monómero, indicando el número de UREs que se quiere unir [Cravero *et al.*, 2016d]. De este modo, se buscaba generar, vía polimerización, las macromoléculas correspondientes a las cadenas poliméricas de peso Mn y Mw correspondientes a cada polímero. Según nuestro conocimiento, este es el único software que ofrece esta función para polímeros de tipo comerciales. También, permite elegir el tipo de unión por ejemplo cabeza-cola. El programa procesa la polimerización de a una URE a la vez, aunque tiene un límite en la cantidad que depende del tipo de URE (ver Figura 5.1). Luego, la macromolécula debe ser estabilizada energéticamente, es esta etapa la de mayor consumo de tiempo y recursos [Cravero *et al.*, 2016c].

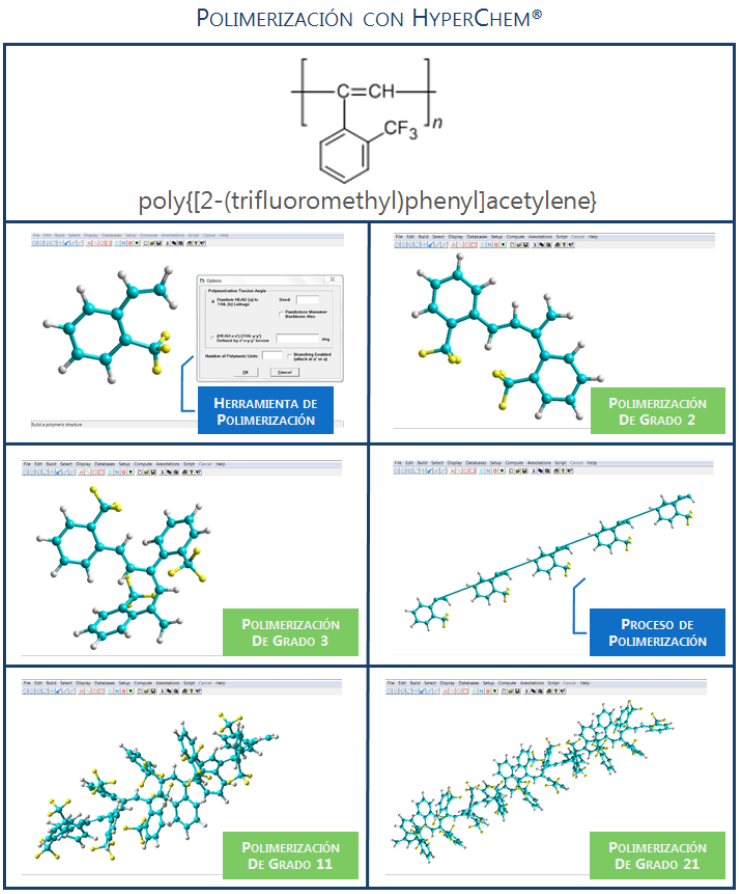


FIGURA 5. 1. EJEMPLO DEL USO DEL PAQUETE DE POLIMERIZACIÓN DE HYPERCHEM USANDO LA URE DE POLÍMERO ID24: POLY{[2- (TRIFLUOROMETHYL)PHENYL]ACETYLENE}.

Lamentablemente, HyperChem no logra ejecutar las polimerizaciones correspondientes a los pesos moleculares Mn y/o Mw, quedando muy lejos del objetivo de generar los modelos moleculares Mn y Mw para la representación de nuestra base de datos. En la Tabla 5.1 se observan algunos casos a modo de ejemplo de la diversidad estructural y del grado de polimerización (GP) necesarias para alcanzar los pesos promedios y, finalmente, alcanzada por HyperChem. Para la

realización de esta experimentación se utilizó un procesador Intel® Core™ i5-3340 CPU @ 3.10 Ghz y 4GB de RAM, como equipo de cómputo. El limitante fue el software HyperChem y no la arquitectura del equipo, ya que se comprobó que con equipos que triplicaban la disponibilidad de memoria y con mayor procesamiento de cómputo tampoco se alcanzaban los pesos deseados. En cada uno de los casos presentes en la Tabla 5.1 tanto la polimerización como la posterior estabilización molecular consumieron más de 400hs de cómputo.

TABLA 5. 1. EJEMPLOS DEL GRADO DE POLIMERIZACIÓN (GP) Y PESO MOLECULAR (PM) ALCANZADO POR HYPERCHEM EN COMPARACIÓN CON LOS PESOS PROMEDIOS (MN Y MW) DESEADOS.

	Mn		**Mw**		**HyperChem**	
	PM [g/mol]	**GP**	**PM [g/mol]**	**GP**	**PM [g/mol]**	**GP**
ID3	765000	7204	880000	8287	21342	201
ID24	190000	1103	690000	4007	25996	151
ID30	750000	2970	2200000	8714	111044	201
ID64	154000	212	332500	454	102210	141
ID72	63000	77	115000	141	46227	57

Frente a estas limitaciones, comenzó a investigarse la posibilidad de representar los polímeros mediante código SMILES (Especificación Simplificada de Entrada Molecular en Línea de texto), a pesar de que al partir de un grafo como representación de la molécula (estructura desplegada con ciclos abiertos) permite más de un tipo de unión posible entre UREs y es necesario establecer correctamente los átomos que representan la cabeza y la cola de la URE para no transformar la estructura que se quiere representar en otra. Este último aspecto se explica en detalle en el siguiente punto.

5.1.2. POLYMAS: POLIMERIZACIÓN BASADA EN SMILES

El código SMILES consiste en un método basado en teoría de grafos, que genera una cadena de caracteres que codifica estructuras moleculares de forma unívoca. El primer objetivo es reconocer en el código SMILES de una URE el inicio (cabeza) y el fin (cola) de la misma, para lograr realizar una polimerización *in silico* del tipo cabeza-cola. Al trabajar con SMILES canónicas de la URE se identifica tanto la cabeza como la cola de la molécula con el símbolo asterisco (*). Esta representación permite localizar en el código SMILES los indicadores de los extremos de la cadena principal de la molécula. Así, la polimerización basada en SMILES se realiza al reemplazar el segundo asterisco en el SMILES de una URE por el SMILES de otro, de forma recursiva [Cravero *et al.,* 2017c]. En la Figura 5.2 se presenta un ejemplo de esta polimerización *in silico* usando como ejemplo al estireno.

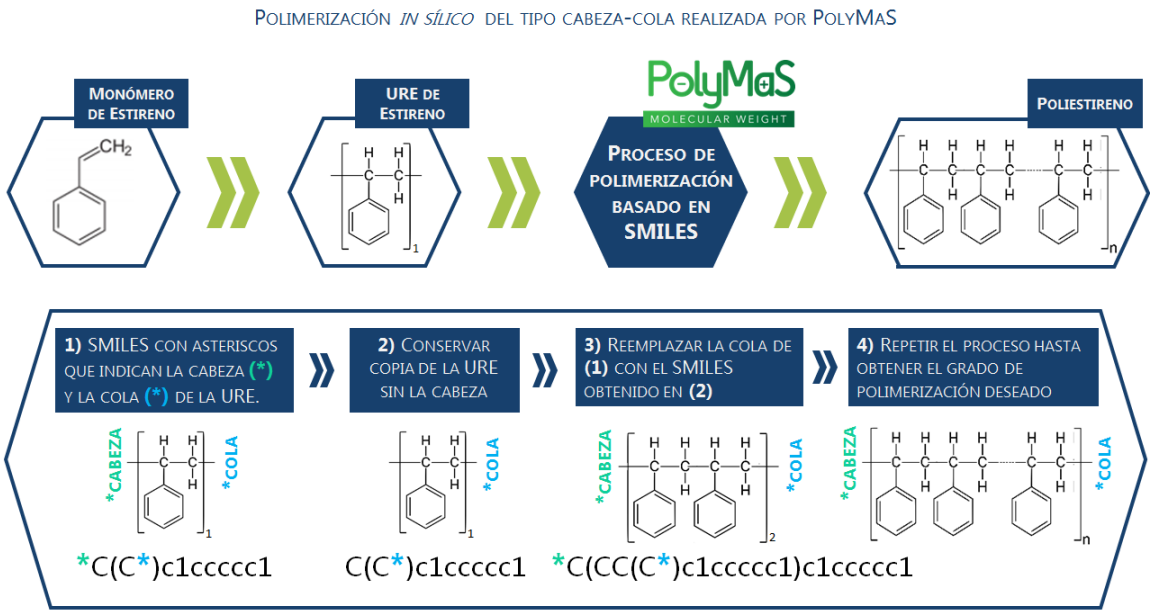


FIGURA 5. 2. ESQUEMA DE LA POLIMERIZACIÓN *IN SILICO* DEL TIPO CABEZA-COLA BASADA EN SMILES USANDO COMO EJEMPLO AL ESTIRENO.

PolyMaS (*Polymer Maker Smiles-based*) es una herramienta desarrollada en nuestro grupo que imita una polimerización cabeza-cola basada en SMILES. PolyMaS une el extremo de una URE con el extremo de otra URE, este proceso se repite hasta obtener el peso o largo de cadena deseado. Para esto es necesario ingresar el número de UREs que se desean unir. El algoritmo de PolyMaS realiza la polimerización en paralelo, la cantidad de ejecuciones paralelas (hilos) dependen del largo de cadena deseado (en términos de cantidad de URE), es decir, el nivel de paralelismo crece proporcionalmente con la cantidad de UREs que conformarán el polímero final. De esta manera se obtiene en paralelo una cierta cantidad de cadenas (substrings) en formato SMILES, las cuales son unidas siguiendo la misma lógica. Es decir, en lugar de unirse una única URE a una cadena conformada por varias UREs, se unen cada una de las cadenas (substrings) entre sí, hasta conformar el polímero completo [Schustik *et al.*, 2019b].

Hasta el momento no se han encontrado limitantes en la capacidad de polimerización de PolyMaS, ya que se ha logrado polimerizar polímeros con pesos en el orden de los 100000 [g/mol] con consumos de tiempo y recursos reducidos. En la Tabla 5.2 se muestran 5 ejemplos de polímeros con sus pesos moleculares promedios (Mn y Mw) y el Grado de Polimerización (GP) necesario para alcanzar ese peso. Además, se muestra el tiempo, medido en segundos [s], de ejecución de PolyMaS necesarios para llevar a cabo dichas polimerizaciones. La ejecución de PolyMas se realizó con las siguientes especificaciones de hardware: Intel CPU® Core™ i3-7100 3.90 GHz y 4 GB de RAM.

TABLA 5. 2. EJEMPLOS DEL GRADO DE POLIMERIZACIÓN (GP) Y PESO MOLECULAR (PM) Y LOS TIEMPOS DE EJECUCIÓN DE POLYMAS, MEDIDOS EN SEGUNDOS [S], PARA ALCANZARLOS.

	Mn			Mw		
	PM [g/mol]	**GP**	**Tiempo [s]**	**PM [g/mol]**	**GP**	**Tiempo [s]**
ID3	765000	7204	1.205	880000	8287	2.095
ID24	190000	1104	0.354	690000	4007	0.800
ID38	11000	17	0.062	26500	41	0.001
ID56	42000	38	0.135	94000	85	0.002
ID77	43000	69	0.090	158000	525	0.005

5.2. LIMITACIONES DE LA REPRESENTACIÓN POR URE

Hasta el momento se han presentado los motivos por los cuales usar la URE como modelo molecular para materiales poliméricos no es adecuada. Además, se avanzó en la representación computacional de moléculas que alcanzan los pesos promedios de los polímeros en nuestra base de datos. El paso siguiente es demostrar que efectivamente existen limitaciones cuando se crean modelos QSPR con la representación por URE de los polímeros. Las preguntas de investigación específicas que se plantean para saber cuáles son las limitaciones de usar URE se listan más abajo. En su conjunto, serán referidas como Segunda Pregunta de Investigación, aunque el lector encontrará que se derivan en subpreguntas **a** y **b**. Nota: la Primera Pregunta de Investigación fue tratada en el capítulo 4.

a). ¿Son los modelos QSPR basados en el modelo molecular URE efectivos cuando se testean sobre modelos moleculares de alto peso?

b). Los descriptores moleculares que fueron seleccionados en los modelos QSPR basados en el modelo molecular URE ¿pueden resultar de utilidad para inferir nuevos modelos QSPR a partir de base de datos de otras instancias univaluadas de representación de mayor peso que URE (Mn y Mw)?

5.2.1. MODELADO QSPR PARA URE

Para saber si la representación usada hasta ahora en la literatura, modelo molecular URE, es suficiente para generar modelos QSPR predictivos realistas para las tres propiedades mecánicas en estudio (Módulo de Tensión, Elongación a la Rotura y Resistencia a la Rotura), se diseñó un experimento cuya metodología puede verse en la Figura 5.3. Para comprender mejor el diseño experimental, su explicación se ha dividido en 4 etapas consecutivas que se describen a continuación.

Etapa 1: Creación de bases de datos.

Esta etapa se describe en la parte superior de la figura. Partiendo de 57 polímeros (tomados desde la base de datos original de 77), se calcularon 57 descriptores moleculares sobre las representaciones basadas en: URE, Mn y Mw. Puede verse un ejemplo en la figura para el poliestireno, donde se muestran los valores calculados para dos descriptores (D_1: nAtom –número de átomos- y D_{57}: nAromRing –número de anillos aromáticos-). Nótese, que para cada una de las tres instancias de peso el mismo descriptor toma valores diferentes.

Etapa 2: Modelado QSPR de URE: A- Selección de Características, B- Aprendizaje Maquinal.

Esta etapa se describe en la columna de la izquierda en sombreada en tono verde lima. *A*- Sobre la base de datos correspondiente al modelo molecular URE (verde), se realizó la fase de selección de características con 5 métodos diferentes provistos por WEKA (en la figura se muestra esta etapa recuadrada en azul bajo el título Selección de DMs). Esto dio como resultado 5 subconjuntos de DMs diferentes (representados en la figura por cubos y reglas verdes). *B*- Sobre cada uno de estos 5 subconjuntos se aplicaron 4 métodos de Aprendizaje Maquinal para el entrenamiento de los modelos QSPR (cuadro azul bajo el título Modelado QSPR). El resultado de esta fase dio 20 modelos QSPR, entrenados y validados para URE (Triángulos verdes: el de mayor tamaño representa al subconjunto de entrenamiento y el de menor tamaño al subconjunto de validación externa).

Etapa 3: Pregunta de Investigación ***a***

Esta etapa está representada en la última parte de la columna izquierda de la figura. Consistió en validar externamente los 20 modelos QSPR inferidos para URE (triángulos grandes verdes), con los dos subconjuntos reservados para validación externa, correspondientes a las bases de datos de Mn y Mw de Etapa 1 (Triángulos pequeños en colore celeste para Mn y azul para Mw).

Etapa 4: Pregunta de Investigación ***b***

Esta etapa está representada en derecha de la figura. Consistió en inferir modelos QSPR para Mn y Mw usando los descriptores seleccionados en etapa 2-A (cubos verdes con reglas en color celeste para Mn y azul para Mw, como una representación gráfica de la instancia de peso sobre la que se miden o calcularon los DMs). Es decir, no se realizó nuevamente el proceso de Selección de Características, sino que directamente se infirieron los modelos QSPR para cada instancia y se los validó externamente en la misma (ambos triángulos celestes para

Mn y ambos triángulos azules para Mw). Nótese que, si bien los descriptores son los mismos, los valores que estos toman no son iguales, como se explicó en Etapa 1 (recuadro con línea punteada en azul bajo el título Bases de Datos en la parte superior de la figura).

Al desarrollar la Etapa 1 es válido aclarar que, si bien el problema de representación de moléculas de alto peso molecular efectivamente pudo superarse usando PolyMaS, el cálculo de descriptores aún representa una brecha tecnológica. Por este motivo, la base de datos utilizada para esta experimentación contiene solamente 57 polímeros y 57 DMs clásicos para cada polímero. A continuación, se irá explicando cómo fue armada, los criterios considerados y los problemas que se fueron presentando, que obligaron a reducir desde los 77 polímeros iniciales a los 57 polímeros utilizados en esta experimentación. La primera reducción de la bases de datos que se hace es la eliminación del polímero ID1, ya que no posee valor experimental para una de las propiedades a predecir (Módulo de Tensión). Para el cálculo de descriptores se utilizó la librería RCDK [Guha, 2007], que permite obtener 302 descriptores moleculares. Para el modelo molecular URE se pudieron calcular todos (302) los DMs para todos (76) los polímeros, no así para los modelos moleculares de Mn y Mw, donde solo pudo trabajarse con 72 polímeros, porque los 4 restantes eran computacionalmente intratables para el paquete RCDK.

Otra aclaración importante de esta etapa es que las bases de datos deben homogenizarse, es decir, establecer cuáles serán los polímeros que las integran y cuáles serán los descriptores moleculares incluidos, que para las tres bases de datos (URE, Mn y Mw), deben ser los mismos. Muchos de los descriptores calculados por RCDK produjeron valores NA (sin respuesta), y para tratar este problema, fue necesario realizar un método de filtrado de la base de datos que priorizara la cantidad de polímeros, sin dejar de lado la diversidad de DMs. Es decir, el filtro debía realizar una relación de compromiso entre la cantidad de polímeros que dejaba y la cantidad de descriptores moleculares, sin embargo, debido a las características de la base de datos era importante conservar la mayor cantidad de polímeros posibles. Esta reducción dio como resultado 61 polímeros y 57 DMs clásicos para las tres instancias de representación (modelos moleculares URE, Mn y Mw).

Además, para un correcto tratamiento de datos, se dividió a cada base de datos en cuatro subconjuntos que mantuvieran la diversidad estructural de la base de datos original. A partir de este tratamiento, se identificaron 4 de esos 61 polímeros como atípicos estructuralmente (familia química no representada en la base de datos, debido a la eliminación de los otros 21 polímeros), y por lo tanto,

fueron eliminados. Finalmente, como resultado general de la Etapa 1, se generaron 3 bases de datos con 57 polímeros y 57 descriptores cada una. Dividida cada una de ellas en un conjunto de entrenamiento compuesto por tres de los cuatro subconjuntos (utilizado en la Selección de Características y en el entrenamiento de los modelos QSPR) y el cuarto subconjunto se reservó para ser utilizado como conjunto de validación externa.

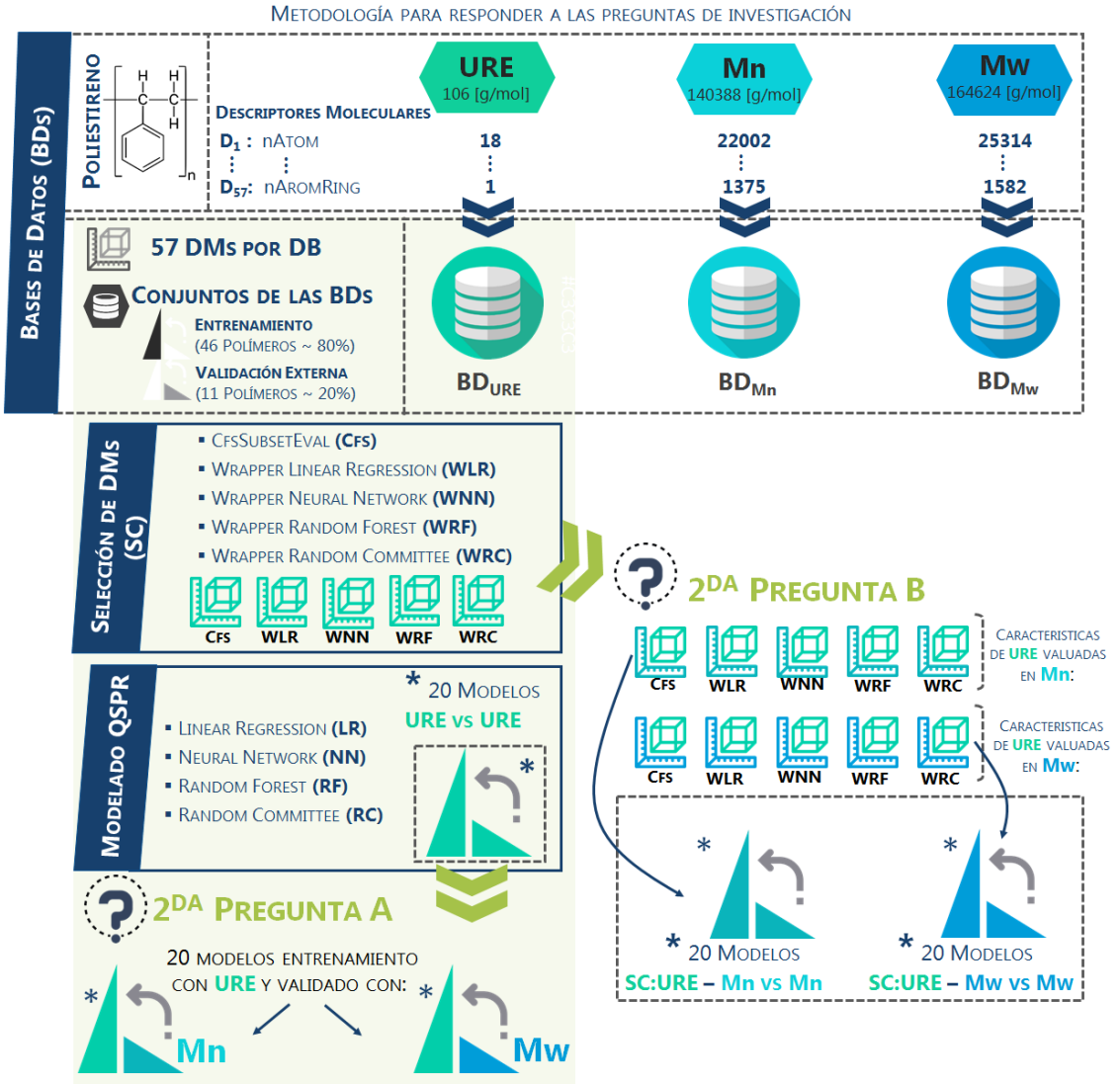


FIGURA 5. 3. ESQUEMA DE DA METODOLOGÍA SEGUIDA PARA RESPONDER A LAS PREGUNTAS a Y b.

Con la Etapa 1 resuelta, se continuó con la Etapa 2-A que consiste en la selección de los descriptores para la URE, resultando en 5 subconjuntos de DMs, uno por cada método: *CfsSubsetEval* (Cfs), *Wrapper Linear Regression* (WLR), *Wrapper Neural Network* (WNN), *Wrapper Random Forest* (WRF), y *Wrapper Random Committee* (WRC). La Tabla 5.3 muestra los DMs seleccionados para cada una de las propiedades (Módulo de Tensión, Elongación a la Rotura y Resistencia a

la Rotura). Este paso se realizó utilizando el conjunto de datos de entrenamiento, que está formado por 46 polímeros (ver Figura 5.3).

TABLA 5. 3 DESCRIPTORES SELECCIONADOS POR 5 MÉTODOS DE SELECCIÓN DE CARACTERÍSTICAS PARA EL MODELO MOLECULAR BASADO EN URE. RESULTADOS PARA LAS 3 PROPIEDADES MECÁNICAS ESTUDIADAS.

	Método	Cardinalidad	Descriptores Moleculares para URE
Módulo de Tensión	Cfs	8	tpsaEfficiency_URE, khs.dsCH_URE, khs.dsssP_URE, khs.ssS_URE, ddssS_URE, Kier3_URE, HybRatio_URE, nHBDon_URE
	WLR	8	MLogP_URE, nAtomLAC_URE, Kier3_URE, HybRatio_URE, nHBDon_URE, nHBAcc_URE, nAtom_URE, nAromBond_URE
	WRC	3	Kier3_URE, nHBAcc_URE, C2SP2_URE,
	WRF	4	khs.sCH3_URE, khs.dO_URE, Kier3_URE, nHBAcc_URE
	WNN	7	VAdjMat_URE, MLogP_URE, khs.ssNH_URE, khs.ssS_URE, Kier3_URE, HybRatio_URE, nAcid_URE
Elongación a la Rotura	Cfs	3	tpsaEfficiency_URE, khs.dsssP_URE, ALogP_URE
	WLR	12	tpsaEfficiency_URE, khs.sssCH_URE, khs.sOH_URE, khs.dO_URE, khs.dsssP_URE, khs.ddssS_URE, HybRatio_URE, nHBDon_URE, nHBAcc_URE, C3SP2_URE, C1SP3_URE, C3SP3_URE, nAcid_URE
	WRC	4	tpsaEfficiency_URE, khs.sssCH_URE, khs.dssC_URE, nHBAcc_URE
	WRF	2	tpsaEfficiency_URE, nHBAcc_URE
	WNN	7	khs.dsCH_URE, khs.sssCH_URE, khs.ssNH_URE, khs.ssO_URE, khs.dsssP_URE, nHBAcc_URE, nAromBond_URE
Resistencia a la Rotura	Cfs	6	nAromRings_URE, khs.dsCH_URE, khs.sssN_URE, khs.ssS_URE, khs.ddssS_URE, nAcid_URE
	WLR	6	nAromRings_URE, MW_URE, khs.ssNH_URE, khs.dsssP_URE, khs.ssS_URE, AMR_URE
	WRC	6	nRings4_URE, khs.ssNH_URE, khs.dO_URE, khs.ssS_URE, khs.ddssS_URE, nHBAcc_URE
	WRF	6	nSmallRings_URE, nRings4_URE, nAtomP_URE, khs.sCH3_URE, khs.ssNH_URE, khs.ssS_URE
	WNN	3	nSmallRings_URE, MW_URE, khs.dsssP_URE

Posteriormente en la Etapa 2-B, se entrenaron los modelos QSPR y se testearon usando un conjunto de datos de validación externa, compuesto por 11 polímeros,) donde los polímeros también están representados por el modelo molecular URE. La Tabla 5.4 muestra los resultados de este proceso. Son 20 los modelos QSPR obtenidos por propiedad, ya que los mismos son entrenados con 4 métodos de aprendizaje maquinal: *Linear Regression* (LR), *Neural Network* (NN), *Random Forest* (RF) y *Random Committee* (RC). El criterio utilizado para evaluar los modelos QSPR en términos de rendimiento (R^2) tanto para la etapa de entrenamiento como de validación externa, probados en el mismo tamaño de representación molecular, es el siguiente: Bajo: $R^2 < 0.6$; Intermedio: $0.6 \leq R^2 < 0.80$; Alto: $R^2 \geq 0.80$.

TABLA 5.4 RENDIMIENTOS DE LA VALIDACIÓN EXTERNA PARA URE DE LOS MODELOS ENTRENADOS CON URE A PARTIR DE LOS DMS SELECCIONADOS PARA URE.

URE vs. URE

	Módulo de Tensión				**Elongación a la Rotura**				**Resistencia a la Rotura**			
	Cfs				Cfs				Cfs			
	LR	NN	RF	RC	LR	NN	RF	RC	LR	NN	RF	RC
R^2	0.88	0.95	0.96	0.95	0.58	0.63	0.49	0.45	0.82	0.84	0.81	0.80
MAE	0.24	0.20	0.20	0.21	2.18	3.24	2.41	3.43	8.43	8.13	9.73	10.47
RMSE	0.35	0.27	0.30	0.30	2.83	5.37	3.63	5.84	10.85	9.94	11.60	12.16
	WLR				WLR				WLR			
	LR	NN	RF	RC	LR	NN	RF	RC	LR	NN	RF	RC
R^2	0.94	0.89	0.91	0.88	0.90	0.71	0.57	0.54	0.80	0.84	0.83	0.80
MAE	0.22	0.25	0.26	0.29	1.98	1.37	1.35	1.45	10.95	8.53	9.55	10.00
RMSE	0.32	0.32	0.35	0.36	2.22	1.78	1.71	1.90	12.80	11.17	11.13	12.29
	WNN				WNN				WNN			
	LR	NN	RF	RC	LR	NN	RF	RC	LR	NN	RF	RC
R^2	0.88	0.92	0.97	0.85	0.72	0.57	0.65	0.53	0.68	0.77	0.76	0.79
MAE	0.26	0.24	0.20	0.30	1.78	1.57	1.46	1.67	14.41	11.32	11.26	10.62
RMSE	0.37	0.30	0.30	0.39	2.15	1.96	1.65	2.03	16.90	14.62	14.80	14.03
	WRF				WRF				WRF			
	LR	NN	RF	RC	LR	NN	RF	RC	LR	NN	RF	RC
R^2	0.93	0.75	0.96	0.94	0.73	0.62	0.46	0.39	0.74	0.79	0.78	0.77
MAE	0.30	0.45	0.22	0.22	1.77	2.44	1.52	1.62	10.13	11.77	10.35	10.80
RMSE	0.41	0.55	0.31	0.30	2.10	3.28	1.93	2.32	12.98	12.96	12.20	13.13
	WRC				WRC				WRC			
	LR	NN	RF	RC	LR	NN	RF	RC	LR	NN	RF	RC
R^2	0.84	0.91	0.93	0.91	0.88	0.74	0.47	0.49	0.81	0.85	0.86	0.85
MAE	0.35	0.22	0.24	0.26	1.92	2.12	1.48	1.60	9.82	8.45	8.24	8.47
RMSE	0.41	0.30	0.33	0.33	2.14	2.46	1.92	2.11	12.00	9.67	9.56	9.70

Módulo de Tensión:

Todos los rendimientos de los modelos QSPR para Módulo de Tensión se consideran altos en términos de R^2, dado que están por encima de 0,80. Excepto por el modelo entrenado con NN para el subconjunto de características obtenido por WRF, que tiene un rendimiento de nivel intermedio (R^2= 0,75). También en la Tabla 5.4, se muestran dos métricas de errores: el error absoluto medio (MAE) y el error cuadrático medio (RMSE), siendo todos muy bajos. Los modelos poseen alto desempeño en general (95% arriba de R^2= 0.80).

Elongación a la Rotura:

Todos los modelos QSPR para Elongación a la Rotura entrenados con RC tienen bajo rendimiento. Aquellos entrenados con RF también tienen bajo

rendimiento, excepto para el subconjunto provisto por WNN. Exactamente lo opuesto sucede con los modelos entrenados con NN. Para los entrenados con LR, los rendimientos son intermedios para Cfs, WNN y WEF; y son altos para los subconjuntos de WLR y WRC. En general, los errores son bajos, aunque podría deberse a la amplitud del rango de valores de la propiedad. Sin embargo, los modelos QSPR obtenidos tienen un rendimiento *Intermedio* (10% arriba de R^2= 0.80, 65% por debajo de R^2=0.60).

Resistencia a la Rotura:

Ninguno de los modelos QSPR de Tensión a la Rotura tiene un rendimiento bajo. Los subconjuntos Cfs, WLR y WRC entrenados con los cuatro métodos de aprendizaje alcanzan valores altos de rendimiento. Los valores de los errores son superiores a los de los casos de las propiedades anteriores, pero esto podría relacionarse con la mayor amplitud en el rango de los valores de esta propiedad. En tal sentido, en esta propiedad la amplitud de valores es 95.5, mientras que para Elongación a Rotura y Módulo de Tensión es en 38.7 y 3.87 respectivamente. En cuanto a la precisión, todos los modelos poseen un R^2 superior al 0.60 y, además, el 50% de los mismos están incluso por encima del 0.80 de R^2.

5.2.2. RESPONDIENDO LA SEGUNDA PREGUNTA DE INVESTIGACIÓN

En las etapas 3 y 4 se intenta responder a las preguntas de investigación formuladas, comparando el rendimiento de los modelos originales (Tabla 5.4: URE vs. URE) y los resultados obtenidos por los modelos inferidos para la experimentación de estas etapas, explicada en la subsección 5.2.1 (Fig. 5.3). La comparación se realizó mediante la diferencia entre los valores de R^2 obtenidos por los nuevos modelos y los valores de R^2 obtenidos por los modelos originales. De esta manera, cuando esta diferencia es positiva el rendimiento aumenta y cuando el valor de R^2 es menor en el nuevo modelo (diferencia negativa) el rendimiento disminuye. El criterio de evaluación del desempeño de los modelos fue discretizado en las siguientes categorías, usando un color para identificar cada una, mostrados en la Figura 5.4:

- *Disminución Abrupta* (Rojo): R^2 disminuye más del 40%;
- *Disminución Moderada* (Anaranjado): disminución de R^2 entre 5%-40%;
- *Invariante* (Amarillo): R^2 varía entre ± 5%;
- *Incremento* (Verde): R^2 aumenta más del 5%.

Para comprender la Figura 5.4, se deben observar primero las dos columnas de la izquierda (5.4.a) que responden la pregunta **a** y más adelante las dos de la

derecha (5.4.b) que responden la pregunta **b**, todas brindan la información utilizando los códigos de color mencionados justo antes. En las dos primeras columnas (Figura 5.4.a), se presentan los resultados para la evaluación de los modelos QSPR entrenados usando el modelo molecular URE y probados sobre polímeros representados con el modelo molecular Mn (URE vs Mn) y la misma información para las pruebas efectuadas sobre polímeros representados con el modelo molecular Mw (URE vs Mw). En la parte superior de la imagen se muestran los resultados para Módulo de Tensión, en la parte media para Elongación a la Rotura y en la parte inferior, para Resistencia a la Rotura.

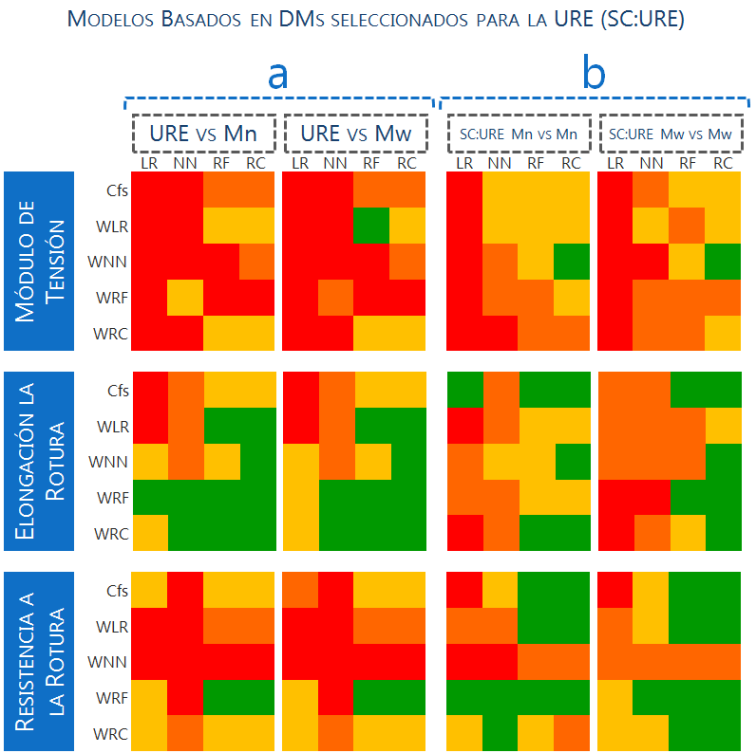


FIGURA 5. 4 Resumen gráfico de la calidad en la precisión alcanzada, en comparación, por los modelos basados en los DMs seleccionados para la URE para responder las preguntas **a** y **b**.

Módulo de Tensión:

En lo referido a la pregunta **a**, observando las dos primeras columnas de la izquierda de la Figura 5.4 para el bloque de Módulo de Tensión, se puede observar que la precisión predictiva, en términos de R^2, para los modelos QSPR entrenados con representación URE y evaluados en Mn tuvo una caída en casi todos los casos (60% con *Disminución Abrupta* y 15% con *Disminución Moderada*) mientras que el resto se mantiene estable (25% con *Invariante*) y ninguno crece en desempeño (0% con *Incremento*). Se observó un comportamiento similar cuando se evaluaron estos modelos en Mw: 85% de caída (65% *Disminución Abrupta* y 20% *Disminución Moderada*), 10% permanece constante (*Invariante*) y 5% de aumento (*Incremento*).

Además, se puede observar en la Tabla A.5.1, en la sección Anexos del capítulo 5, que en todos los casos los errores aumentan significativamente denotando la mala calidad de los modelos QSPR. Frente a estos resultados se puede concluir que tanto para Mn como para Mw no sería recomendable utilizar los modelos QSPR entrenados con los polímeros representados con el modelo molecular URE. Esta conclusión parcial queda a la vista en el resumen gráfico mostrado en la Figura 5.4 bloque horizontal Módulo de Tensión, columnas de la izquierda, porque el color rojo es preponderante.

Respecto a la pregunta **b**, en las dos columnas de la derecha de la Figura 5.4.b se pueden ver los resultados de los modelos QSPR cuando estos se entrenan y validan con polímeros representados con el mismo modelo molecular. Esto significa que los modelos QSPR se construyeron a partir de los mismos descriptores seleccionados usando la representación URE (Tabla 5.3), pero se recalcularon para los modelos moleculares Mn y Mw. Luego, estos descriptores fueron utilizados para entrenar y validar los modelos QSPR para polímeros representados con su correspondiente modelo molecular: Mn y Mw (Figura 5.3 columna derecha). En este caso, al hacer el entrenamiento y validación de los modelos QSPR para la misma instancia de representación (modelo molecular), se espera obtener mejores resultados que en los ensayos anteriores, donde los modelos QSPR fueron entrenados para otra instancia de representación (URE). Esta hipótesis se comprueba ya que para Mn/Mw, el 30%/30% de los modelos tienen una *Disminución Abrupta*, especialmente para LR, mientras que 25%/35% solo presentan *Disminución Moderada*, 40%/30% se mantiene con rendimiento *Invariante* y 5%/5% presentan *Incrementos* (ver las dos columnas de la derecha de la Figura 5.4). Estas conclusiones quedan evidenciadas visualmente en la Figura 5.4 ya que el color rojo, el cual era preponderante en la pregunta **a**, ha disminuido en la pregunta **b**.

Elongación a la Rotura:

En lo referido a la pregunta **a**, la precisión predictiva en términos de R^2 para los modelos QSPR entrenados con representación URE y evaluados en Mn tuvo una caída de 25% (10% con *Disminución Abrupta* y 15% con *Disminución Moderada*) mientras que el 25% se mantiene *Invariante* y la mitad crece en desempeño (50% con *Incremento*). Se observó un comportamiento similar cuando se evaluaron estos modelos en Mw: 25% de caída (10% *Disminución Abrupta* y 15% *Disminución Moderada*), 30% permanece constante (*Invariante*) y 45% de mejora (*Incremento*). La mejora en ambos casos es superior a la esperada (color verde preponderante en Figura 5.4.a bloque horizontal Elongación a la Rotura). Esto merece una discusión

adicional, ya que los modelos QSPR de referencia (URE vs URE) presentados en Tabla 5.4 tienen un rendimiento *Intermedio* para esta propiedad, lo cual podría significar que superarlos no requeriría mucho esfuerzo. Más aun, aunque fueron superados por los modelos QSPR URE vs Mn y Mw, los rendimientos de estos no son buenos, entonces aunque el color verde denota una mejora, no implica que el resultado final es bueno. Además, existen algunos valores de correlación negativa, y en términos de errores, tanto MAE como RMSE, aumentan significativamente para la mayoría de los modelos QSPR URE vs Mn y Mw (Tabla A.5.2 de la sección Anexos del capítulo 5).

En cuanto a la pregunta **b**, al analizar la Figura 5.4.b puede observarse que los rendimientos caen con respecto a los observados en la Figura 5.4.a (color preponderante anaranjado). Para Mn/Mw, solo el 10%/15% de los modelos tienen una *Disminución Abrupta*, mientras que 40%/45% presentan *Disminución Moderada*, 20%/10% se mantiene con rendimiento *Invariante,* y 30%/30% presentan *Incrementos*. Sin embargo, los errores bajan considerablemente (Tabla A.5.2 de la sección Anexos del capítulo 5), por lo que aún podría considerarse una mejor práctica testear los modelos QSPR con aquellos modelos moleculares para los que fueron entrenados.

Resistencia a la Rotura:

En lo referido a la pregunta **a**, para los modelos entrenados con URE pero validados externamente con los modelos moleculares correspondientes a Mn se observa, en términos de R^2, una caída de 55% (40% con *Disminución Abrupta* y 15% con *Disminución Moderada*). El 35% de los modelos QSPR se mantiene *Invariante* y solo presenta *Incremento* el 10%. En cuanto a los modelos QSPR validados externamente con los modelos moleculares correspondientes a Mw: el 55% presenta caída (35% *Disminución Abrupta* y 20% *Disminución Moderada*), el 35% permanece constante (*Invariante*) y el 10% mejora (*Incremento*). Visualmente queda comprobada la preponderancia de los colores rojo y amarillo denotando *Disminución Abrupta* e *Invariante* para Mn y Mw (Figura 5.5a), es decir que el color verde es minoritario. A su vez, los errores MAE y RMSE, evaluados en esta experimentación (Tabla A.5.3 de la sección Anexos del capítulo 5), aumentan y especialmente lo hacen para aquellos modelos entrenados a partir de una Regresión Lineal (LR).

En cuanto a la pregunta **b**, para Mn/Mw, solo el 15%/5% de los modelos tienen una *Disminución Abrupta*, el 25%/25% presentan *Disminución Moderada*, mientras que el 15%/25% se mantiene con rendimiento *Invariante,* y finalmente el

45%/45% presentan *Incrementos*. Visualmente se comprueba todo esto con la preponderancia del color verde en Figura 5.4.b. En cuanto a los errores, tienen el mismo comportamiento que los modelos inferidos para responder la pregunta **a**, para esta misma propiedad en estudios (Tabla A.5.3 de la sección Anexos del capítulo 5).

Conclusiones sobre preguntas a y b:

Concluyendo, los modelos QSPR generados a partir de modelos moleculares URE han demostrado no ser precisos cuando se aplican sobre cadenas poliméricas de alto peso molecular. Además, a la luz de estos experimentos, tampoco resulta recomendable emplear DMs seleccionados desde experimentos con modelos moleculares URE en modelado QSPR de cadenas poliméricas de alto peso molecular. Sin embargo, este primer intento de usar modelos moleculares representativos de los pesos moleculares promedios de polímeros de alto peso molecular para inferir modelos QSPR nos brindó la base para el planteo de nuevos modelos moleculares univaluados. En tal sentido, esta primera experimentación muestra que los modelos entrenados y testeados para la misma instancia de peso funcionan mejor que aquellos que son testeados con pesos mayores. Estos resultados sugieren que modelos QSPR entrenados con modelos moleculares más representativos (Mn y Mw) podrían ser más robustos y mejorar la calidad predictiva.

5.3. PRIMERA PROPUESTA ALTERNATIVA

La precisión y la confiabilidad de los modelos generados por aprendizaje maquinal dependen, en gran medida, de la forma en que se representan las configuraciones atómicas, es decir, la elección de los descriptores utilizados como entrada [Imbalzano *et al.*, 2018]. Suele decirse que en el aprendizaje maquinal, el modelo es tan bueno como lo son los datos con los que se lo entrena. En Informática de Materiales, las características (o descriptores) que capturan los patrones en la estructura, la química o la unión de una composición química dada, son cruciales [Balachandran *et al.*, 2018]. Como ya se discutió antes, para los materiales poliméricos es clave representar adecuadamente la conectividad de las UREs y alcanzar pesos relativos al peso real del material (pesos moleculares promedios).

La selección de características es de suma importancia para cualquier algoritmo de aprendizaje maquinal. Cuando el conjunto de características seleccionado es deficiente puede generar problemas asociados a información incompleta, características irrelevantes o ruidosas, que luego harán que el

algoritmo de aprendizaje utilizado pueda experimentar bajas en las precisiones de predicción debido al aprendizaje de información irrelevante [Piramuthu, 2004]. Para evitar este tipo de problemas, se plantea la Primera Propuesta Alternativa al uso de la URE como modelo molecular de los polímeros que consiste en usar cadenas poliméricas que alcancen los pesos promedios de los polímeros.

Esta Primera Propuesta alternativa surge de las dificultades observadas y no resueltas del uso de la URE como modelo molecular estudiadas en la sección anterior, y busca responder a una Tercera Pregunta de Investigación que se desprende como una proyección o suerte de adecuación de la Segunda Pregunta de Investigación. De este modo la Tercera Pregunta de Investigación es: ¿los modelos moleculares Mn y Mw pueden resultar más efectivos en la caracterización de materiales poliméricos para modelado QSPR que el modelo molecular URE? Para responder esta pregunta se siguió la misma metodología explicada en 4 etapas en la Subsección 5.2.1, pero intercambiando las bases de datos, como se muestra en Figura 5.5.

FIGURA 5. 5. ESQUEMA GRÁFICO DE LA INTEGRACIÓN DE LA SEGUNDA CON LA TERCERA PREGUNTA DE INVESTIGACIÓN

En otras palabras, a diferencia de los experimentos propuestos para responder a la Segunda Pregunta de Investigación, aquí se plantea la necesidad de atravesar también un proceso de Selección de Características empleando los modelos moleculares pertenecientes a Mn y Mw. En contraste con la Etapa 2-A de Subsección 5.2.1, donde la Selección de característica se hizo empleando los modelos moleculares pertenecientes a URE. Hasta donde sabemos, este es uno de los primeros intentos de investigar la posibilidad de predecir propiedades mecánicas derivadas del ensayo de tensión para polímeros de alto peso molecular mediante el uso de técnicas de relación cuantitativa estructura-propiedad (QSPR) considerando modelos moleculares basados en el peso molecular promedio real de polímeros.

5.3.1. REPRESENTACIÓN COMPUTACIONAL BASADA EN PESOS MOLECULARES PROMEDIOS

Análogamente a la segunda pregunta de investigación (**a** y **b**), para esta nueva propuesta se plantea la Tercera Pregunta de Investigación compuesta con 4 interrogantes o subpreguntas: que siguen el orden alfabético anterior para evitar confusión:

- **c).** ¿Son los modelos QSPR basados en el modelo molecular Mn efectivos cuando se testean sobre modelos moleculares de otro peso (URE y Mw)?
- **d).** Los descriptores moleculares que fueron seleccionados en los modelos QSPR basados en el modelo molecular Mn ¿pueden resultar de utilidad para inferir nuevos modelos QSPR a partir de bases de datos de otras instancias univaluadas de representación (URE y Mw)?
- **e).** ¿Son los modelos QSPR basados en el modelo molecular Mw efectivos cuando se testean sobre modelos moleculares de otro peso (URE y Mn)?
- **f).** Los descriptores moleculares que fueron seleccionados en los modelos QSPR basados en el modelo molecular Mw ¿pueden resultar de utilidad para inferir nuevos modelos QSPR a partir de bases de datos de otras instancias univaluadas de representación (URE y Mn)?

5.3.1.a. MODELADO QSPR PARA PESO PROMEDIO EN NÚMERO (MN)

Para responder las preguntas **c** y **d**, se debió realizar primero la Selección de Características y luego el modelado QSPR (análogo a Etapas 2-A y 2-B). En la Tabla 5.5 se presentan los descriptores seleccionados por cada uno de los 5 métodos empleados: *CfsSubsetEval* (Cfs), *Wrapper Linear Regression* (WLR), *Wrapper Neural Network* (WNN), *Wrapper Random Forest* (WRF) y *Wrapper Random Committee* (WRC), para el modelo molecular Mn para cada una de las propiedades: Módulo de Tensión, Elongación a la Rotura y Resistencia a la Rotura.

TABLA 5. 5. DESCRIPTORES SELECCIONADOS PARA EL MODELO MOLECULAR BASADO EN Mn.

	Método	Cardinalidad	Descriptores Moleculares para Mn
Módulo de Tensión	Cfs	4	tpsaEfficiency_Mm, khs.dsssP_Mm, khs.ssS_Mm, HybRatio_Mn
	WLR	9	nSmallRings_Mm, tpsaEfficiency_Mm, khs.ssCH2_Mm, khs.aaaC_Mm, khs.ssNH_Mm, khs.dsssP_Mm, khs.ssS_Mm, nHBDon_Mm, C1SP3_Mm
	WRC	5	nAtomLAC_Mm, khs.sssCH_Mm, khs.ssNH_Mm, khs.aaO_Mm, khs.ssS_Mm

Elongación a la Rotura	WRF	11	nSmallRings_Mm, khs.aaCH_Mm, khs.ssNH_Mm, khs.aasN_Mm, khs.dO_Mm, khs.dsssP_Mm, khs.ddssS_Mm, nHBDon_Mm, C2SP2_Mm, C3SP3_Mm, nAcid_Mm
	WNN	7	nRings5_Mm, tpsaEfficiency_Mm, khs.sOH_Mm, khs.dsssP_Mm, khs.ssS_Mm, HybRatio_Mm, C2SP2_Mm
	Cfs	4	tpsaEfficiency_Mm, VAdjMat_Mm, khs.dsssP_Mm, HybRatio_Mm
	WLR	3	khs.ssO_Mm, khs.dsssP_Mm, C3SP2_Mm
	WRC	2	nAtomP_Mm, C3SP2_Mm
	WRF	8	nRings4_Mm, nRings7_Mm, nAtomLAC_Mm, khs.dsCH_Mm, khs.ssS_Mm, khs.ddssS_Mm, nHBDon_Mm, C3SP2_Mn
	WNN	8	tpsaEfficiency_Mm, Zagreb_Mm, khs.aaaC_Mm, khs.sssN_Mm, khs.dsssP_Mm, khs.ssS_Mm, HybRatio_Mm, C1SP2_Mm
Resistencia a la Rotura	Cfs	3	khs.sssN_Mm, khs.aaO_Mm, khs.ssS_Mm, Kier3_Mm
	WLR	7	nRings6_Mm, nAtomP_Mm, khs.aaCH_Mm, khs.aasN_Mm, khs.aaO_Mm, khs.ssS_Mm, C3SP2_Mm
	WRC	6	nAtomLAC_Mm, khs.ssCH2_Mm, khs.sssN_Mm, khs.dsssP_Mm, nHBDon_Mm, apol_Mm
	WRF	11	nAromBlocks_Mm, nRings4_Mm, nRings5_Mm, nRings6_Mm, Zagreb_Mm, nAtomLAC_Mm, khs.sssN_Mm, khs.ssO_Mm, khs.ssS_Mm, nHBDon_Mm, C3SP3_Mm
	WNN	7	nRings5_Mm, khs.ssCH2_Mm, khs.ssNH_Mm, khs.ssO_Mm, khs.ssS_Mm, HybRatio_Mm, nHBDon_Mm

Estos subconjuntos fueron utilizados para entrenar 20 modelos con el modelo molecular Mn usando los mismos 4 métodos de aprendizaje maquinal que en la sección 5.2.1*: Linear Regression* (LR), *Neural Network* (NN), *Random Forest* (RF) y *Random Committee* (RC). Los resultados para la validación externa se presentan en términos de R^2, MAE y RMSE en la Tabla 5.6. Para clasificar los modelos se utiliza el mismo criterio, según R^2, que en la Tabla 5.4: *Bajo*: $x < 0.6$; *Intermedio*: $0.6 \leq x < 0.80$; *Alto*: $x \geq 0.80$.

TABLA 5.6 RENDIMIENTOS DE LA VALIDACIÓN EXTERNA PARA Mn DE LOS MODELOS ENTRENADOS CON Mn A PARTIR DE LOS DMS SELECCIONADOS PARA Mn.

Mn vs. Mn

	Módulo de Tensión				**Elongación a la Rotura**				**Resistencia a la Rotura**			
	Cfs				Cfs				Cfs			
	LR	NN	RF	RC	LR	NN	RF	RC	LR	NN	RF	RC
R^2	0.74	0.85	0.94	0.92	-0.16	0.64	0.76	0.78	0.74	0.83	0.91	0.87
MAE	0.37	0.45	0.22	0.23	2.75	1.63	1.12	0.98	10.58	8.58	6.96	8.54
RMSE	0.50	0.54	0.32	0.33	3.19	2.01	1.34	1.36	13.33	10.91	7.78	10.17
	WLR				WLR				WLR			
	LR	NN	RF	RC	LR	NN	RF	RC	LR	NN	RF	RC
R^2	0.73	0.79	0.98	0.98	-0.48	-0.52	0.66	0.47	0.77	0.81	0.94	0.93
MAE	0.35	0.42	0.15	0.14	3.05	2.65	1.25	1.57	10.75	9.56	5.66	6.80
RMSE	0.52	0.57	0.19	0.16	3.97	3.28	1.75	2.89	11.85	11.26	6.45	7.81
	WNN				WNN				WNN			
	LR	NN	RF	RC	LR	NN	RF	RC	LR	NN	RF	RC

R^2	0.00	0.85	0.98	0.98	0.25	0.76	0.76	0.81	0.85	0.92	0.92	0.89
MAE	0.35	0.28	0.16	0.13	3.21	1.09	1.05	0.88	11.89	10.30	5.85	6.98
RMSE	0.69	0.44	0.21	0.17	4.00	1.53	1.33	1.25	13.00	11.68	7.16	8.56
	WRF				WRF				WRF			
	LR	NN	RF	RC	LR	NN	RF	RC	LR	NN	RF	RC
R^2	0.00	0.87	0.96	0.98	-0.48	-0.41	0.67	0.71	0.79	0.88	0.95	0.93
MAE	0.37	0.26	0.16	0.12	3.94	2.28	1.23	1.15	10.99	10.44	4.93	6.09
RMSE	0.71	0.34	0.22	0.16	4.98	2.99	1.51	1.56	12.62	12.61	5.99	7.05
	WRC				WRC				WRC			
	LR	NN	RF	RC	LR	NN	RF	RC	LR	NN	RF	RC
R^2	0.47	0.98	0.98	0.98	-0.37	-0.56	0.74	0.79	0.77	0.88	0.93	0.94
MAE	0.40	0.15	0.15	0.11	3.32	2.78	1.12	0.81	12.76	12.40	6.42	5.78
RMSE	0.62	0.21	0.20	0.16	4.13	3.72	1.40	1.40	14.70	15.49	7.07	6.62

Módulo de Tensión:

Como se puede observar en Tabla 5.6, los rendimientos alcanzados para los modelos basados en el modelo molecular Mn son: *Alto* (≥ 0.80) para todos los subconjuntos de DMs entrenados con NN, RF y RC, excepto por una diferencia de 0.01 para el subconjunto de WLR entrenado con NN que alcanza un R^2 = 0.79. El 73.3% de estos modelos supera un R^2 = 0.90. Para los subconjuntos entrenados con LR no sucede lo mismo. Cuando los subconjuntos son seleccionados con WNN y WRF obtienen R^2 = 0, aquel seleccionado con WRC logra mejorar, pero no lo suficiente con un R^2 = 0.47 y sí alcanzan rendimientos de categoría *Intermedio* cuando son seleccionados por Cfs y WLR con R^2 de 0.74 y 0.73, respectivamente. Existe una tendencia marcada en aquellos modelos entrenados con LR a obtener rendimientos menores que para los entrenados con los otros tres métodos de aprendizaje maquinal.

Elongación a la Rotura:

Independientemente del método utilizado para la selección de DMs, los modelos entrenados con LR tienen rendimiento *Bajo* en términos de R^2 (< 0.6). Aquellos entrenados con RF tienen un rendimiento *Intermedio* que varían en este caso para R^2 entre 0.66 y 0.76. El comportamiento para los modelos entrenados con RC es más complejo de analizar ya que tiene tanto rendimientos *Bajo* (DMs seleccionados con WLR), como *Alto* (DMs seleccionados con WNN). En términos generales, los modelos obtenidos tienen un rendimiento *Intermedio* ya que solo el 10% superó R^2=0.8 y el 45% está por debajo de R^2=0.36. Además, existen varias correlaciones negativas (30% de los experimentos) todas presentes en los modelos entrenados con LR y NN.

Resistencia a la Rotura:

Los modelos entrenados con modelos moleculares Mn para Resistencia a la Rotura tienen rendimientos por encima de R^2 =0.70. Exceptuando los modelos entrenados con LR, los demás alcanzan rendimientos *Alto* (R^2 < 0.8). En términos generales, el 95 % de los modelos QSPR alcanza un R^2 de 0.75, un 80% lo hace para R^2 = 0.8 y el 45% supera un R^2 de 0.9.

5.3.1.b. RESPONDIENDO A LA TERCERA PREGUNTA DE INVESTIGACIÓN (PRIMERA PARTE)

En esta sección, se presentan los resultados de la experimentación llevada a cabo para responder las preguntas que fueron identificadas como **c** y **d** (Figura 5.6). En primer lugar, para responder a **c**, se utilizaron los modelos QSPR entrenados para el modelo molecular Mn y se realizó una validación externa utilizando por un lado el modelo molecular URE y por el otro, el modelo molecular Mw (análogo a la Etapa 3, Sección 5.2). Luego, para responder a **d**, se recuperaron las características seleccionadas desde la base de datos correspondiente (URE o Mw) para entrenar y validar los modelos QSPR con cada uno de los modelos moleculares (URE o Mw), análogo a la Etapa 4, Sección 5.2. El criterio utilizado para el análisis de resultados es el mismo que en la Sección 5.2.2.: *Abrupta Disminución* (Rojo) cuando R^2 disminuye más del 40%, *Disminución Moderada* (Anaranjado) cuando la disminución de R^2 varía entre 5%-40%, *Invariante* (Amarillo) si R^2 varía entre ± 5% e *Incremento* (Verde) si R^2 aumenta más del 5%. Nuevamente en el tipo de figura como la 5.7, con elementos de colores, debe leerse en primer lugar las dos columnas de la izquierda y posteriormente las dos de la derecha. Las propiedades mecánicas estudiadas se presentan en los bloques horizontales.

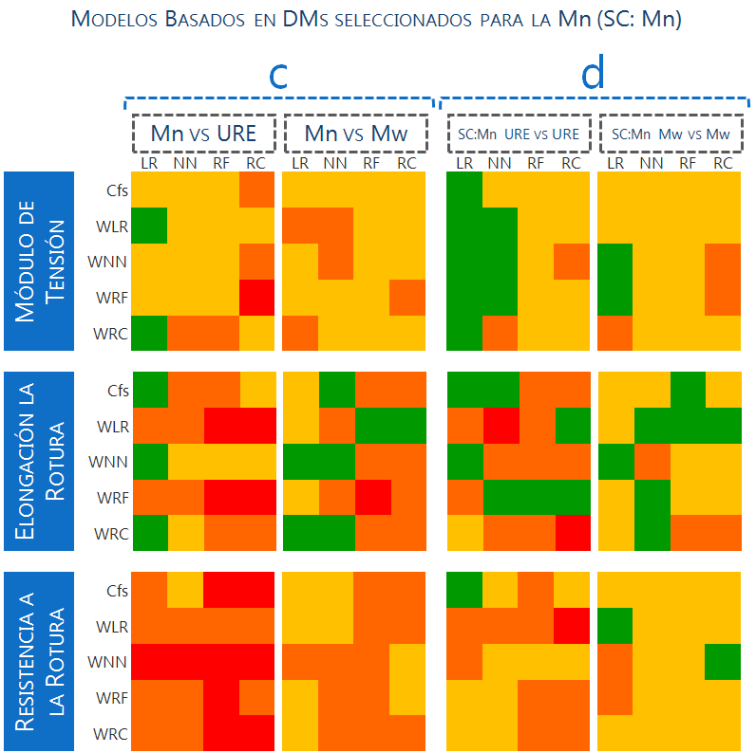


FIGURA 5. 6. RESUMEN GRÁFICO DE LA CALIDAD EN LA PRECISIÓN ALCANZADA POR LOS MODELOS BASADO EN LOS DMS SELECCIONADOS PARA MN PARA RESPONDER LAS PREGUNTAS **c** Y **d**.

Módulo de Tensión:

En lo referido a la pregunta **c**, para los modelos QSPR entrenados con el modelo molecular Mn, pero validados externamente con el modelo molecular correspondiente a URE se observa, en términos de R^2, una caída de 30% (10% con *Disminución Abrupta* y 20% con *Disminución Moderada*). Se mantiene *Invariante* el 60% de los modelos QSPR y solo el 10% presenta *Incremento*. En cuanto a los modelos QSPR validados externamente con los modelos moleculares Mw, en términos de R^2, el 25% presentan caída (0% *Disminución Abrupta* y 25% *Disminución Moderada*), el 75% permanece constante (*Invariante*) y ninguno presenta mejora (0% *Incremento*). Estas conclusiones quedan evidenciadas visualmente en la Figura 5.6.c, ya que el color amarillo es preponderante en las dos columnas de la izquierda, para esta propiedad. Los errores MAE y RMSE (Tabla A.5.4 de la sección Anexos del capítulo 5), evaluados en esta experimentación se mantienen estables con respectos a los modelos de referencia (Tabla 5.6).

En cuanto a la pregunta **d**, para URE/Mw, ninguno de los modelos QSPR tienen una *Disminución Abrupta* para ambos modelos moleculares (0%/0%), solo el 10%/15% presentan *Disminución Moderada*, el 50%/75% se mantiene con rendimiento *Invariante,* mientras que el 40%/10% presentan *Incremento* (color

amarillo preponderante en Figura 5.6.d). En cuanto a los errores, tienen el mismo comportamiento que los modelos inferidos para responder la pregunta **c** en relación al Módulo de Tensión (Tabla A.5.4 de la sección Anexos del capítulo 5).

Elongación a la Rotura:

En lo referido a la pregunta **c** (columnas de la izquierda en Figura 5.6), para los modelos QSPR entrenados con el modelo molecular Mn, pero validados externamente con los modelos moleculares correspondientes a URE para Elongación a la Rotura, se observa en términos de R^2, una caída de 60% (20% con *Disminución Abrupta* y 40% con *Disminución Moderada*). El 25% de los modelos QSPR se mantiene *Invariante* y el 15% presenta *Incremento* (preponderancia de color anaranjado en la primera columna de la izquierda de Figura 5.6.c). En cuanto a los modelos QSPR validados externamente con los modelos moleculares Mw, en términos de R^2, el 50% presentan caída (5% *Disminución Abrupta* y 45% *Disminución Moderada*), el 15% permanece constante (*Invariante*) y el 35% presenta *Incremento* (preponderancia de color anaranjado en la segunda columna de la izquierda de Figur 5.7c). Los errores MAE y RMSE (Tabla A.5.5 de la sección Anexos del capítulo 5), evaluados en esta experimentación aumentan significativamente con respectos a los modelos de referencia (Tabla 5.6).

En cuanto a la pregunta **d**, para los modelos moleculares URE/Mw, el 10%/0% de los modelos QSPR tienen una *Disminución Abrupta*, el 50%/15% presentan *Disminución Moderada*, mientras que el 5%/50% se mantiene con rendimiento *Invariante,* finalmente el 35%/35% presentan *Incremento* (preponderancia de color anaranjado/amarillo en las columnas de la derecha de Figura 5.6.d). Tanto MAE como RMSE (Tabla A.5.5 de la sección Anexos del capítulo 5) aumentan notablemente en comparación con el modelo de referencia de la Tabla 5.6.

Resistencia a la Rotura:

Para Resistencia a la Rotura, en lo referido a la pregunta **c**, para los modelos QSPR entrenados con el modelo molecular Mn, pero validados externamente con los modelos moleculares correspondientes a URE, se observa en términos de R^2: una caída de 95% (45% con *Disminución Abrupta* y 50% con *Disminución Moderada*). El restante 5% de los modelos QSPR se mantiene *Invariante* y ninguno (0%) presenta *Incremento* (preponderancia de colores rojo y anaranjado en la primera columna de la izquierda de Figura 5.6.c). En cuanto a los modelos QSPR validados externamente con los modelos moleculares Mw, en términos de R^2: el 60% presentan caída (0% *Disminución Abrupta* y 60% *Disminución Moderada*), el

40% permanece constante (*Invariante*) y nuevamente ninguno (0%) presenta *Incremento* (preponderancia de color anaranjado en la segunda columna de la izquierda de Figura 5.6.c). Los errores MAE y RMSE (Tabla A.5.6 de la sección Anexos del capítulo 5) aumentan con respectos a los modelos de referencia (Tabla 5.6).

En cuanto a la pregunta **d**, para los modelos moleculares URE/Mw, el 5%/0% de los modelos QSPR tienen una *Disminución Abrupta*, el 45%/10% presentan *Disminución Moderada*, mientras que el 45%/80% se mantiene con rendimiento *Invariante,* finalmente el 5%/10% presentan *Incremento* (preponderancia de colores anaranjado y amarillo/amarillo en las columnas de la derecha de Fig. 5.7d). Tanto MAE como RMSE (Tabla A.5.6 de la sección Anexos del capítulo 5) aumentan considerablemente en comparación con el modelo de referencia de la Tabla 5.6.

Conclusiones sobre preguntas c y d:

Se concluye, una vez más como se esperaba, que los modelos QSPR entrenados y evaluados en la misma instancia de peso (pregunta **d**) tienen un mejor rendimiento que cuando se utiliza el modelo solo para testear (pregunta **c**), lo que queda evidenciado mirando los colores de la Figura 5.6.c vs 5.6.d. Comparando los resultados de la pregunta **a** (basada en URE) vs la **c** (basada en Mn), podemos concluir que hay menos disminuciones de rendimiento para cuando trabajamos con pesos más realistas, y esto es coherente con la hipótesis planteada. Mn parecería ser un modelo molecular más representativo para los polímeros de alto peso molecular según nuestras condiciones de experimentación (base de datos y tipo de propiedades). Es notorio que los modelos generados aquí funcionan mejor cuando se testean con Mw que cuando lo hacen con URE. Esto también era una hipótesis de trabajo, ya que tanto numérica como fisicoquímicamente estas instancias de peso (Mn y Mw) están más relacionadas semánticamente entre ellas que con la URE.

5.3.1.c. MODELADO QSPR PARA PESO PROMEDIO EN PESO (MW)

En la Tabla 5. 7 se presentan los descriptores seleccionados para el último modelo molecular propuesto (Mw) como primera alternativa al uso del modelo molecular URE (análogo a la Etapa 2-A). Se emplearon los mismos 5 métodos que en los casos anteriores para poder hacer una justa comparación: CfsSubsetEval (Cfs), Wrapper Linear Regression (WLR), Wrapper Neural Network (WNN), Wrapper Random Forest (WRF) y Wrapper Random Committee (WRC). Siguiendo la

metodología descripta anteriormente, los 5 métodos se aplicaron para las tres propiedades estudiadas.

TABLA 5. 7. DESCRIPTORES SELECCIONADOS PARA EL MODELO MOLECULAR BASADO EN MW.

	Método	Cardinalidad	Descriptores Moleculares para Mw
Módulo de Tensión	Cfs	6	tpsaEfficiency_Mw, khs.sssCH_Mw, khs.dsssP_Mw, khs.ssS_Mw, nHBDon_Mw, C1SP3_Mw
	WLR	12	nRings4_Mw, tpsaEfficiency_Mw, MW_Mw, MLogP_Mw, khs.aaaC_Mw, khs.ssNH_Mw, khs.dsssP_Mw, khs.ssS_Mw, HybRatio_Mw, C1SP3_Mw, nB_Mw, nAtom_Mw
	WRC	6	nAtomLAC_Mw, khs.ssNH_Mw, khs.aaO_Mw, khs.ssS_Mw, khs.ddssS_Mw, C3SP3_Mw
	WRF	7	tpsaEfficiency_Mw, khs.sssCH_Mw, khs.dsssP_Mw, khs.ssS_Mw, khs.ddssS_Mw, HybRatio_Mw, C3SP3_Mw
	WNN	8	nAromBlocks_Mw, nRings5_Mw, khs.aaCH_Mw, khs.ssNH_Mw, khs.ssS_Mw, C1SP2_Mw, C3SP2_Mw, C1SP3_Mw
Elongación a la Rotura	Cfs	6	nRingBlocks_Mw, tpsaEfficiency_Mw, nAtomP_Mw, khs.dsssP_Mw, khs.ssS_Mw, C3SP2_Mw
	WLR	6	tpsaEfficiency_Mw, nAtomP_Mw, khs.aasN_Mw, khs.sOH_Mw, nHBDon_Mw, nAcid_Mw
	WRC	6	nAtomP_Mw, khs.dsCH_Mw, nHBAcc_Mw, C3SP2_Mw, apol_Mw, ALogP_Mw
	WRF	2	nAtomP_Mw, C3SP2_Mw
	WNN	11	tpsaEfficiency_Mw, VAdjMat_Mw, MLogP_Mw, nAtomP_Mw, khs.dsCH_Mw, khs.ssNH_Mw, khs.dsssP_Mw, Kier1_Mw, HybRatio_Mw, ALogP_Mw, nAcid_Mw
Resistencia a la Rotura	Cfs	5	nRings5_Mw, nAtomLAC_Mw, khs.sssN_Mw, khs.aaO_Mw, khs.ssS_Mw
	WLR	4	khs.sssN_Mw, khs.aaO_Mw, khs.ssS_Mw, ALogp2_Mw
	WRC	12	nRings4_Mw, nRings7_Mw, nAtomLAC_Mw, khs.sssCH_Mw, khs.dssC_Mw, khs.sssN_Mw, khs.sOH_Mw, khs.ssS_Mw, khs.ddssS_Mw, C2SP3_Mw, C3SP3_Mw, nAcid_Mw
	WRF	8	nRings5_Mw, nAtomLAC_Mw, khs.sssN_Mw, khs.sOH_Mw, khs.ssS_Mw, khs.ddssS_Mw, nHBAcc_Mw, C3SP3_Mw
	WNN	8	nRings5_Mw, nAtomP_Mw, khs.sOH_Mw, khs.ssS_Mw, khs.ddssS_Mw, HybRatio_Mw, nHBAcc_Mw, nAcid_Mw

Para cada una de las propiedades, se entrenó cada uno de los subconjuntos con 4 métodos de aprendizaje maquinal: Linear Regression (LR), Neural Network (NN), Random Forest (RF) y Random Committee (RC). De este paso se obtuvieron 20 modelos QSPR por propiedad (análogo a la Etapa 2-B). Los resultados para la validación externa se presentan en términos de R^2, MAE y RMSE en la Tabla 5.8 y se utiliza el mismo criterio para clasificar los modelos, según R^2, que en los casos anteriores: Bajo: $x < 0.6$; Intermedio: $0.6 \leq x < 0.80$; Alto: $x \geq 0.80$.

TABLA 5. 8 RENDIMIENTOS DE LA VALIDACIÓN EXTERNA PARA MW DE LOS MODELOS ENTRENADOS CON MW A PARTIR DE LOS DMS SELECCIONADOS PARA MW.

Mw vs. Mw

	Módulo de Tensión				**Elongación a la Rotura**				**Resistencia a la Rotura**			
	Cfs				Cfs				Cfs			
	LR	NN	RF	RC	LR	NN	RF	RC	LR	NN	RF	RC
R2	0.65	0.90	0.98	0.98	0.21	-0.07	0.72	0.74	0.76	0.90	0.89	0.88
MAE	0.36	0.52	0.15	0.13	2.78	1.89	1.02	1.10	11.11	7.20	6.78	7.46
RMSE	0.51	0.60	0.20	0.16	3.07	2.41	1.48	1.52	12.13	8.66	8.44	8.80
	WLR				WLR				WLR			
	LR	NN	RF	RC	LR	NN	RF	RC	LR	NN	RF	RC
R^2	0.63	0.82	0.92	0.94	0.87	0.64	0.62	0.58	0.76	0.82	0.92	0.86
MAE	0.37	0.53	0.20	0.18	1.74	2.01	1.35	1.39	11.05	8.61	6.33	7.80
RMSE	0.65	0.68	0.27	0.22	2.11	2.85	1.64	1.86	13.03	10.48	7.58	9.70
	WNN				WNN				WNN			
	LR	NN	RF	RC	LR	NN	RF	RC	LR	NN	RF	RC
R^2	0.39	0.89	0.93	0.89	0.58	0.61	0.61	0.75	0.82	0.85	0.92	0.90
MAE	0.43	0.33	0.21	0.25	1.91	1.66	1.45	1.63	12.92	8.66	7.04	7.65
RMSE	0.74	0.39	0.26	0.31	2.40	2.27	2.11	2.15	13.93	10.85	8.05	8.86
	WRF				WRF				WRF			
	LR	NN	RF	RC	LR	NN	RF	RC	LR	NN	RF	RC
R^2	0.87	0.89	0.98	0.97	-0.41	-0.65	0.62	0.66	0.90	0.86	0.94	0.94
MAE	0.31	0.42	0.16	0.15	3.26	2.57	1.22	1.18	11.38	8.63	6.67	6.34
RMSE	0.41	0.48	0.20	0.18	4.03	3.54	1.64	1.72	12.58	11.24	8.01	7.28
	WRC				WRC				WRC			
	LR	NN	RF	RC	LR	NN	RF	RC	LR	NN	RF	RC
R^2	0.30	0.97	0.97	0.96	-0.41	0.09	0.52	0.31	0.58	0.88	0.91	0.81
MAE	0.38	0.16	0.16	0.16	3.26	3.73	1.83	3.12	13.28	7.76	6.65	8.70
RMSE	0.70	0.18	0.21	0.21	4.03	5.67	2.75	5.42	16.68	9.54	8.04	11.95

Módulo de Tensión:

Los resultados del rendimiento estadístico de los modelos para Módulo de Tensión entrenados y evaluados externamente en el modelo molecular correspondiente a Mw pueden verse en la Tabla 5.8. En términos generales son modelos con alto rendimiento. El 80% de ellos alcanza un R^2 mayor a 0.8 y sólo el 10% está por debajo de 0.5. Los modelos entrenados mediante Regresión Linear (LR) tienen menor rendimiento independientemente del conjunto de DMs del que se parta, cuando se los compara con los otros métodos de aprendizaje maquinal. Los errores expresados tanto en MAE como RMSE no se modifican en relación a los reportados para las instancias de URE (Tabla 5.4) y Mn (Tabla 5.6).

Elongación a la rotura:

Los modelos para Elongación la Rotura presentados en la Tabla 5.8 tienen un rendimiento intermedio. Solo el 5% de los modelos alcanza un R^2 = 0.80. El 55% de

los modelos QSPR tiene un rendimiento; *Intermedio*: $0.6 \leq R^2 < 0.80$. Finalmente, el restante 40% no supera un R^2=0.6. Independientemente del método de aprendizaje maquinal utilizado para su entrenamiento todos los modelos QSPR inferidos a partir del subconjunto de DMs seleccionado con WRC tiene rendimiento *Bajo*. Los errores (MAE y RMSE) bajan con respecto a los modelos entrenados y evaluados para el modelo molecular correspondiente a Mn (Tabla 5.6), pero aumentan con respecto al de URE (Tabla 5.4).

Resistencia a la Rotura:

El rendimiento para los modelos de Resistencia a la Rotura entrenados y evaluados en la Mw es alto. El 85% de los modelos supera un R^2 = 0.80. Es decir, en términos generales los modelos tienen rendimientos altos. Solo el 5%, es decir un único modelo, tiene un rendimiento *Bajo*, aunque es superior a 0.5 (valor random). Al igual que para los modelos de Elongación a la Rotura, los errores tanto MAE como RMSE bajan con respecto a Mn (Tabla 5.6) y aumentan en cuanto a URE (Tabla 5.4).

5.3.1.d. RESPONDIENDO A LA TERCERA PREGUNTA DE INVESTIGACIÓN (SEGUNDA PARTE)

En la Figura 5.7 se presentan de forma gráfica los resultados de los experimentos que responden a las preguntas **e** y **f**. Del mismo modo que en las Figuras 5.4 y 5.6 los colores representan el comportamiento de los rendimientos de los modelos en términos de R^2 en comparación con los modelos QSPR de referencias (Tabla 5.8). Se espera que el comportamiento del modelo molecular Mw sea más similar al de Mn que al de URE. Sin embargo, es válido experimentar con ambos modelos moleculares, es decir, para ambos pesos moleculares ya que a la polidispersión de pesos moleculares es necesario caracterizarla por más de un representante de peso. Aunque, se esperan algunas diferencias ya que el Peso Promedio en Número (Mn) es sensible a la mezcla de moléculas de baja masa molecular y el Peso Promedio en Peso (Mw) lo es a las de gran masa molecular.

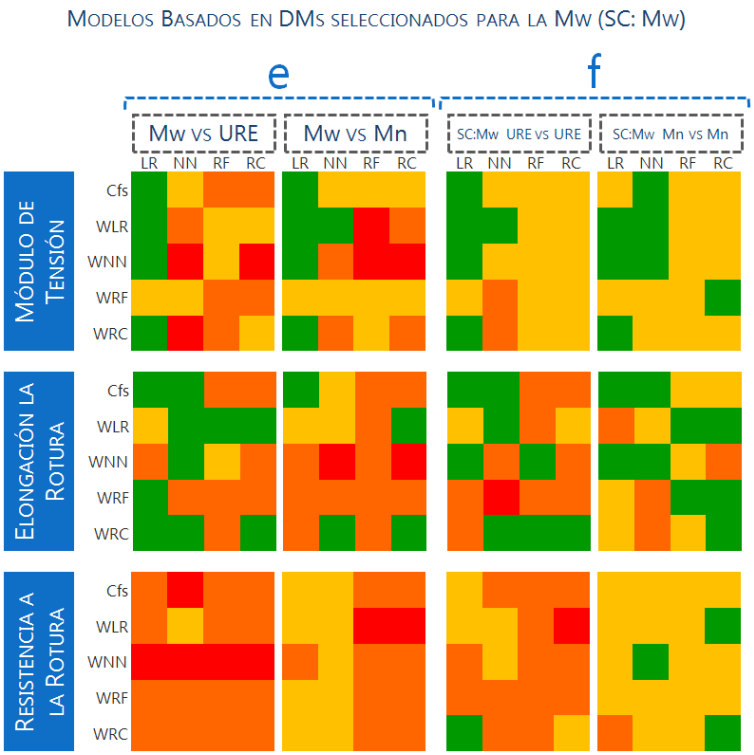


FIGURA 5. 7 RESUMEN GRÁFICO DE LA CALIDAD EN LA PRECISIÓN ALCANZADA POR LOS MODELOS BASADO EN LOS DMS SELECCIONADOS PARA MW PARA RESPONDER LAS PREGUNTAS **e** Y **f**.

Módulo de Tensión:

En lo referido a la pregunta **e**, la precisión predictiva, en términos de R^2, para los modelos QSPR entrenados con modelos moleculares Mw y evaluados en URE tuvieron una caída en el 45% los casos (15% con *Disminución Abrupta* y 30% con *Disminución Moderada*), mientras que el 35% se mantiene estable (*Invariante*) y finalmente el 20% crece en desempeño (*Incremento*). En la primera columna de la izquierda de la Figura 5.7.e se ve preponderancia de los colores anaranjado y amarillo. Se observó un comportamiento similar cuando se evaluaron estos modelos en Mn: 35% de caída (15% *Disminución Abrupta* y 20% *Disminución Moderada*), 40% permanece constante (*Invariante*) y 25% de aumento (*Incremento*), (preponderancia de color amarillo en la segunda columna de la izquierda de Fig. 5.8e). Además, en términos generales, los errores de los modelos QSPR se mantienen (Tabla A.5.7 de la sección Anexos del capítulo 5).

Respecto a la pregunta **f** los resultados de los modelos QSPR inferidos a partir de los descriptores seleccionados usando la representación Mw, pero que fueron entrenados y validados para polímeros representados con el modelo molecular URE y el modelo molecular Mn son los siguientes (URE/Mn): el 0%/0% de los modelos tienen una *Disminución Abrupta*, mientras que el 10%/0% solo presentan *Disminución Moderada*, el 65%/65% se mantiene con rendimiento *Invariante,* y el

25%/35% presentan *Incremento*. Para ambos modelos los errores evaluados, es decir, MAE y RMSE (Tabla A.5.7 de la sección Anexos del capítulo 5), se mantienen o tienden a bajar.

Elongación a la Rotura:

En lo referido a la pregunta **e**, la precisión predictiva, en términos de R^2, para los modelos QSPR entrenados con modelos moleculares Mw y evaluados en URE tuvieron una caída en el 40% los casos (0% con *Disminución Abrupta* y 40% con *Disminución Moderada*) mientras que solo el 10% se mantiene estable (*Invariante*) y finalmente la mitad crece en desempeño (50% con *Incremento*), preponderancia de color verde en la primera columna de la izquierda de Figura 5.7.e. Se observó un comportamiento diferente cuando se evaluaron estos modelos en Mn: 65% de caída (10% *Disminución Abrupta* y 55% *Disminución Moderada*), 20% permanece constante (*Invariante*) y el 15% presenta aumento (*Incremento*), preponderancia de color anaranjado en la segunda columna de la izquierda de Figura 5.7.e. Los errores evaluados tienden a aumentar, aunque solo mínimamente (Tabla A.5.8 de la sección Anexos del capítulo 5).

Respecto a la pregunta **f**, los resultados de los modelos QSPR inferidos a partir de los DMs seleccionados usando la representación Mw, pero que fueron entrenados y validados para polímeros representados con el modelo molecular URE y el modelo molecular Mn son los siguientes (URE/Mn): el 5%/0% de los modelos tienen una *Disminución Abrupta*, mientras que el 45%/20% presentan *Disminución Moderada*, el 10%/35% se mantiene con rendimiento *Invariante,* y finalmente el 40%/45% presentan *Incremento*, preponderancia de color anaranjado y verde/verde en la columna de la derecha (Figura 5.8f). Al igual que para los modelos entrenados para predecir el Módulo de Tensión, los errores se mantienen mayoritariamente constantes (Tabla A.5.8 de la sección Anexos del capítulo 5).

Resistencia a la Rotura:

En lo referido a la pregunta **e**, en términos de R^2 la precisión predictiva para los modelos QSPR entrenados con modelos moleculares Mw y evaluados en URE tuvieron una caída en el 95% los casos (45% con *Disminución Abrupta* y 50% con *Disminución Moderada*) mientras que el restante 5% se mantiene estable (*Invariante*) y finalmente ningún modelo QSPR crece en desempeño (0% con *Incremento*), preponderancia de colores rojo y anaranjado en la primera columna de la izquierda de Figura 5.7.e. Se observó un comportamiento similar cuando se evaluaron estos modelos QSPR en los modelos moleculares Mn: el 55% de caída

(10% *Disminución Abrupta* y 45% *Disminución Moderada*), el restante 45% permanece constante (*Invariante*) y ningún modelo QSPR presenta aumento (0% con *Incremento*), preponderancia de colores anaranjado y amarillo en la segunda columna de la izquierda de Figura 5.7.e. Como es esperable frente a estos resultados los errores (MAE y RMSE) aumentan (Tabla A.5.9 de la sección Anexos del capítulo 5).

Respecto a la pregunta **f**, los resultados en términos de R^2 para los modelos QSPR inferidos a partir de los DMs seleccionados para la representación Mw, pero que fueron entrenados y validados para polímeros representados con modelos moleculares URE/Mn son los siguientes: el 0%/0% de los modelos tienen una *Disminución Abrupta*, el 70%/5% presentan *Disminución Moderada*, mientras que el 25%/80% se mantiene con rendimiento *Invariante,* y finalmente el 5%/15% presentan *Incremento*, preponderancia de colores anaranjado/amarillo en la columna de la derecha (Figura 5.8f). En cuanto a los errores, ambos MAE y RMSE, se mantienen en términos generales (Tabla A.5.9 de la sección Anexos del capítulo 5).

Conclusiones sobre preguntas e y f:

Concluyendo, dados los resultados de este bloque de experimentación puede notarse nuevamente que aquellos modelos entrenados y validados usando el mismo modelo molecular tienen mejor rendimiento que aquellos que fueron entrenados para una instancia de peso y evaluados en otra. Por otro lado, es interesante notar que la hipótesis de los modelos generados para responder las preguntas **e** y **f** usando como testeo el Mn deberían tener un mejor rendimiento que aquellos usando la URE para evaluar, fue confirmada.

5.3.2. CONCLUSIONES SOBRE LA PRIMERA PROPUESTA ALTERNATIVA

Este capítulo está enfocado a responder lo que llamamos Segunda y Tercera Pregunta de Investigación en esta tesis, ambas están relacionadas con el modelado QSPR para materiales poliméricos que presentan polidispersidad, además de alto peso molecular. La Segunda Pregunta de Investigación está formada por dos interrogantes: por un lado, queríamos saber si los modelos QSPR, para predecir propiedades mecánicas, que están basados en representaciones simplificadas como la URE, eran útiles para predecir estas propiedades utilizando bases de datos de polímeros representados a partir de sus pesos promedios Mn o Mw (sustancialmente mayores al de la URE). Además, queríamos investigar si los subconjuntos de características seleccionados a partir de los modelos moleculares URE eran útiles para aprender el modelo QSPR, utilizando luego para predecir

sobre bases de datos valuadas en Mn o Mw. Finalmente, se evalúan los mismos interrogantes, en la Tercera Pregunta de Investigación, partiendo de instancias de peso mayores, Mn y Mw, con el objetivo de verificar si estas otras representaciones univaluadas de los polímeros resultan en modelos moleculares más informativos que permitan la generación de modelos QSPR más robustos.

Las respuestas a estos interrogantes revelaron que los modelos QSPR inferidos para una representación mínima o simplificada (URE) no son aconsejables para aplicarlos a los pesos promedio de un polímero polidisperso (representaciones mayores). Esta conclusión está en concordancia con la complejidad del campo de la Informática de Polímeros, en el cual solo representaciones breves han sido consideradas, debido a que los resultados obtenidos no son uniformes. Cuando se utilizan los modelos inferidos a partir de Mn para predecir propiedades mecánicas en otras instancias (URE y Mw) puede observarse que los resultados tampoco son uniformes, con muy pocos casos en los que el desmejoramiento es alto (mayoritariamente cuando se aplica en URE). Finalmente, cuando se utilizan los modelos aprendidos para Mw en las instancias menores de peso, el comportamiento es similar al caso anterior (Mn). Esto es esperable y podría demostrar que Mn y Mw son modelos moleculares más similares entre sí, de lo que ambos lo son con URE. Dicha similaridad podría deberse al hecho de que estos modelos moleculares (Mn y Mw) pueden captar alguna característica de la conectividad entre las UREs que el modelo molecular URE no puede por estar compuesto por un único bloque estructural.

Además, los resultados han demostrado que el rendimiento de los modelos QSPR cuando son entrenados y testeados para el mismo modelo molecular, mejora con respecto a aquellos modelos entrenados y evaluados en distintas instancias de peso. La intención aquí fue investigar si las características importantes, es decir, seleccionadas para los diferentes modelos moleculares eran adecuadas para entrenar modelos con otras instancias de peso. En términos generales podemos decir que la información relevante para una representación mínima no lo es para modelos moleculares más grandes. Pero las características de instancias mayores, tienen rendimientos estables, es decir no desmejoran para la instancia menor (URE). Por lo que esto podría revelar que las instancias mayores de peso aportan mayor información.

Todos los modelos fueron desarrollados con descriptores moleculares de una única instancia de peso, es decir, los descriptores utilizados son univaluados en los modelos moleculares URE, Mn o Mw. Los resultados parecen confirmar que el uso de la URE como modelo molecular de polímeros es una sobresimplificación. La

experimentación desarrollada fue una exploración para determinar si existía otro modelo molecular univaluado, no explorado en la literatura, capaz de captar de mejor manera la complejidad estructural de los polímeros polidispersos. Existe una tendencia hacia una mejora de desempeño de los modelos QSPR generados usando representaciones computacionales basadas en instancias de peso mayores.

Cabe mencionar que esta experimentación se realizó de forma acotada en relación a las limitaciones tecnológicas existentes, como por ejemplo la baja cantidad de descriptores moleculares, 57 en total. Esto último podría haber generado subconjuntos de DMs poco variables durante los procesos de selección de características llevados a cabo. Sin embargo, esto no sucedió, ya que, si bien existen características compartidas por cada subconjunto, ninguno es igual. Por lo que se puede concluir que cada instancia de peso está aportando información diferente (DMs diferentes), lo que soporta la idea de comenzar a modelar los polímeros con más de una instancia de peso, tema que se presenta en los próximos capítulos.

Síntesis y Conclusiones del Capítulo 5

Hasta el momento se ha trabajado con representaciones simplificadas de polímeros para predecir sus propiedades, tanto en la literatura como en esta tesis. Por lo general, se utiliza la Unidad Repetitiva Estructural (URE) o el monómero como modelo molecular. El objetivo de este capítulo fue responder varias preguntas, que en su conjunto fueron nombradas como Segunda y Tercera Pregunta de Investigación. Por un lado, se buscó revelar si la URE es una representación sobresimplificada de un polímero polidisperso. Es decir, investigar si las características seleccionadas para entrenar los modelos QSPR (o los modelos finales) que predicen propiedades mecánicas son relevantes al aplicarlas en otras representaciones de pesos mayores (pesos promedios: Mn o Mw), ya que una variable relevante en un dominio (instancia de peso) puede convertirse en irrelevante o redundante en otro, y así, agregar más ruido que información útil al describir el concepto de interés.

Todos los modelos fueron desarrollados con descriptores de una única instancia de peso, es decir, los descriptores moleculares utilizados son univaluados en URE, Mn o Mw. Los resultados parecen confirmar que el uso de la URE como modelo molecular de polímeros es una sobresimplificación. Los resultados obtenidos a partir de los otros modelos moleculares (Mn y Mw) parecen ser más robustos que los de enfoque sintético (URE). En cuanto a las características seleccionadas, por cada uno de los métodos utilizados para las diferentes instancias de peso, existen algunas características compartidas sobre todo entre las de Mn y Mw, pero ningún subconjunto de descriptores moleculares es igual, a pesar de la baja disponibilidad de los mismos (57 totales). Por lo tanto, puede concluirse que cada instancia de peso está aportando información diferente, lo que refuerza la hipótesis acerca de que modelar los polímeros con más de una instancia de peso resultaría más adecuado.

MODELADO QSPR CON DMS TRIVALUADOS

CAPÍTULO 6

La mayoría de los modelos QSPR disponibles en la literatura usan representaciones computacionales simplificadas (univaluadas) de polímeros basados en su unidad repetitiva estructural. El objetivo de este capítulo es evaluar el efecto de esta simplificación y analizar nuevas estrategias para lograr caracterizaciones alternativas que capturen el fenómeno de la polidispersión, usando un modelo QSPR con Descriptores Moleculares trivaluados.

6.1. IMPORTANCIA DE LA POLIDISPERSIÓN

El primer paso en el Aprendizaje Maquinal aplicado al modelado QSPR es codificar la estructura de la molécula en una forma que sea apta para dicho aprendizaje. Una representación útil debe codificar características que sean relevantes. Entonces, ¿una representación univaluada es adecuada para caracterizar a los polímeros polidispersos? [Wu *et al.*, 2016]. Los materiales poliméricos son muy diferentes de otros materiales, como las cerámicas o los metales, debido a su naturaleza macromolecular [Meijer & Govaert, 2005]. El peso molecular y la distribución del peso molecular de un polímero pueden afectar notablemente sus propiedades mecánicas [Martin *et al.*, 1972]. Por esta razón, la polidispersión que presentan desalienta el uso de una única molécula como modelo molecular.

Tanto las bolsas de basura como las fibras de alto rendimiento usadas para la fabricación de chalecos antibalas, están hechas de polietileno, lo que cambia entre un material y otro es la curva de distribución de pesos. Por lo tanto, una misma estructura química (URE), que es base de un material, puede dar perfiles de aplicación muy diferentes si la polidispersión es muy distinta, así como también si varía su procesamiento [Meijer & Govaert, 2005]. La influencia de la polidispersión en las propiedades macromoleculares de los polímeros, como lo son las mecánicas, se reconoce desde hace décadas. Sin embargo, ha sido relativamente difícil de comprobar empíricamente por varias razones. Uno de los problemas más importantes fue la caracterización de la curva completa de distribución de pesos moleculares mediante métodos de fraccionamiento, porque la recolección de fracciones y la determinación de los diversos promedios de pesos moleculares de los polímeros llevan mucho tiempo [Martin *et al.*, 1972].

La polidispersión es un atributo distintivo de los materiales poliméricos y, como consecuencia de este, cada descriptor molecular de un material polimérico también está asociado a una distribución discreta de valores. Esta distribución se obtiene calculando el descriptor molecular para los diferentes largos de cadenas poliméricas presentes en la distribución de pesos moleculares de un polímero y asociándolo a la frecuencia de ese largo de cadena. Sin embargo, como se presentó en los capítulos anteriores, los enfoques tradicionales de modelado QSPR propuestos para predecir propiedades de polímeros en Informática de Polímeros no tienen en cuenta la polidispersión, simplificando así en demasía la representación computacional de cada material polimérico, en general, usando su Unidad Repetitiva Estructural (URE) como modelo molecular único [Cravero *et al.*, 2018a].

6.2. SEGUNDA PROPUESTA ALTERNATIVA

En este capítulo, el objetivo es analizar la Segunda Propuesta Alternativa como estrategia para abordar caracterizaciones de los materiales que capturen, al menos parcialmente, el fenómeno de la polidispersión. Nuestra contribución es evaluar el efecto de una caracterización univaluada del polímero y abordar el problema de selección de descriptores moleculares (DMs) en un contexto de polidispersión, utilizando DMs valuados en tres modelos moleculares o largos diferentes de cadena correspondientes a la URE y a los pesos moleculares promedios en número (Mn) y en peso (Mw) [Cravero *et al.*, 2019b].

Los descriptores moleculares son variables que caracterizan la estructura de los compuestos químicos, y el primer paso para inferir un modelo QSPR es identificar cuáles descriptores están más relacionados con la propiedad en estudio. Las herramientas de software para el cálculo de descriptores moleculares pueden calcular miles de variables, pero, en general, un modelo QSPR de regresión solo requiere un pequeño número para estimar la propiedad objetivo. Por lo tanto, la selección de descriptores en el modelado QSPR es una instancia del problema tradicional de Selección de Características. En particular, en este contexto de modelado QSPR con DMs trivaluados, intentamos responder lo que denominamos Cuarta Pregunta de Investigación: *Los modelos QSPR inferidos mediante el uso de representaciones trivaluada, ¿producen estimaciones más precisas que los modelos QSPR generados a partir de representaciones univaluadas?*

A su vez, esta pregunta puede dividirse en las preguntas que se describen más abajo, lo cual permite abordarlas de manera modular y plantear los experimentos necesarios para responderlas:

a). ¿Existen modelos moleculares basados en sus pesos moleculares promedios, de los materiales, que den como resultado modelos QSPR predictivos más precisos que los obtenidos por modelos moleculares URE?

b). ¿Es aconsejable integrar en una única base de datos los descriptores moleculares correspondientes a modelos moleculares de los diferentes pesos característicos relacionados con las curvas de distribución de peso molecular de los materiales?

6.2.1. REPRESENTACIÓN COMPUTACIONAL TRIVALUADA

La hipótesis central a explorar, relacionada con las preguntas de investigación previamente establecidas, es: *los modelos QSPR inferidos mediante el uso de información estructural correspondiente a varias longitudes de cadenas poliméricas de diferentes pesos característicos (modelos moleculares URE, Mn y Mw) de estos materiales deberían producir estimaciones más precisas que modelos QSPR generados a partir solo del modelo molecular URE.* Como se expuso en capítulos anteriores, obtener bases de datos de materiales es una tarea compleja; más aún, si la base de datos debe estar integrada por polímeros relacionados con el estudio de las propiedades mecánicas asociadas a ensayos de tracción. Por esta razón, se trabajó a partir de una base de datos interna propia (ver Capítulo 4.3. Base de Datos utilizada).

La base de datos inicial (BD) consta de 77 polímeros representados por su URE, cuyos rangos de longitud de cadena para obtener Mn son [5–7205 UREs] y para Mw, [24–10773 UREs]. Sin embargo, solo pudo trabajarse con 57 de ellos, calculándoles los descriptores moleculares para cada material para las tres longitudes de cadena polimérica correspondientes a las tres instancias utilizadas: URE (modelo molecular URE), Mn (modelo molecular Mn) y Mw (modelo molecular Mw), tal cual se había realizado para estudiar la Primera Propuesta Alternativa (ver capítulo 5.3). Así, se obtuvieron tres bases de datos (DB_{URE}, DB_{Mn} y DB_{Mw}) integradas por 57 polímeros, pero esta vez con 108 Descriptores Moleculares cada una, de los cuales 57 son DMs clásicos calculados con la herramienta RCDK [Guha, 2007] (ídem capítulo 5) y los 51 restantes corresponden a los propuestos por Palomba y denominados en esta tesis como DMs de visión macro [Palomba, 2014], los cuales aportan una visión extra a los clásicos sumando descripción. Una vez calculados

todos los DMs para las tres bases de datos, se unieron en una única base datos llamada DB_{Global}. Un esquema gráfico es presentado en la Figura 6.1 para facilitar la comprensión de la construcción de estas bases de datos. Esta cuarta base de datos contiene la información asociada a tres instancias (modelos moleculares) diferentes de pesos moleculares y constituye un primer enfoque para caracterizar materiales poliméricos mediante la captura de parte de su polidispersión [Cravero *et al.*, 2019b].

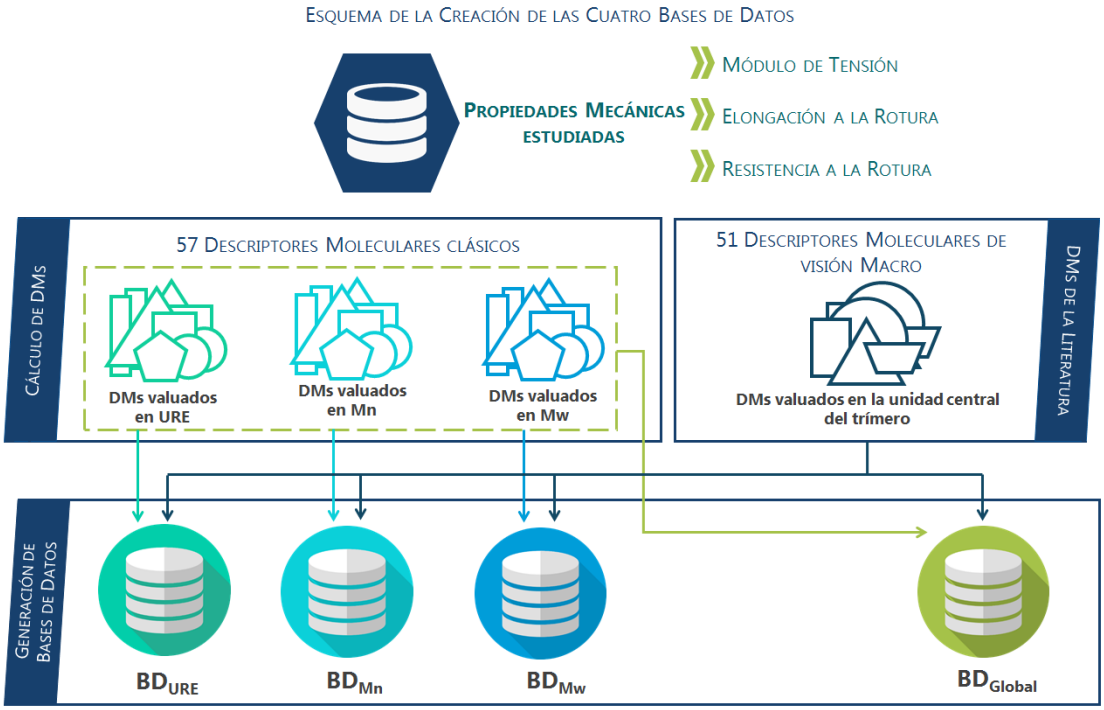


FIGURA 6. 1. ESQUEMA GRÁFICO DE LA CREACIÓN DE LAS CUATRO BASES DE DATOS UTILIZADAS EN ESTE CAPÍTULO. TRES UNIVALUADAS: BD_{URE}, BD_{MN} Y BD_{MW} Y UNA LLAMADA BD_{GLOBAL} QUE INCLUYE LOS DESCRIPTORES TRIVALUADOS.

En este punto es válido aclarar que la base de datos inicial contiene información para las tres propiedades mecánicas con las que se viene trabajando desde el capítulo 4 de esta tesis: Módulo Elástico, Elongación a la Rotura y Resistencia a la rotura y, por tal motivo, todas las bases de datos resultantes también lo hacen, por lo que se tuvieron en cuenta estas tres propiedades en los experimentos realizados. Esto permite, además, una mayor diversidad en los datos y la posibilidad de evaluar la hipótesis de trabajo en varios escenarios, aportando así a favorecer la generalización de las conclusiones de los experimentos.

6.2.2. MODELADO QSPR CON DMS TRIVALUADOS

Una base de datos de 57 polímeros podría considerarse pequeña, sin embargo, nueve familias químicas diferentes están representadas y por esta razón,

el uso de técnicas de muestreo aleatorio en la división de los datos no es recomendable para garantizar la preservación de las características globales de la misma (diversidad estructural). Por esta razón, cada una de las bases de datos fue dividida en cuatro conjuntos de manera que cada uno de ellos conserve la representatividad de la diversidad estructural de la base de datos original (ídem a capítulo 5). Para cada experimento, las bases de datos se dividieron en dos partes, una parte consistió en tres de los cuatro conjuntos iniciales es llamada conjunto de datos de entrenamiento, equivale al ~80% de la base de datos (46 moléculas), y se usó para la fase de entrenamiento: selección de descriptores y entrenamiento de modelos QSPR utilizando el enfoque Validación Cruzada Dejando Uno Fuera (LOOCV). La otra parte, el cuarto pliegue, fue utilizada como conjunto de datos de validación externa en la fase de prueba para la cual se reservó este pliegue equivalente al ~20% (11 moléculas) de la base de datos (Figura 6.2.).

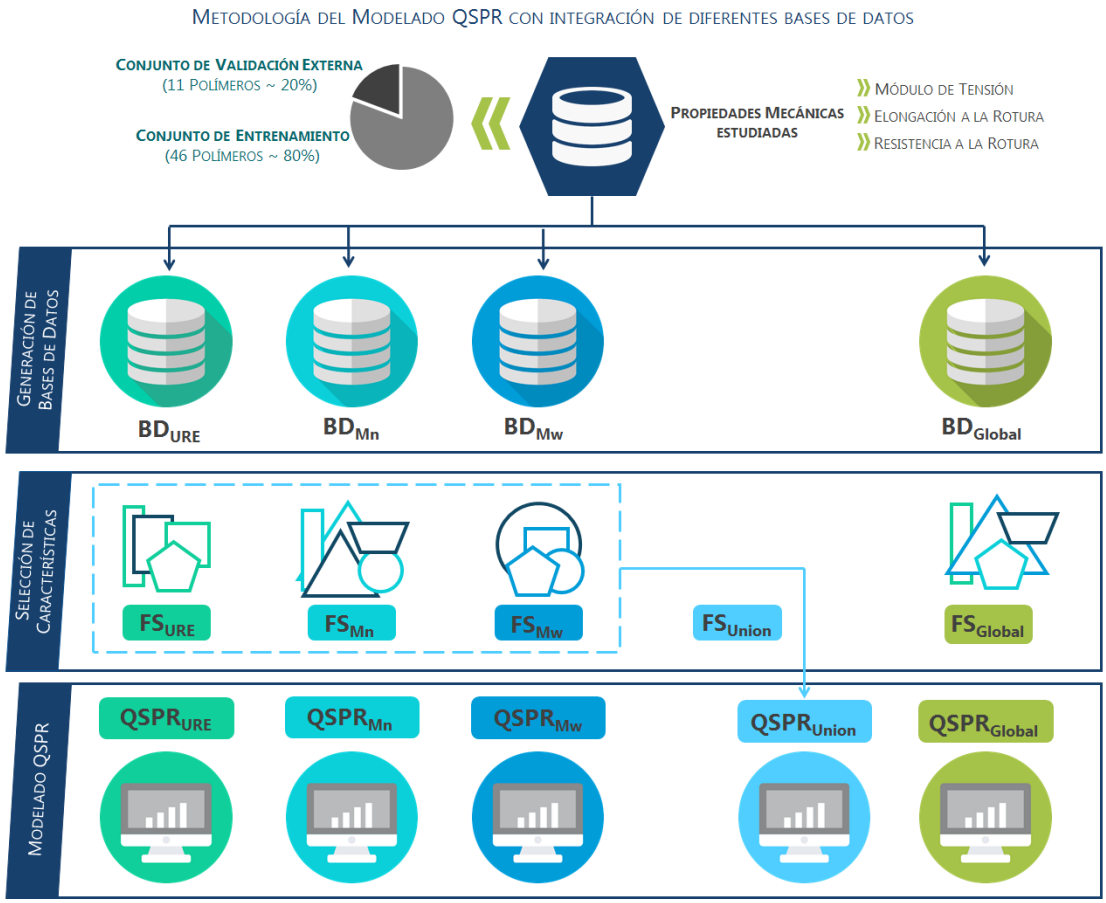


FIGURA 6. 2. ESQUEMA GRÁFICO DE LA METODOLOGÍA UTILIZADA PARA LA CONSTRUCCIÓN DE LOS MODELOS QSPR UTILIZADOS PARA RESPONDER LAS PREGUNTAS DE INVESTIGACIÓN FORMULADAS.

6.2.2.a. SELECCIÓN DE CARACTERÍSTICAS

Para responder las preguntas de investigación planteadas anteriormente (ver 6.2 Segunda Propuesta Alternativa) se diseñaron varios experimentos. En cuanto a

la selección de los descriptores moleculares más relevantes de cada base de datos (en los conjuntos de datos de entrenamiento), se utilizó la herramienta WEKA [Hall *et al.,* 2009], el método *Wrapper* con *Best First* como algoritmo de búsqueda y se emplearon cuatro técnicas de selección: Regresión lineal (W-LR), Redes Neuronales (W-NN), Bosques Aleatorios (W-RF) y Comités Aleatorios (W-RC), obteniendo así, cuatro subconjuntos de DMs para cada base de datos. Luego, considerando la cardinalidad y el equilibrio entre las diferentes clases de descriptores (clásicos y de visión macro) se eligieron los dos mejores subconjuntos para cada una.

A continuación, se presentan estos resultados para las cuatro bases de datos (BD_{URE}, BD_{Mn}, BD_{Mw} y BD_{Global}) para Módulo de Tensión en la Tabla 6.1, para Elongación a la Rotura en la Tabla 6.2 y para Resistencia a la Rotura en la Tabla 6.3. Los DMs clásicos de los distintos modelos moleculares (URE, Mn y Mw) presentes en la DB_{Global} fueron denominados con un sufijo que identifica la instancia de peso de la que proviene. Por ejemplo, nAcid_Mn es el DM para el número de grupos ácidos (nAcid) calculado para el modelo molecular Mn, mientras que nAcid_Mw es el mismo DM pero calculado para el modelo molecular Mw. Los descriptores seleccionados para más de una instancia de peso se resaltan en negrita en las tablas siguientes.

TABLA 6. 1. DESCRIPTORES MOLECULARES (DMS) PARA LOS DOS MEJORES SUBCONJUNTOS SELECCIONADOS EN CADA BASE DE DATOS PARA MÓDULO DE TENSIÓN. LOS DMS COMPARTIDOS POR DOS O MÁS SUBCONJUNTOS ESTÁN RESALTADOS EN NEGRITA.

BD	Método de Selección de DMs	DMs Clásicos	DMs de vision Macro	Cardinalidad
BD_{URE}	W-RF	**khs.ssNH**, **khs.sssN**, C3SP3, **C4SP3**	nBondsM.(Mn/MW), **nSA_{MC}/nSA_{SC}**, M_{SC}	7
	W-RC	**khs.ssNH**, **khs.sssN**, **khs.sssCH**, nRing4	Mn/MW, nP_{MC}, **nSA_{MC}/nSA_{CS}**	7
BD_{Mn}	W-RF	C2SP3, **C4SP3**, **khs.ssNH**	**nSA_{MC}/nSA_{SC}**	4
	W-LR	**khs.aaaC**, **khs.sssP**	Mn/SA_{MC}, **nP_{MC}/nP_{SC}**, V_{MC}	5
BD_{Mw}	W-NN	C2SP2, khs.aaO, khs.ddssS, **khs.dsssP**, **khs.ssNH**, nAromRings, nRings6	$nLogP_{SC}$, **nP_{MC}/nP_{SC}**, nP_{SC}, IPD	11
	W-RF	khs.aaCH, khs.aaO, **khs.ssNH**, **khs.sssCH**, **khs.sssN**, nAcid	**M_{SC}**, **nSA_{MC}/nSA_{SC}**	8
BD_{Global}	W-NN	Kier3_URE, **nAromBond_Mn**, **nAromBond_Mw**, nsmallRings_Mn, tpsaEfficiency_Mw, tpsaEfficiency_URE	**nSA_{MC}/nSA_{SC}**	7
	W-RF	**C4SP3_URE**, khs.ssCH2_Mn, khs.ssNH_URE, **khs.sssN_Mn**	**nSA_{MC}/nSA_{SC}**	5

TABLA 6. 2. DESCRIPTORES MOLECULARES (DMS) PARA LOS DOS MEJORES SUBCONJUNTOS SELECCIONADOS EN CADA BASE DE DATOS PARA ELONGACIÓN A LA ROTURA. LOS DMS COMPARTIDOS POR DOS O MÁS SUBCONJUNTOS ESTÁN RESALTADOS EN NEGRITA.

BD	Método de Selección de DMs	DMs Clásicos	DMs de vision Macro	Cardinalidad
BD_{URE}	W-NN	**ALogP**, **C2SP2**, nRing4	**nR_{MC}/nR_{SC}**, nSA_{SC}	5
	W-LR	**nRing5**, Zagreb, kier3, khs.dsssP, khs.aasC	**CHS**, $LogP_{SC}$, nSA_{MC}, **nSA_{SC}**, nR_{SC}, nV_{SC}, IPD	12
BD_{Mn}	W-NN	khs.dssC, **nRing5**	$nLogP_{MC}$, **nSA_{SC}**, **nR_{MC}/nR_{SC}**	5
	W-RF	**khs.aaO**, **khs.sssCH**, **nRing7**	**CHS**	4
BD_{Mw}	W-LR	nAtomP, **Khs.aaO**, khs.dsCH, VAdjMat	AMR, $LogP_{MC}$/$LogP_{SC}$, nP_{MC}/nP_{SC}, **nR_{MC}/nR_{SC}**	8
	W-RF	**khs.aaO**, **khs.sssCH**, **nRing7**	**CHS**	4
BD_{Global}	W-NN	**ALogP_URE**, **C2SP2_URE**, **khs.aaO_Mw**	**nSA_{SC}**, **nR_{MC}/nR_{SC}**	5
	W-RF	**khs.aaO_Mw**, khs.ddssS_Mn, **khs.sssCH_URE**, khs.ssS_URE, nAcid_Mw, **nRings4_URE**, **nRings7_Mw**	**CHS**	8

TABLA 6. 3. DESCRIPTORES MOLECULARES (DMS) PARA LOS DOS MEJORES SUBCONJUNTOS SELECCIONADOS EN CADA BASE DE DATOS PARA RESISTENCIA A LA ROTURA. LOS DMS COMPARTIDOS POR DOS O MÁS SUBCONJUNTOS ESTÁN RESALTADOS EN NEGRITA.

BD	Método de Selección de DMs	DMs Clásicos	DMs de vision Macro	Cardinalidad
BD_{URE}	W-NN	**khs.dsssP**, **MW**, **nSmallRings**	**nV_{MC}**, **nSA_{MC}/nSA_{SC}**	5
	W-RC	Khs.ddssS, nAromBond, **nAromRings**	$LogP_{MC}$, **nP_{MC}**, nR_{MC}	6
BD_{Mn}	W-LR	C1SP3, **khs.aaO**, **khs.ssS**	**nM_{MC}**, $LogP_{MC}$/$LogP_{SC}$, IPD	6
	W-NN	C1SP2**,** **khs.dO**, **khs.ssS**, **nAcid**, **nRings5**, tpsaEfficecincy	nSA_{MS}, **nV_{MC}**, IPD	9
BD_{Mw}	W-LR	**khs.aaO**, **khs.ssS**, khs.ssssC, **nAromRings**	**nM_{MC}**, **nSA_{MC}/nSA_{SC}**, P_{MC}	7
	W-NN	**khs.dO**, **khs.sssP**, **khs.ssS**, **nRings5**	**nSA_{MC}/nSA_{SC}**, **nP_{MC}**	6
BD_{Global}	W-NN	**nSmallRings_URE**, nRings4_URE, nRings6_URE, **MW_URE**, VAdjMat_URE, khs.aaaC_URE, **khs.dsssP_URE**, khs.ssNH_Mn, **khs.aaO_Mw**	**-**	9
	W-RF	C3SP3_URE, khs.sCH3_Mn, khs.ssNH_Mw, **khs.ssS_Mn**, **khs.ssS_Mw**, khs.sssCH_URE, khs.sssN_Mn**,** **nAcid_Mw**	**nP_{MC}**	9

La Tabla 6.1, Tabla 6.2 y Tabla 6.3 muestran 24 subconjuntos de DMs con rangos de cardinalidad de 4–11, 4–12 y 5-9 para Módulo de Tensión, Elongación a la Rotura y Resistencia a la Rotura, respectivamente. Algunos DMs parecen ser especialmente informativos, ya que son elegidos en la mayoría de los subconjuntos, en particular CHS (velocidad del ensayo de tensión). Este es un parámetro de ensayo importante ya que afecta fuertemente las tres propiedades. Los polímeros presentan comportamientos muy diferentes dependiendo de la velocidad de ensayo, es decir, de CHS (*Cross Head Speed*). Otro descriptor frecuente es el DM de visión macro nSA_{MC}/nSA_{SC}, que representa al tamaño de la molécula como una relación de las cadenas principales y laterales. Finalmente, el descriptor IPD (Índice de Polidispersión) es otro de los frecuentemente seleccionados, este DM contiene información macro relacionada con la distribución de pesos moleculares. De todos modos, muchos otros MDs aparecen específicamente dependiendo de la propiedad en estudio.

6.2.2.b. ENTRENAMIENTO DE MODELOS QSPR

Los ocho mejores subconjuntos de descriptores moleculares por propiedad (Tablas 6.1, 6.2 y 6.3), fueron entrenados para inferir modelos QSPR con cuatro métodos de Aprendizaje Maquinal (AM): Regresión lineal (LR), Redes Neuronales (NN), Bosques Aleatorios (RF) y Comités Aleatorios (RC). Por lo tanto, se obtuvieron un total de 32 modelos por propiedad, donde 8 corresponden a cada base de datos: DB_{URE}, BD_{Mn}, BD_{Mw} y BD_{Global}. Después de analizarlos teniendo en cuenta tanto el rendimiento estadístico en términos de R^2 como error absoluto medio (MAE) y error cuadrático medio (RMSE), un solo subconjunto de DMs fue elegido para cada base de datos, obteniendo así cuatro subconjuntos finales que fueron llamados: FS_{URE}, FS_{Mn}, FS_{Mw} y FS_{Global} (Figura 6.2).

A continuación, se presentan todos los resultados obtenidos durante el proceso de modelado QSPR en la fase de validación externa. Se resaltan en negrita aquellos subconjuntos de DMs (FS) seleccionados en cada base de datos para cada propiedad. En la Tabla 6.4 se encuentran los valores obtenidos para Módulo de Tensión, en la Tabla 6.5, para Elongación a la Rotura y finalmente en la Tabla 6.5 están los relacionados a Resistencia a la Rotura. El subconjunto llamado FS_{Union} obtenido mediante la unión de los descriptores de los tres subconjuntos finales pertenecientes a FS_{URE}, FS_{Mn}, FS_{Mw} (Figura 6.2), tiene como objetivo capturar las características de los tres modelos moleculares involucrados, de forma tal que compitan con el FS_{Global} que también contienen información sobre los tres pesos característicos. Luego, el subconjunto FS_{Union} fue entrenado con los cuatro métodos

de Aprendizaje Maquinal. Los resultados obtenidos son mostrados en las correspondientes tablas.

TABLA 6. 4. RESULTADOS PARA LA VALIDACIÓN EXTERNA DE LOS MODELOS QSPR INFERIDOS PARA EL MÓDULO DE TENSIÓN. SE RESALTAN LOS MEJORES MODELOS OBTENIDOS POR CADA BASE DE DATOS.

BD	Método de Selección de DMs	Método de Aprendizaje Maquinal	R^2	MAE	RMSE	QSPR Seleccionado
BD_{URE}	W-RF	NN	0.9566	0.2126	0.2659	
		LR	0.5275	0.3371	0.5975	
		RF	0.9635	0.1773	0.2185	
		RC	0.9606	0.1636	0.2078	
	W-RC	NN	0.9617	0.1733	0.2248	
		LR	0.9451	0.1700	0.2230	
		RF	**0.9679**	0.1720	0.2026	**$QSPR_{URE}$**
		RC	0.9597	0.1845	0.2087	
BD_{Mn}	**W-LR**	NN	**0.9629**	0.1409	0.1806	**$QSPR_{Mn}$**
		LR	0.9321	0.1895	0.2639	
		RF	0.9691	0.1785	0.2129	
		RC	0.9677	0.1645	0.1867	
	W-RF	NN	0.8705	0.2633	0.3748	
		LR	0.3842	0.3496	0.6650	
		RF	0.9386	0.1857	0.2469	
		RC	0.8853	0.1892	0.3158	
BD_{Mw}	W-NN	NN	0.9063	0.2714	0.3723	
		LR	0.9138	0.2316	0.3050	
		RF	0.9070	0.2178	0.2886	
		RC	0.8524	0.2491	0.3561	
	W-RF	NN	0.8917	0.2776	0.3756	
		LR	0.2502	0.6007	1.0446	
		RF	**0.9700**	0.1576	0.2102	**$QSPR_{Mw}$**

		RC	0.9554	0.1627	0.2171	
BD_{Global}	**W-NN**	NN	0.9561	0.1986	0.2195	
		LR	0.8352	0.3689	0.4354	
		RF	0.9648	0.1737	0.2193	
		RC	**0.9725**	0.1583	0.1801	**$QSPR_{Global}$**
	W-RF	NN	0.9633	0.2358	0.3877	
		LR	0.5397	0.3812	0.5763	
		RF	0.9461	0.1933	0.2373	
		RC	0.943	0.1818	0.2297	
	Union	NN	0.9149	0.203	0.3015	
		LR	0.8562	0.3173	0.4315	
		RF	**0.9813**	0.1542	0.1875	**$QSPR_{Union}$**
		RC	0.9672	0.1655	0.1964	

TABLA 6. 5. RESULTADOS PARA LA VALIDACIÓN EXTERNA DE LOS MODELOS QSPR INFERIDOS PARA EL ELONGACIÓN A LA ROTURA. SE RESALTAN LOS MEJORES MODELOS OBTENIDOS POR CADA BASE DE DATOS.

BD	Método de Selección de DMs	Método de Aprendizaje Maquinal	R^2	MAE	RMSE	QSPR Seleccionado
BD_{URE}	W-RF	NN	0.7340	1.2352	1.5411	
		LR	0.5001	1.8209	2.3655	
		RF	0.7411	1.1513	1.4346	
		RC	0.7833	1.0911	1.3368	
	W-RC	NN	**0.8330**	1.0497	1.2064	**$QSPR_{URE}$**
		LR	0.5154	3.1633	3.7118	
		RF	0.6280	1.5599	2.1261	
		RC	0.5968	1.3498	2.0884	
BD_{Mn}	**W-LR**	NN	**0.6899**	1.3693	1.7031	**$QSPR_{Mn}$**
		LR	0.2727	2.4716	2.8574	

		RF	0.6862	1.1106	1.4948	
		RC	0.6558	1.1521	1.6777	
	W-RF	NN	0.5271	2.3210	3.1044	
		LR	0.5473	4.3929	5.7501	
		RF	0.4892	1.9268	2.7512	
		RC	0.6357	1.4413	2.2263	
BD_{Mw}	**W-NN**	NN	0.6541	1.4819	2.0096	
		LR	0.2157	2.2988	3.0954	
		RF	**0.7509**	1.0524	1.3414	**$QSPR_{Mw}$**
		RC	0.4624	1.5828	2.5243	
	W-RF	NN	0.3996	3.2927	4.3099	
		LR	0.4617	4.1389	5.3196	
		RF	0.5923	1.7089	2.1705	
		RC	0.6226	1.7352	2.4145	
BD_{Global}	**W-NN**	NN	**0.8008**	1.4625	1.6278	**$QSPR_{Global}$**
		LR	0.5154	3.1633	3.7118	
		RF	0.6151	1.6294	2.2596	
		RC	0.5849	1.6202	2.3897	
	W-RF	NN	0.3804	2.4766	3.4411	
		LR	0.2567	2.9293	3.7689	
		RF	0.4200	1.5948	2.1920	
		RC	0.4345	1.4770	2.4448	
	Union	NN	0.3421	2.2817	3.7070	
		LR	0.4259	2.0996	2.8563	
		RF	**0.7529**	1.0485	1.3408	**$QSPR_{Union}$**
		RC	0.5209	1.6769	2.8484	

TABLA 6. 6. RESULTADOS PARA LA VALIDACIÓN EXTERNA DE LOS MODELOS QSPR INFERIDOS PARA EL RESISTENCIA A LA ROTURA. SE RESALTAN LOS MEJORES MODELOS OBTENIDOS POR CADA BASE DE DATOS.

BD	Método de Selección de DMs	Método de Aprendizaje Maquinal	R^2	MAE	RMSE	QSPR Seleccionado
BD_{URE}	**W-NN**	NN	0.8073	11.8861	14.7842	
		LR	0.6774	13.3514	15.9744	
		RF	**0.8377**	8.3722	10.6909	**$QSPR_{URE}$**
		RC	0.7924	9.0928	11.5932	
	W-RC	NN	0.7507	11.2030	13.3103	
		LR	0.8056	8.9981	11.3524	
		RF	0.8214	9.5269	11.0806	
		RC	0.7948	10.3411	12.4258	
BD_{Mn}	**W-LR**	NN	0.8938	8.8613	10.6211	
		LR	0.8951	10.0132	11.1302	
		RF	0.9049	7.0414	8.0105	
		RC	**0.9267**	6.0900	7.2808	**$QSPR_{Mn}$**
	W-NN	NN	0.8859	10.4284	13.9697	
		LR	0.8433	9.5370	10.9208	
		RF	0.9067	6.8196	7.8821	
		RC	0.9021	7.0133	8.1724	
BD_{Mw}	**W-LR**	NN	0.8875	8.9193	10.1277	
		LR	0.9019	8.0678	9.5589	
		RF	0.9287	6.4345	6.9805	
		RC	**0.9386**	6.0861	6.6046	**$QSPR_{Mw}$**
	W-NN	NN	0.8402	10.1787	13.3127	
		LR	0.8453	8.5572	10.1181	
		RF	0.9392	6.1239	6.9786	
		RC	0.9158	7.5842	8.5045	
BD_{Global}	**W-NN**	NN	**0.9370**	8.1382	9.4827	**$QSPR_{Global}$**
		LR	0.8936	7.8275	9.7808	
		RF	0.8703	8.6735	9.9669	
		RC	0.8861	7.5038	9.7898	
	W-RF	NN	0.7941	12.3004	16.2356	

	LR	0.8674	9.5058	10.9746	
	RF	0.9064	6.8495	8.1909	
	RC	0.9142	6.6555	8.3283	
Union	NN	0.8909	9.6300	12.8077	
	LR	0.8800	8.7362	10.7207	
	RF	0.9086	7.1507	8.0140	
	RC	**0.9170**	6.0731	7.7099	**$QSPR_{Union}$**

En resumen, los cinco subconjuntos finales (FS_{URE}, FS_{Mn}, FS_{Mw}, FS_{Global} y FS_{Union}) fueron entrenados con los cuatro métodos de aprendizaje maquinal LR, NN, RF y RC, obteniendo 20 modelos. Estos modelos se evaluaron siguiendo los mismos criterios estadísticos (mayor R^2 y menor error). Finalmente se seleccionó un solo modelo QSPR de cada subconjunto final: $QSPR_{URE}$; $QSPR_{Mn}$; $QSPR_{Mw}$; $QSPR_{Global}$ y $QSPR_{Union}$, para responder las preguntas de investigación planteadas en la sección 6.2, Segunda Propuesta Alternativa. Este proceso de modelado se repitió para cada propiedad y se presenta en la Tabla 6.7 un resumen de los resultados hasta aquí obtenidos.

TABLA 6. 7. PARA LOS MODELOS QSPR CORRESPONDIENTES A CADA PROPIEDAD MECÁNICA SE PRESENTAN: LA CARDINALIDAD, R^2, LOS ERRORES MAE Y RMSE, ASÍ COMO TAMBIÉN EL MÉTODO DE SELECCIÓN DE CARACTERÍSTICAS (M. FS) Y EL DE APRENDIZAJE MAQUINAL (M. ML) EMPLEADOS.

Propiedad Mecánica	Modelo QSPR	Cardinalidad	M. FS	M. ML	R^2	MAE	RMSE
Módulo de Tensión	$QSPR_{URE}$	7	W-RC	RF	0.9679	0.1720	0.2026
	$QSPR_{Mn}$	5	W-LR	NN	0.9629	0.1409	0.1806
	$QSPR_{Mw}$	8	W-RF	RF	0.9700	0.1576	0.2102
	$QSPR_{Global}$	7	W-NN	RC	**0.9725**	0.1583	0.1801
	$QSPR_{Union}$	19	Union	RF	**0.9813**	0.1542	0.1875
	$QSPR_{AWI}$	-	-	-	0.9609	0.1568	0.1982
Elongación a la Rotura	**$QSPR_{URE}$**	5	W-NN	NN	**0.8330**	1.0497	1.2064
	$QSPR_{Mn}$	5	W-NN	NN	0.6899	1.3693	1.7031
	$QSPR_{Mw}$	8	W-LR	RF	0.7509	1.0524	1.3414
	$QSPR_{Global}$	5	W-NN	NN	**0.8008**	1.4625	1.6278
	$QSPR_{Union}$	15	Union	RF	0.7529	1.0485	1.3408
	$QSPR_{AWI}$	-	-	-	0.7304	1.1571	1.4324
Resistencia a la Rotura	$QSPR_{URE}$	5	W-NN	RF	0.8377	8.3722	10.6909
	$QSPR_{Mn}$	6	W-LR	RC	0.9267	6.0900	7.2808
	$QSPR_{Mw}$	7	W-LR	RC	**0.9386**	6.0861	6.6046
	$QSPR_{Global}$	9	W-NN	NN	**0.9370**	8.1582	9.4827
	$QSPR_{Union}$	16	Union	RC	0.9170	6.0731	7.7099
	$QSPR_{AWI}$	-	-	-	0.9002	6.8495	8.3851

Respondiendo la pregunta a

La pregunta de investigación **a**, planteada en este capítulo de tesis, busca responder a la incógnita de la existencia de una representación univaluada de un material polímero que sea mejor a la utilizada hasta el momento en la literatura (modelo molecular URE). Dado que las bases de datos utilizadas en este capítulo se diferencian de las del capítulo 5 (estas últimas no incluían DMs de visión macro), es necesario analizar nuevamente estos resultados para comenzar a responder a la pregunta: ¿Existen modelos moleculares basados en sus pesos moleculares promedios, de los materiales, que den como resultado modelos QSPR predictivos más precisos que los obtenidos por modelos moleculares URE?

A partir de estos nuevos resultados, es posible concluir nuevamente que el modelo molecular URE no garantiza siempre un alto rendimiento, es decir, para todos los casos aquí estudiados (tres propiedades mecánicas). En particular, el modelo $QSPR_{URE}$ inferido para predecir la Resistencia a la Rotura es superado claramente por los modelos QSPR obtenidos utilizando representaciones más altas (modelos moleculares Mn y Mw) (Tabla 6.7). Por lo tanto, es posible afirmar que, en algunos casos, el modelo molecular URE puede no ser suficiente para inferir modelos QSPR precisos. De manera similar, puede decirse que para Módulo de Tensión y para Resistencia a la Rotura los modelos QSPR inferidos a partir del modelo molecular Mw, alcanzan mejores rendimientos que los obtenidos por URE y Mn. Es posible concluir entonces, que para la Elongación a la Rotura los modelos moleculares basados en Mn y Mw están claramente dominados por el modelo $QSPR_{URE}$. En consecuencia, ninguna de las representaciones alternativas univaluadas propuestas en este capítulo, basadas en el uso de los modelos moleculares Mn y Mw, alcanzaron mejores rendimientos que los logrados por el modelo molecular URE en todos los casos de estudios. Aunque, tampoco han tenido siempre los peores rendimientos.

Respondiendo la pregunta b

Los resultados analizados hasta aquí, permiten avanzar hacia la pregunta de investigación **b**, donde se evalúa la conveniencia, o no, de utilizar modelos trivaluados para emular la polidispersión de un material polimérico, al menos parcialmente. En este sentido, el análisis de los resultados obtenidos por los modelos $QSPR_{Global}$ y $QSPR_{Union}$ es útil para responder la pregunta **b**: ¿Es aconsejable integrar en una única base de datos los descriptores moleculares correspondientes a modelos moleculares de los diferentes pesos característicos relacionados con las curvas de distribución de peso molecular de los materiales?

Como se explicó anteriormente, en la base de datos global (DB_{Global}) se integró toda la información estructural relacionada con los tres modelos moleculares de un material polimérico utilizados (URE, Mn y Mw). Además, se han llevado a cabo dos estrategias de selección de DMs diferentes, la primera estrategia, llamada FS_{Global}, ejecuta un nuevo procedimiento de Selección de Características para obtener un subconjunto de descriptores moleculares del conjunto completo de DMs incluidos en la DB_{Global}. En la segunda estrategia, llamada FS_{Union}, no existe un procedimiento de Selección de Características, sino una unión de los descriptores (evitando las repeticiones) incluidos en los subconjuntos: FS_{URE}, FS_{Mn}, FS_{Mw}. Estos pasos metodológicos se han sido incluidos en la Figura 6.2.

Los modelos $QSPR_{Global}$ logran un rendimiento competitivo en términos de R^2 (Tabla 6.7) para las tres propiedades mecánicas, respaldando la hipótesis de que los modelos QSPR inferidos haciendo uso de la información de varias cadenas poliméricas de diferentes pesos pueden proporcionar estimaciones más precisas de los valores de las propiedades mecánicas. Para la Elongación a la Rotura, el modelo $QSPR_{URE}$ logró un rendimiento mejor que el modelo $QSPR_{Global}$, incluso cuando la base de datos DB_{Global} incluye todos los MDs de DB_{URE}. Este resultado puede explicarse considerando que el problema de selección de DMs es un caso particular del problema de Selección de Características (*Feature Selección*) y es un problema perteneciente a la clase NP-complejo (*NP-hardness*) en términos de complejidad computacional [Amaldi & Kann, 1998]. Por esta razón, cualquier procedimiento de optimización combinatorial para seleccionar subconjuntos de MDs solo puede garantizar selecciones subóptimas. Esta observación también es válida para los resultados presentados para la Resistencia a la Rotura cuando se contrastan los rendimientos de $QSPR_{Global}$ y $QSPR_{Mw}$.

Con respecto al modelo $QSPR_{Union}$, su rendimiento es ligeramente más alto para el Módulo de Tensión, y más bajo tanto para Elongación a la Rotura como para Resistencia a la Rotura. Esto podría resultar algo inesperado teniendo en cuenta que los modelos $QSPR_{Union}$ utilizan más descriptores moleculares, es decir, más información que los modelos $QSPR_{Global}$ (ver cardinalidad en Tabla 6.7.). Sin embargo, la unión de los subconjuntos de DMs seleccionados para las bases de datos de DB_{URE}, DB_{Mn} y DB_{Mw} podría estar combinando información redundante en los modelos $QSPR_{Unión}$ y esto podría derivar en sobreajuste y falta de generalizabilidad del modelo QSPR. Reduciendo así, la capacidad del modelo QSPR para hacer predicciones precisas en datos desconocidos (validación externa).

Finalmente, es posible responder la pregunta **b** concluyendo que la integración de la información estructural que corresponde a modelos moleculares

de diferentes pesos característicos relacionados con las curvas de distribución de pesos moleculares de los materiales poliméricos, es una práctica aconsejable para el modelado QSPR en Informática de Polímeros. Esta estrategia trivaluada de representación puede capturar parcialmente la polidispersión inherente a estos materiales, beneficiando la precisión y la generalizabilidad de los modelos QSPR inferidos sin requerir un número significativamente mayor de descriptores moleculares seleccionados (cardinalidad del modelo QSPR) que para aquellos modelos QSPR obtenidos a partir del modelo molecular URE.

Hasta el momento, la generalizabilidad de los modelos $QSPR_{Global}$ y $QSPR_{Union}$ se comparó con la cuantificación unificada de la generalizabilidad de los modelos QSPR inferidos a partir de instancias únicas de representación (DB_{URE}, DB_{Mn} y DB_{Mw}). Sin embargo, es necesario proveer un escenario justo de comparación, por este motivo se propone obtener de una manera agregada las métricas estadísticas (R^2 y errores) de los diferentes modelos moleculares (URE, Mn y Mw). Para esto fue necesario computar nuevas métricas que fueron asignadas a un nuevo modelo QSPR llamado $QSPR_{AWI}$, AWI por las siglas en inglés para Todas las Instancias de Peso (*All Weight Instances*). Por lo tanto, los resultados de de las validaciones externas (R^2 y los errores) obtenidas por los tres modelos QSPR univaluados ($QSPR_{URE}$, $QSPR_{Mn}$ y $QSPR_{Mw}$) se gestionan y se interpretan como el de un único modelo QSPR unificado ($QSPR_{AWI}$). En la Figura 6.3 se presentan estos resultados de forma comparativa con $QSPR_{Global}$ y $QSPR_{Union}$. Además, en la Tabla 6.7 se muestran los resultados obtenidos por $QSPR_{AWI}$ para las tres propiedades en estudio.

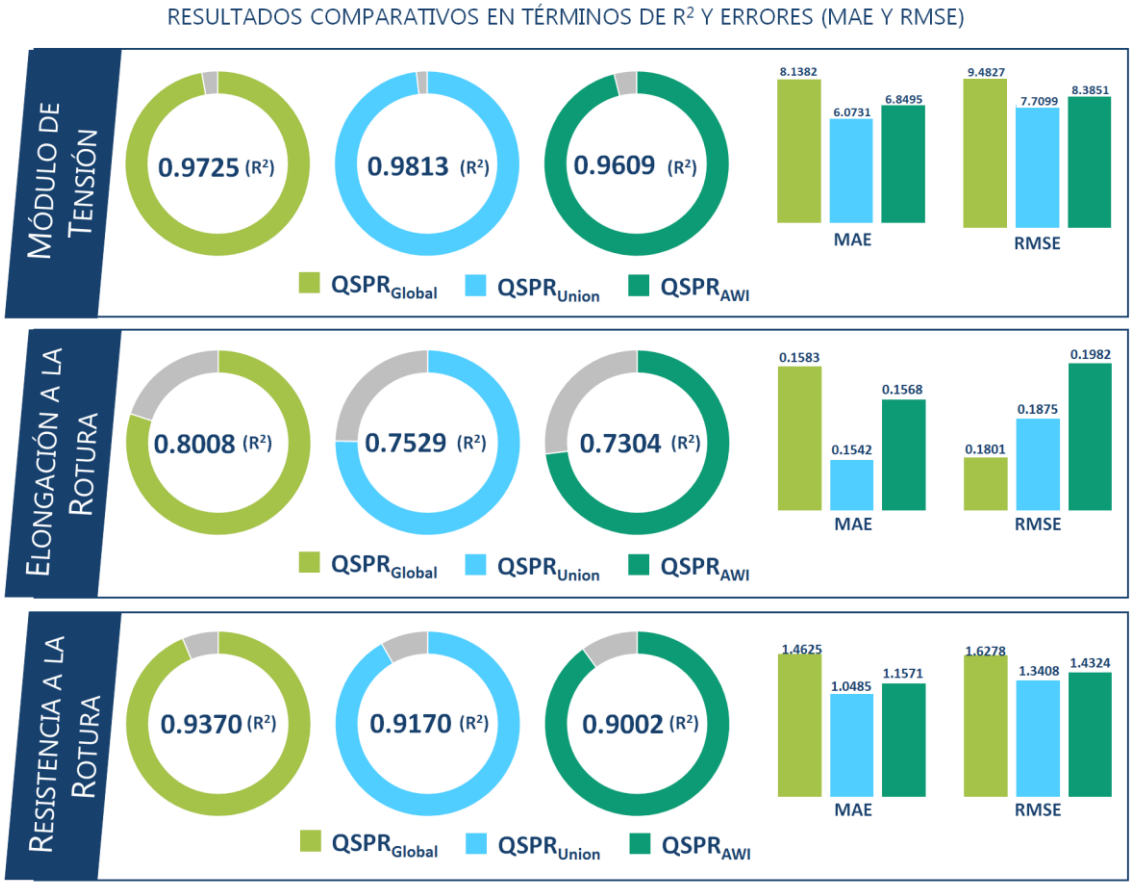


FIGURA 6. 3 RESULTADOS COMPARATIVOS EN TÉRMINOS DE R^2 Y ERRORES (MAE Y RMSE) PARA $QSPR_{GLOBAL}$, $QSPR_{UNIÓN}$, Y $QSPR_{AWI}$ PARA EL CONJUNTO DE VALIDACIÓN EXTERNA DE LAS TRES PROPIEDADES ESTUDIADAS.

El valor R^2 alcanzado por el modelo $QSPR_{AWI}$ para el Módulo de Tensión es 0.9609 y se obtuvo calculando el valor R^2 que corresponde al conjunto completo de resultados de predicción obtenidos por $QSPR_{URE}$, $QSPR_{Mn}$ y $QSPR_{Mw}$ para la misma propiedad cuando estos modelos se prueban en sus propios conjuntos de datos de validación externa. Se aplicó el mismo procedimiento para calcular las demás métricas de los modelos $QSPR_{AWI}$. Contrastando la precisión estadística, en términos de R^2, de $QSPR_{Unión}$ y $QSPR_{Global}$ con la de $QSPR_{AWI}$ para las tres propiedades, está claro que $QSPR_{Global}$ y $QSPR_{Union}$ superan el rendimiento de los modelos QSPR aprendidos de las bases de datos de polímeros correspondientes a una sola instancia de peso (univaluadas). Por lo tanto, puede concluirse que los modelos QSPR inferidos a partir de varios modelos moleculares, es decir, varias instancias de peso (trivaluadas en estos estudios) tienen mejores propiedades de generalización que aquellos modelos QSPR obtenidos basándose en un único modelo molecular o única instancia de peso.

6.3. CONCLUSIONES SOBRE LA SEGUNDA PROPUESTA ALTERNATIVA

Los objetivos de este capítulo fueron evaluar el efecto de esta visión simplificada de la complejidad estructural del polímero y proponer nuevas ideas para lograr otras caracterizaciones de polímeros que capturen el fenómeno de polidispersión, al menos parcialmente. En particular, nos enfocamos en explorar una hipótesis clave: *los modelos QSPR inferidos mediante el uso de información estructural correspondiente a varias longitudes de cadena polimérica de diferentes pesos característicos de estos materiales (modelos moleculares URE, Mn y Mw) deberían producir estimaciones más precisas que aquellos modelos QSPR generados únicamente a partir del modelo molecular URE.*

Se evaluaron diferentes representaciones computacionales en combinación con varias técnicas de aprendizaje maquinal. Estas se utilizaron tanto para la Selección de Características con el objetivo de seleccionar los DMs más relevantes relacionados con cada propiedad objetivo, como para inferir los modelos QSPR, entrenándolos con cuatro diferentes métodos de aprendizaje maquinal. Con respecto a atravesar, o no, un proceso de Selección de Características para incluir las características trivaluadas, podemos concluir que cuando efectivamente se lo atraviesa no solo disminuye notablemente la cardinalidad de los subconjuntos de DMs, sino que también los modelos $QSPR_{Global}$ logran mejor o igual desempeño estadístico y habilidades de generalizabilidad que aquellos que se limitan a unir las características que habían sido capturadas para modelos moleculares univaluados ($QSPR_{Unión}$).

Los resultados obtenidos, tanto en el capítulo anterior como en este, muestran claramente que los modelos moleculares de polímeros basados en URE o alternativas univaluadas son simplificaciones excesivas y, en general, una práctica desaconsejable para el modelado QSPR, al menos dentro del alcance los casos de estudio abordados en esta tesis. Con respecto a la segunda propuesta alternativa de representación computacional de materiales poliméricos e hipótesis central de este capítulo, contribuimos con una base de datos de representaciones basadas en el cálculo de los descriptores moleculares para tres modelos moleculares por cada material (URE, Mn y Mw), que logra un alto rendimiento y captura la polidispersión parcialmente. En particular, podemos concluir que los modelos QSPR inferidos a partir de bases de datos que incluyen diferentes instancias de peso de los polímeros alcanzan mejores propiedades de generalizabilidad.

Síntesis y Conclusiones del Capítulo 6

Uno de los temas más complejos para el modelado computacional de materiales poliméricos está relacionado con la polidispersión que caracteriza a estas estructuras macromoleculares. Los modelos QSPR propuestos en la literatura para predecir las propiedades de polímeros evitan este problema simplificando su representación computacional utilizando como modelo molecular únicamente su Unidad Repetitiva Estructural (URE). Recientemente, en el Capítulo 5, planteamos lo que en esta tesis llamamos Primera Propuesta Alternativa de representación computacional de polímeros para el modelado QSPR, basándonos en modelos moleculares de cadenas poliméricas que alcanzan los pesos promedios en número (Mn) y peso (Mw). Hasta aquí, la selección de características se realizaba para modelos univaluados, es decir, modelos moleculares URE, Mn o Mw. La hipótesis central a explorar en este capítulo, y que encarna la segunda propuesta alternativa de representación computacional de polímeros es: *los modelos QSPR inferidos mediante el uso de información estructural correspondiente a varias longitudes de cadena polimérica de diferentes pesos característicos de estos materiales (modelos moleculares URE, Mn y Mw) deberían producir estimaciones más precisas que aquellos modelos QSPR generados a partir únicamente del modelo molecular URE.*

Con respecto a la hipótesis planteada, esta efectivamente se verifica y para poder demostrarla se generó una base de datos que contiene información trivaluada de los valores de los Descriptores Moleculares (DMs). A través del análisis de los resultados obtenidos puede concluirse que, atravesar un proceso de selección de características multivaluadas ($QSPR_{Global}$) permite la generación de modelos QSPR con mejor precisión estadística y habilidades de generalizabilidad que aquellos que se infieren a partir de la unión de DMs que fueron seleccionados para una única instancia de peso a la vez ($QSPR_{Unión}$). Además, se realizó una evaluación con respecto a cómo se comportan estos modelos frente a los modelos univaluados tomados de forma agregada para poder realizar una comparación justa ($QSPR_{AWI}$). Estos modelos univaluados, para ninguna de las propiedades logran los mejores rendimientos estadísticos. A partir de esta base de datos trivaluada (BD_{Global}) se atravesó un proceso de Selección de Características en busca de los DMs más relevantes para cada propiedad. De este paso se obtuvo que varios de los descriptores seleccionados estaban valuados en más de una instancia, lo cual habilita a pensar que continuar los esfuerzos por conseguir modelos QSPR basados en características multivaluadas no sería, en principio, en vano.

SELECCIÓN DE CARACTERÍSTICAS MULTIVALUADAS

CAPÍTULO 7

Los pesos moleculares promedios aportan información útil, pero estos no caracterizan completamente a los materiales poliméricos. Por eso, conocer la curva de distribución de pesos moleculares y ser capaces de valuar los descriptores moleculares (DMs) en ella, constituyen pasos deseables para la caracterización de polímeros en una forma más realista. Estos DMs podrían aportar más y mejor información respecto de los univaluados. Sin embargo, este abordaje es el primero en su tipo y este capítulo presenta una propuesta acerca de cómo tratar la Selección de Características de estos DMs multivaluados expresados como una distribución probabilística discreta.

7.1. IMPORTANCIA DE LA FRECUENCIA

El diseño y el descubrimiento de nuevos materiales están siendo impulsados cada vez más por métodos de Aprendizaje Maquinal (ML), que extraen patrones de conjuntos de datos preexistentes [Jennings *et al.*, 2019]. En Informática de Polímeros, el modelado QSPR es particularmente desafiante, ya que el objetivo que persigue es avanzar hacia la comprensión y el diseño de nuevos materiales poliméricos personalizados [Adams, 2010; Audus & de Pablo, 2017]. El modelado QSPR de polímeros exige un cauteloso tratamiento de representación computacional de las macromoléculas, ya que cada polímero está formado por varias cadenas poliméricas que a su vez consisten en la sucesión de muchas Unidades Repetitivas Estructurales (URE). Es decir, ninguno de los materiales poliméricos con que se trabaja en esta tesis está formado por una única URE, sino que está formado por un conjunto de cadenas (uniones de UREs) de varios largos (cantidad de URE).

En la Figura 7.1 puede observarse un esquema de una curva de dispersión de pesos moleculares. En el eje de las abscisas se representan los valores de los pesos moleculares en orden creciente y en el eje de las ordenadas están las frecuencias con que esos pesos (o largos de cadena) aparecen en el material polimérico. La mayoría de los estudios publicados sobre modelado QSPR de polímeros utilizan modelos moleculares sintéticos; es decir, caracterizan a los polímeros a través de DMs calculados en una sola URE [Cao & Lin, 2003; Duce *et al.*, 2006; Yu *et al.*, 2008; Liu & Cao, 2009; Toropova *et al.*, 2014; Jabeen *et al.*, 2017; Chen *et al.*, 2018] o en la

unidad central del trímero [Katritzky *et al.*, 1998; Palomba *et al.,* 2012b; Cravero *et al.,* 2019a]. Sin embargo, la polidispersión no ha sido considerada en la mayoría de las contribuciones al modelado QSPR en Informática de Polímeros [Wu *et al.*, 2016].

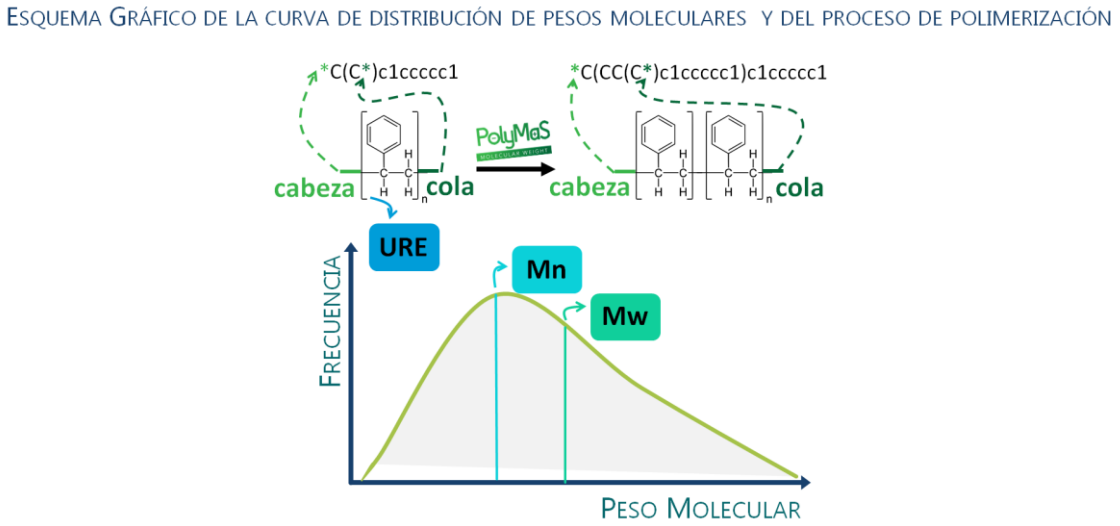


FIGURA 7. 1. REPRESENTACIÓN ESQUEMÁTICA DE LA CURVA DE POLIDISPERSIÓN DE PESOS DE UN MATERIAL POLIMÉRICO, INDICANDO LOS PESOS MOLECULARES PROMEDIOS Y EL PROCESO DE POLIMERIZACIÓN A PARTIR DE LA URE A TRAVÉS DE POLYMAS.

El cálculo de DMs directo sobre las cadenas de polímeros de longitudes realistas aún no es factible [Wu *et al.*, 2016], o al menos, es tecnológicamente complejo hacerlo. Por un lado, debido a que las cadenas poliméricas suelen ser mucho más grandes que las moléculas pequeñas clásicas (por ejemplo drogas o incluso una URE). Por el otro, están los problemas asociados a la recolección de las distintas fracciones y determinación de sus pesos [Martin *et al.*, 1972], sumado a la representación computacional individual para luego, recién, ejecutar el costoso cálculo de descriptores.

La caracterización parcial de la polidispersión ha sido propuesta anteriormente [Cravero *et al.*, 2019b], aunque sin tener en cuenta toda la curva de polidispersión, sino considerando los modelos moleculares de la URE y de los pesos moleculares promedios en número (Mn) y en peso (Mw). Es decir, los DMs de cada polímero se caracterizaron mediante tres valores. Los resultados informados para esta propuesta, confirman que los modelos QSPR predictivos inferidos de bases de datos que incluyen varias instancias de representaciones de polímeros (modelos moleculares) obtienen mejores rendimientos, en términos de generalizabilidad, que los modelos predictivos inferidos a partir de bases de datos de representaciones univaluadas. Esto constituye un antecedente promisorio que respalda los esfuerzos

experimentales para continuar estudiando el impacto de la polidispersión en el modelado QSPR de las propiedades mecánicas en Informática de Polímeros.

7.1.1. RECONSTRUCCIÓN TEÓRICA DE LA CURVA DE POLIDISPERSIÓN

Muchas de las propiedades de los polímeros dependen entre otros factores de la naturaleza química de sus macromoléculas, de los métodos de ensayos, de las condiciones de medición de dichas propiedades y de la curva de distribución de pesos moleculares que presentan [Adams, 2010]. Esta curva define el perfil de comportamiento mecánico del material, entre otras propiedades fisicoquímicas [Meijer & Govaert, 2005]. Por lo tanto, al diseñar un nuevo material polimérico no sólo es importante describir correctamente la estructura química, sino que es indispensable considerar la polidispersión deseada. La importancia de la reconstrucción teórica de la curva de distribución de pesos moleculares reside, justamente, en que más allá de su relevancia no suelen estar disponibles para su acceso, junto con los otros datos que si se reportan en la literatura [Adams, 2010].

Para la reconstrucción de la curva de dispersión de pesos, se propuso emplear una técnica de aprendizaje maquinal supervisado en tres etapas (Figura 7.2). El objetivo fue crear dos funciones capaces de predecir los dos parámetros de la distribución lognormal: la media (μ) y el desvío estándar (σ) [Cravero et al., 2016e]. Se decidió seguir una distribución lognormal porque es la distribución más comúnmente observada dentro de las distribuciones asimétricas, con las cuales los polímeros de nuestra base de datos se identifican, y porque es la más utilizada en la literatura referente a Química de Polímeros [Lorenzini *et al.*, 1992; Brandolin *et al.*, 1996]. Además, la distribución lognormal es la que se usa típicamente para materiales poliméricos y la más modelada. Se puede describir una amplia gama de distribuciones con dos parámetros: peso molecular promedio en número (Mn) y peso molecular promedio en peso (Mw) [Monteiro, 2015].

Para entrenar los algoritmos que deben predecir μ y σ, se generó una nueva base de datos con curvas de distribución de pesos moleculares reales (llamado Dataset 2 en Figura 7.2). Las 22 resinas termoplásticas puras[1] incluidas en el Dataset 2, presentan una distribución de tipo *lognormal* y tienen disponibles tanto los datos de sus respectivas curvas de distribución de pesos reales, como sus pesos promedios: Mn, Mw. En la Etapa 1, a partir de cada curva se estimaron, usando test

[1] Nota de la autora: Agradecemos la colaboración del grupo de Ciencia y Tecnología de Polímeros de PLAPIQUI, en las personas de la Lic. Cristina Frova y el Dr. Jorge Guapacha, por facilitarnos estos datos.

de bondad de ajuste, los dos parámetros de la distribución *lognormal*: media y desvío estándar.

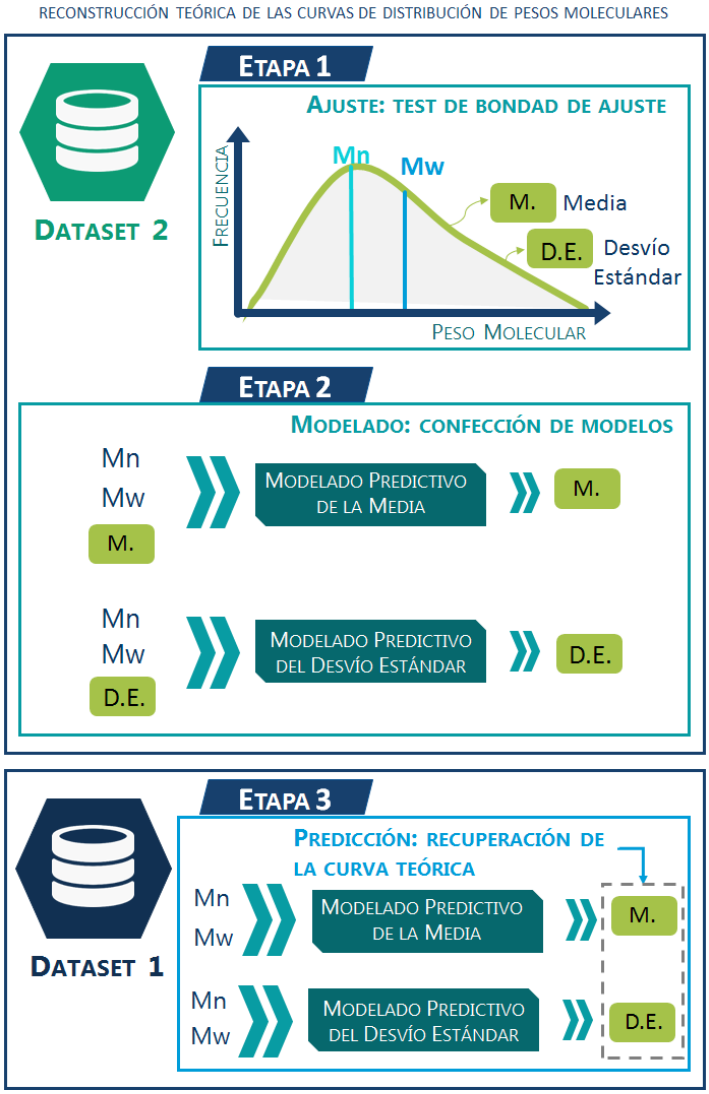


FIGURA 7. 2 METODOLOGÍA DE LA RECONSTRUCCIÓN TEÓRICA DE LAS CURVAS DE DISTRIBUCIÓN DE PESOS MOLECULARES DEL DATASET 1.

Luego, en la Etapa 2, se infirieron dos modelos de aprendizaje maquinal capaces de relacionar Mn y Mw (*input*) con cada parámetro de la *lognormal* usando la herramienta WEKA [Hall *et al.,* 2009]. Los métodos utilizados para inferir los modelos de regresión fueron: Regresión Lineal, Perceptron Multicapa, Perceptron Lineal Simple, Comités Aleatorios, Árboles Aleatorios, Bosques Aleatorios, *M5Rules*, *Decision Stump*, *Arbol M5P* (para más información sobre estos métodos ver Capítulo 1. Conceptos de Aprendizaje Maquinal). La calidad de predicción de cada uno de ellos fue evaluada mediante un enfoque de validación cruzada dejando uno fuera (LOOCV) y validación cruzada de 7 iteraciones (*7-fold cross-validation*). La Etapa 3 consistió en obtener las predicciones de μ y σ para nuestra base de datos

(llamada Dataset 1 en la Figura 7.2), a partir de los modelos inferidos en la Etapa 2, introduciendo los datos de Mn y Mw del Dataset 1. Con estos parámetros, fue posible reconstruir la curva de distribución de pesos a partir de los pesos moleculares promedios.

Concluyendo, el principal aporte de esta sección fue la construcción de un set de curvas de polidispersión de materiales reales, que según nuestro conocimiento, es la primera metodología que intenta abordar esta problemática. Sin embargo, al entrenar los modelos con una base de datos muy pequeña los resultados obtenidos contienen gran porcentaje de errores, por lo que se espera poder aumentar la cantidad de polímeros con datos de sus curvas de polidispersión para volver a entrenar los modelos, lograr disminuir el error y poder utilizar las curvas obtenidas.

7.2. TERCERA PROPUESTA ALTERNATIVA

La naturaleza macro y polidispersa es un aspecto distintivo de los polímeros y, como consecuencia, para su correcto modelado QSPR cada descriptor debería estar asociado a una distribución discreta de valores que podría obtenerse al calcular el descriptor molecular para cada una de las diferentes cadenas poliméricas asociándolo a las frecuencias de las mismas. En otras palabras, se propone avanzar con una Tercera Propuesta Alternativa a la representación computacional de materiales poliméricos, mediante la utilización de distribuciones probabilísticas derivadas de la curva de distribución de pesos moleculares para caracterizar los DMs asociados a todas las cadenas poliméricas presentes [Cravero *et al.*, 2018b]. Esto naturalmente deriva en el siguiente interrogante: *¿La caracterización de los descriptores moleculares mediante distribuciones probabilísticas derivadas de la curva de polidispersión permitiría generar modelos QSPR con desempeño superior a los modelos que no tienen en cuenta la frecuencia de los distintos largos de cadenas poliméricas en el material?* Responder esta pregunta, a diferencia de los capítulos anteriores, plantea tres nuevos desafíos en términos de Ciencias de la Computación: primero, diseñar una representación multivaluada que permita resolver el problema de trasladar el fenómeno de polidispersión a las distribuciones probabilísticas asociadas a los descriptores moleculares; segundo, dado que los algoritmos de Selección de Características empleados hasta ahora en el modelado QSPR solo operan sobre variables numéricas, y no sobre distribuciones de probabilidad, es necesario diseñar un algoritmo de Selección de Características que se adecue a esta problemática; y tercero, diseñar una estrategia de modelado QSPR adecuada para operar con DMs con polidispersión. En el contexto de este capítulo se abordan los primeros dos desafíos, quedando el tercero como parte de los

trabajos a futuro. De esta manera se puede reformular el interrogante anterior en lo que denominaremos la Quinta, y última, Pregunta de Investigación de esta tesis: *¿Es posible identificar con más precisión los DMs más relevantes usando un algoritmo de Selección de Características multivaluadas que usando enfoques tradicionales sobre representaciones univaluadas*?

7.2.1. Representación Computacional Multivaluada

La metodología propuesta para la representación computacional multivaluada de descriptores moleculares parte de la reconstrucción teórica de la curva de distribución de pesos moleculares. Consiste en una primera etapa, donde se eligen n representantes de tamaños del material (Figura 7.3), que básicamente sería seleccionar diferentes largos de cadenas poliméricas, dentro de las opciones que ofrece la curva [Cravero *et al.*, 2016f]. Debido a que los polímeros pueden presentar curvas con anchos diferentes, la selección de representantes puede realizarse, por un lado, dividiendo el ancho de la curva ($peso\ molecular\ mayor$ - $peso\ molecular\ menor$) en n fracciones y tomar el valor del medio de cada una de ellas y, por el otro, dividendo por área bajo la curva. En este último caso, las n fracciones proyectadas en el eje de las abscisas no tendrán el mismo ancho pero de igual manera se puede tomar el valor medio de cada una de ellas para obtener así los n representantes de peso (Figura 7.3).

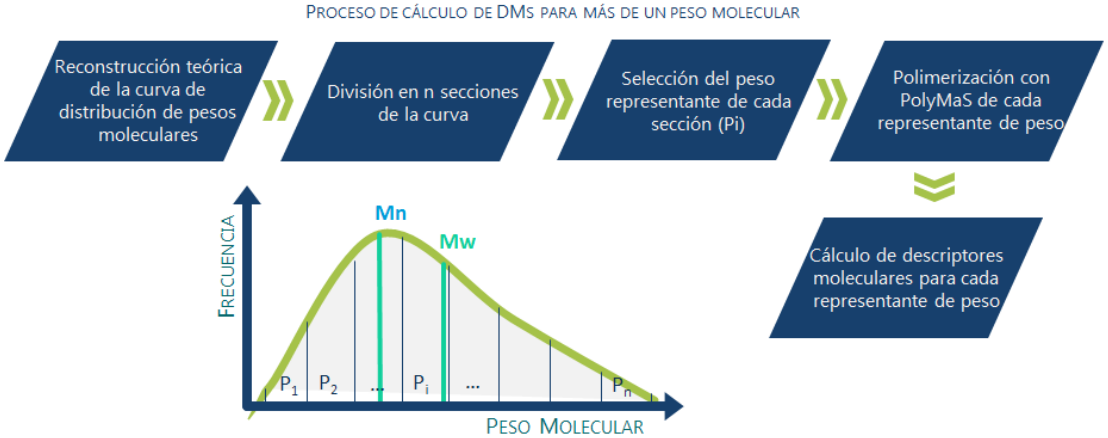


Figura 7. 3. Proceso del cálculo multivaluado de un Descriptor Molecular

Una vez seleccionados los n representantes de peso, se procede a la polimerización *in silico,* con PolyMaS [Schustik *et al.*, 2019b], de las cadenas poliméricas hasta que alcancen esos pesos. A estas cadenas se les calculan los descriptores moleculares y luego, los mismos son ponderados por la frecuencia alcanzada (eje de las ordenadas en la curva de distribución de pesos moleculares) para ese largo de cadena (representantes de peso). Esta etapa aún no ha sido

totalmente abordada en forma experimental con datos reales, porque depende de la aceleración del cálculo de descriptores para moléculas de alto peso, problema aún en etapa de resolución. Como primera aproximación, actualmente, estamos dividiendo a la curva en 10 percentiles, es decir, trabajando con $n = 10$. Luego de obtener estas divisiones se selecciona el peso promedio de cada intervalo (P_i en la Figura 7.3) para polimerizarlo con PolyMaS y posteriormente calcular los DMs, aún en resolución para los pesos más grandes. Dividir la curva en 10 representantes es solo la primera aproximación, se pretende seguir variando n en busca de un óptimo.

7.2.1.a. Modelado del Fenómeno de Polidispersión en Descriptores Moleculares mediante Datos sintéticos

Una problemática reconocida en Informática de Polímeros es la falta de bases de datos de referencia [Ma *et al.*, 2019]. Además, como fue planteado en la sección anterior, la generación de una base de datos con valores reales de DMs para polímeros con alto peso molecular es computacionalmente costosa ya que el principal factor de variabilidad (es decir, la polidispersión) que caracteriza un material polimérico debe ser modelado, y aún quedan desafíos por resolver enfocados en el cálculo de DMs de pesos de muy alto valor. Por esta razón, en este contexto el uso de datos sintéticos se convierte en un enfoque aconsejable para realizar pruebas de concepto. Además, la generación aleatoria de datos sintéticos es más adecuada para evaluar escalabilidad y robustez [Brinkhoff, 2009], que es justamente lo que se pretende hacer, es decir, evaluar la capacidad de un nuevo método de Selección de Características que se propone más adelante en este capítulo de tesis, bautizado $FS4RV_{DD}$ (*Feature Selection for Random Variables with Discrete Distribution*), el cual es capaz de trabajar con DMs multivaluados de forma conjunta [Cravero *et al.*, 2018b; Cravero, *et al.*, 2020].

Para la generación computacional de estas bases de datos sintéticas, debe considerarse que los materiales poliméricos tienen una variabilidad de peso. Es decir, no tienen un único peso molecular, sino una distribución de pesos moleculares. En la primera etapa, seguimos una distribución lognormal para modelarla porque, como se explicó anteriormente, es la distribución más ampliamente usada en la literatura. La idea clave consiste en obtener una distribución discreta de valores para cada descriptor molecular. En este contexto, fue necesario generar polímeros sintéticos caracterizados por esta distribución lognormal para obtener la base de datos sintética. Esta distribución tiene dos parámetros μ y σ que corresponden a la media y al desvío estándar,

respectivamente. De esta manera, para cada material polimérico, dados dos valores generados aleatoriamente correspondientes a los parámetros de distribución de peso molecular, se genera la curva de distribución de peso molecular para cada polímero (Figura 7.4, parte A). A fin de testear la escalabilidad, se crearon tres bases de datos de tamaños diferentes, con 400, 800 y 1600 materiales, respectivamente [Cravero, *et al.*, 2020].

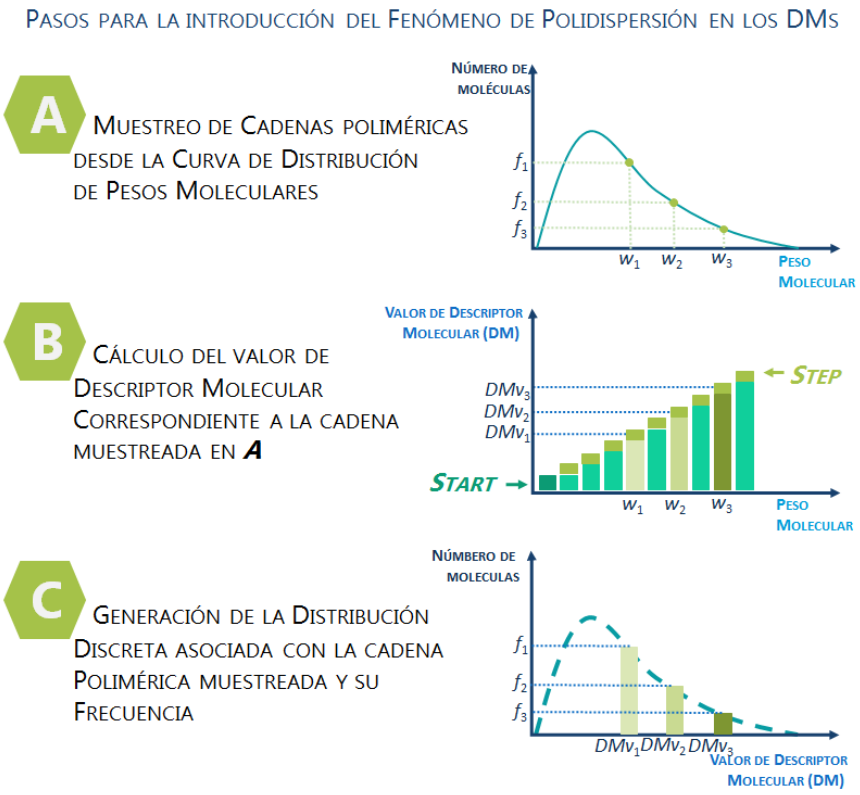


FIGURA 7. 4. DESCRIPCIÓN DE LOS PASOS PARA INTRODUCIR EL FENÓMENO DE POLIDISPERSIÓN EN LA CARACTERIZACIÓN DEL DESCRIPTOR MOLECULAR.

En la segunda etapa se calculó para cada descriptor un vector de valores, el cual se genera mediante dos números aleatorios: $start$ y $step$. Se inició el primer valor de cada vector en $start$ y se obtuvieron los siguientes valores sumando consecutivamente el valor de $step$ hasta completar el largo del vector. Esto representa los valores de los DMs calculados incrementalmente para diferentes pesos de cadena molecular para un material polimérico (Figura 7.4, parte B). En consecuencia, como tercera etapa, se guardó un vector de coordenadas para cada material en la base de datos de materiales (Figura 7.4, parte C). La primera coordenada pertenece a los diferentes pesos moleculares (eje x) y la segunda, representa las frecuencias (eje y). Es decir, la frecuencia de cada valor de descriptor en la distribución se recuperó de la curva de polidispersión (lognormal).

7.2.1.b. Construcción de la Base de Datos Sintética

La base de datos se define como una matriz en la que las filas están asociadas con materiales poliméricos y las columnas, con descriptores (Figura 7.4, parte A). Para cada celda C_{ij}, la curva de polidispersión correspondiente al i-ésimo material M_i se asoció con el vector de valores asociados con el j-ésimo descriptor D_j (Figura 7.4, parte B). Entonces, se podría obtener una distribución discreta de valores para cada descriptor D_j, y para cada material M_i. En particular, estas distribuciones se obtuvieron tomando 100 muestras k de las curvas de dispersión de pesos moleculares generadas en la primera etapa (Figura 7.4, parte C). Cada muestra k corresponde a una instancia de peso molecular y coincide con un valor de descriptor.

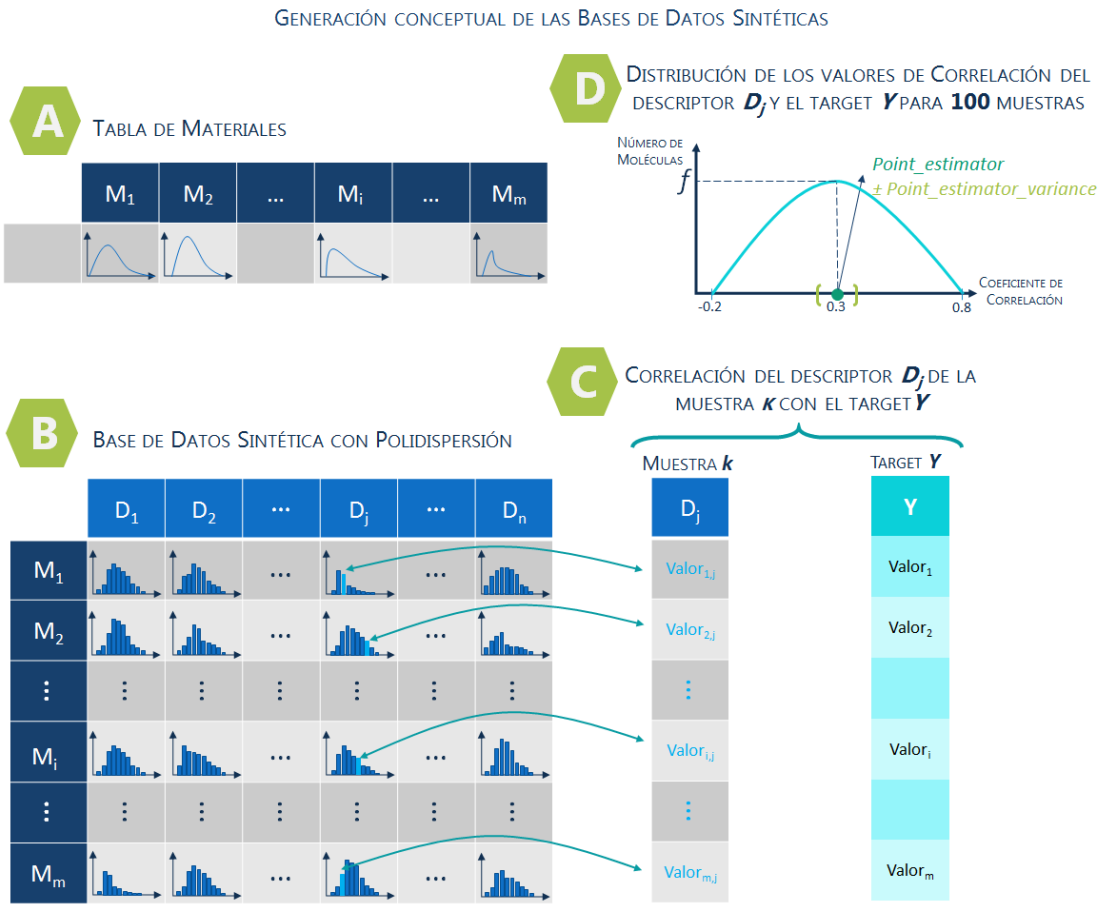


FIGURA 7. 5. Esquema gráfico de la construcción conceptual de las bases de datos. Tenga en cuenta que contiene datos polidispersos y no un valor único.

Aunque, se generaron tres bases de datos de diferentes tamaños (400, 800 y 1600 materiales), nótese que el número total de descriptores incluidos en cada una de ellas se mantuvo fijo en 100, y esto resulta importante en la etapa de análisis de resultados. Para completar la generación de la base de datos sintética, es necesario

asociar cada material incluido a un valor de propiedad objetivo, mediante una correlación entre cada muestra k, expresada por la media y la varianza (valores de $point\_estimator$ y $point\_estimator\_variance$) y los valores de la variable objetivo o target (Figura 4 parte D). Para construir estos valores, en primer lugar se seleccionan al azar una cantidad s fija de DMs. Con la intención de generar diferentes bases de datos esa cantidad s varió en 5, 10 y 20 descriptores, obteniendo así, nueve bases de datos diferentes (incluyendo la variabilidad del número de materiales).

7.2.1.c. CONSTRUCCIÓN DE LA VARIABLE OBJETIVO SINTÉTICA

Se emplearon operaciones matemáticas simples para correlacionar los descriptores seleccionados para construir la variable objetivo (*target*), usando la media aritmética de los descriptores seleccionados. Se crearon dos escenarios: target lineal y target no lineal. En el escenario lineal, el objetivo lineal (Y_L) se generó mediante la fórmula: $Y_L = \sum_{i=1}^{s} D_{p_i'}$ donde s es el número total de descriptores seleccionados. Por otro lado, el target no lineal (Y_{nL}) se calculó usando: $Y_{nL} = \sum_{i=1}^{s}(D_{p_i} * D_{p_{i+1}}) + \cdots + (D_{p_{s-1}} * D_{p_s})$, donde el subíndice en p es la posición del descriptor elegido y s es el número total de descriptores involucrados en la construcción del target. Por ejemplo, cuando los descriptores seleccionados al azar son cinco ($s = 5$), el target no lineal se construye de la siguiente manera: $Y_{nL} = D_{p_1} * D_{p_2} + D_{p_3} * D_{p_4} + D_{p_5}$. La fórmula utilizada para la construcción de los targets se inspiró en las funciones objetivo de etiquetas denominadas *sum y nonlinear* disponibles en RapidMiner para la generación de datos sintéticos [Mierswa & Klinkenberg, 2018]. Un esquema de generación de estos dos escenarios se muestra en la Figura 7.6. En este punto, se generaron bases de datos con 400, 800 y 1600 materiales poliméricos, con un target lineal y no lineal para cada una de las tres cantidades diferentes de descriptores seleccionados (5, 10 y 20). Estas combinaciones proporcionaron 18 escenarios posibles diferentes.

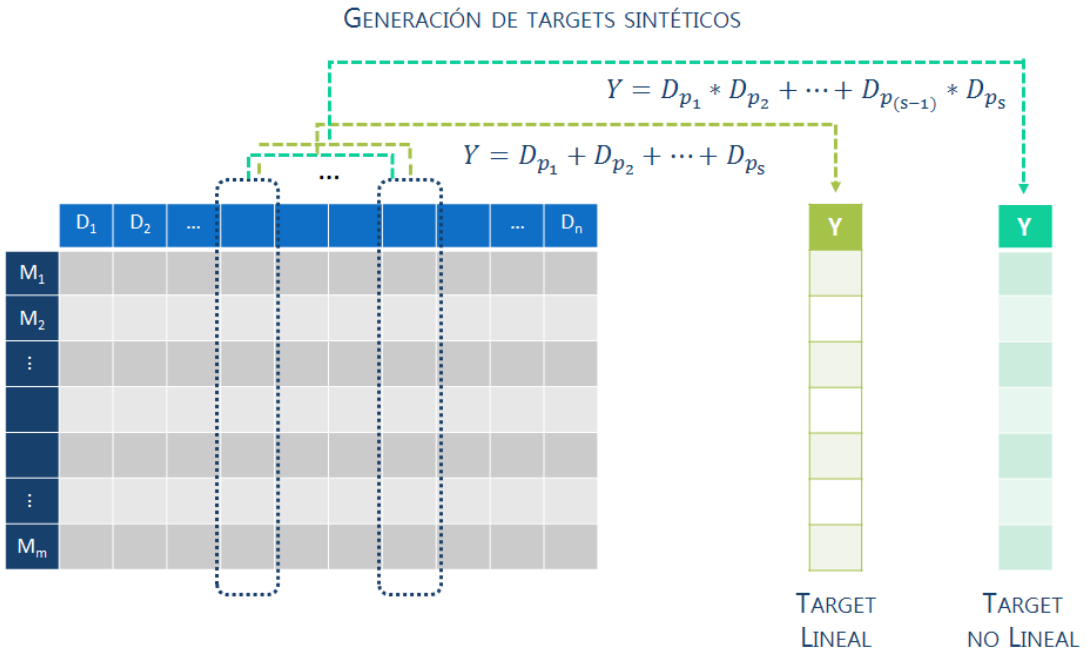


FIGURA 7. 6. GENERACIÓN DE TARGETS LINEALES Y NO LINEALES

7.2.1.d. ESCENARIOS SIN Y CON RUIDO

Para tener más variabilidad en el escenario experimental planteado se decidió agregar ruido a las 18 bases de datos construidas hasta aquí, lo que genera un total de 36 bases de datos diferentes. El escenario con ruido se generó utilizando la biblioteca jitter[2], considerando un porcentaje aleatorio de ruido que varía ± 5%. Se utilizó este porcentaje porque es equivalente al nivel de error de la precisión de cromatografía de exclusión por tamaño utilizada para medir la curva de distribución de pesos moleculares usando el equipo de cromatografía Waters Scientific 150-CV [Pantano *et al.*, 2009]. La incorporación del ruido se realizó para todos los materiales en las bases de datos en un 10% de sus valores de descriptores. Finalmente, en la Figura 7.7 se resume de forma gráfica los principales pasos para la generación de los escenarios creados para evaluar la Tercera Propuesta Alternativa de Representación Computacional de Polímeros [Cravero, *et al.*, 2020].

[2] R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from [https://www.R-project.org]

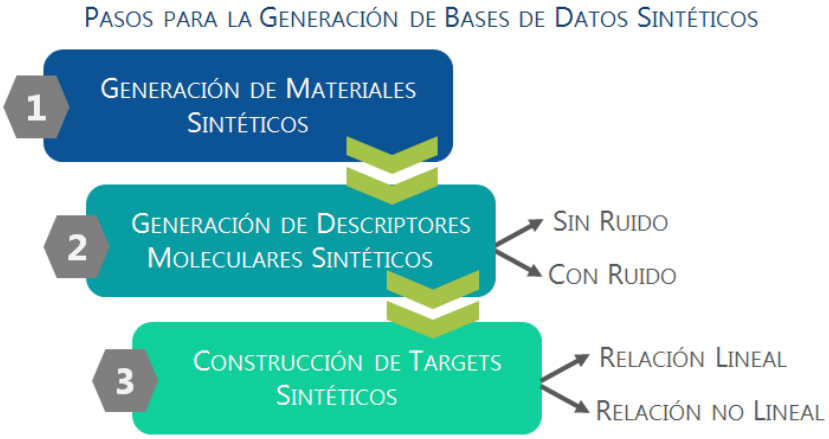


FIGURA 7.7. PASOS PRINCIPALES PARA LA GENERACIÓN DE DATOS SINTÉTICOS.

7.2.2. PROPUESTA DEL ALGORITMO $FS4RV_{DD}$

El primer paso para inferir un modelo QSPR constituye una instancia particular del problema de selección de características. En este caso, las variables bajo análisis presentan incertidumbre y, por lo tanto, es necesario contar con un método de selección de características para tratar con variables caracterizadas por distribuciones. En este sentido, se desarrolló en el marco de esta tesis el algoritmo $FS4RV_{DD}$ para tratar distribuciones discretas como variables de entrada [Cravero *et al.*, 2018b]. Recibe el nombre de $FS4RV_{DD}$ por las iniciales en inglés para Selección de Características para Variables Aleatorias con Distribución Discreta (*Feature Selection for Random Variables with Discrete Distribution*). El algoritmo consta de dos fases claves: la clasificación de los DMs y la eliminación de los DMs en función de la correlación entre ellos. La primera fase se genera en base a las correlaciones lineales entre cada uno de los DMs y la propiedad objetivo. En la segunda, se eliminan los DMs que están altamente correlacionados entre sí (Figura 7.8).

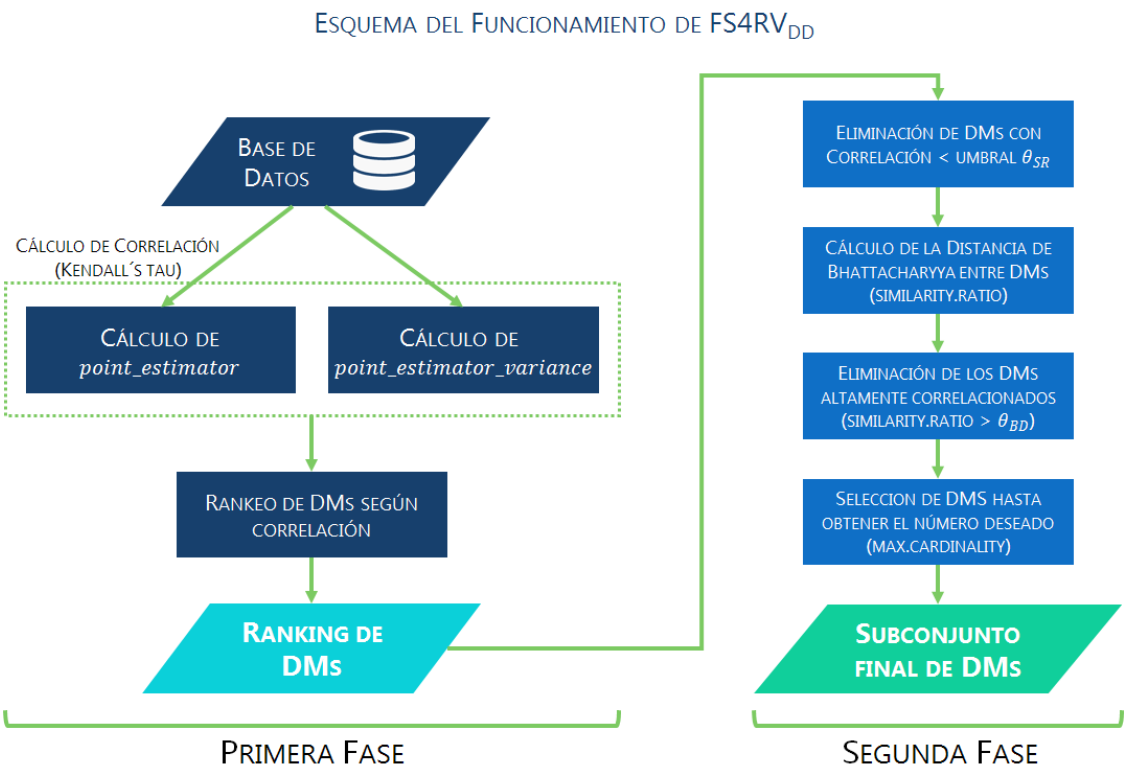


FIGURA 7. 8. METODOLOGÍA GLOBAL DE FS4RV$_{DD}$ PARA LA SELECCIÓN DE CARACTERÍSTICAS.

Para explicar en detalle este algoritmo, deben introducirse las siguientes definiciones. Sea Mat una matriz con m filas (número de materiales) y n columnas (número de descriptores moleculares), donde cada celda C_{ij} contiene la distribución discreta DD_{ij} asociada con el i-ésimo material M_i y el j-ésimo descriptor molecular D_j. Además, sea Y el vector de variable objetivo (target) de longitud m, de modo que y_i sea el valor del target para M_i. Por lo tanto, un modelo QSPR puede definirse como una función de regresión f, tal que $f\ (Mat)\ \cong Y$.

En la primera fase del algoritmo se mide la correlación entre cada DM y el target. Esta correlación se calcula utilizando la distancia Tau de Kendall [McLeod, 2015], en este contexto, el uso de Tau de Kendall es una elección más sólida que, por ejemplo, la correlación de Pearson porque es bien sabido que no es aconsejable su uso cuando las variables tienen distribuciones diferentes [Newson, 2002; Chok, 2010; Mukaka, 2012]. Cada descriptor D_j está representado por k muestras de las distribuciones discretas que el descriptor tiene asociado con cada material polimérico M_i. Luego, se calcula la correlación existente entre r muestras y el valor objetivo, obteniendo una distribución de valores de correlación. Es decir, se hace un muestreo de la distribución, donde r es el número de muestras que se recuperan. A continuación, el método obtiene una medida de correlación para cada descriptor expresada por la media aritmética ($point\_estimator$) y el desvío estándar

($point\_estimator\_variance$) de la distribución de correlación entre el descriptor D_j y el objetivo y calculada para $r = 100$ valores (Figura 7.5, parte D). Finalmente, se genera un ranking de descriptores, ordenándolos de mayor a menor según el valor de $point\_estimator$. Este ranking es el resultado de la primera fase.

En la segunda fase los DMs que tienen un valor de correlación ($point\_estimator$) inferior a un umbral (θ_{SR}) predefinido se eliminan y no forman parte del ranking (Figura 7.8). Luego, el objetivo es obtener un subgrupo de DMs con un nivel de redundancia bajo entre ellos. Para esto, se los compara a fin de identificar aquellos DMs que presentan distribuciones discretas similares. La distancia de Bhattacharyya [Bhattacharyya, 1943] se aplica para comparar las distribuciones de cada par de DMs para cada material. Esta distancia es ampliamente utilizada para estimar la diferencia (distancia) entre dos distribuciones de probabilidad. Si dos DMs presentan un grado de similitud mayor al umbral (θ_{BD}) antes fijado, se elimina aquel DM ubicado más abajo en el ranking.

La selección de características es un campo, del Aprendizaje Maquinal y del reconocimiento de patrones, que consiste en reducir la dimensionalidad de los datos al eliminar aquellas características que son ruidosas, redundantes o irrelevantes para un determinado caso de estudio. Puede pensarse a la selección de características como un problema de ponderación, donde se asignan los pesos 0 o 1 a cada característica [Boyán, 2010]. Como es sabido, la evaluación exhaustiva de la combinatoria de características es prohibitiva, debido a lo que Bellman llamó, en 1957, la maldición de la dimensionalidad (*curse of dimensionality*) [Bellman & Kalaba, 1965]. El criterio de detención, que condiciona la cardinalidad, para este tipo de algoritmos es un tema delicado. Finalmente, se decidió, para facilitar el análisis del rendimiento del método $FSRV_{DD}$, que la cardinalidad del subgrupo final de descriptores seleccionados se defina teniendo en cuenta el número de características utilizadas en la construcción del target.

A continuación se presenta el pseudocódigo del algoritmo *Feature Selection for Random Variables with Discrete Distribution* ($FS4RV_{DD)}$:

ALGORITHM 1: $FS4RV_{DD}$ METHOD

```
Input:
  Mat: training dataset,
  Y: target property values,
  max.card: maximum cardinality,
  min.corr: minimum correlation threshold,
  k: number of values sampled from the discrete distributions,
```

```
  θBD: Bhattacharyya Distance threshold,
  θSR: Similarity ratio threshold.
Output:
  SSF: subset of selected features.
```

PHASE 1: CORRELATION RANKING

```
point.estimators <- array [1..n] of null;
point.estimator.variances <- array [1..n] of null;
For each j varying from 1 to n do:
  correlations <- array [1..K] of null;
  For each k varying from 1 to K do:
    sMDj <- array [1..K] of null;
    For each i varying from 1 to m do:
      sMDj[i] <- random.sampling (Cij)
    end-for
    correlations[k] <- correlation(sMDj, Y) # Kendall's tau
    end-for
  point.estimators[j] <- 1/K ∑_{k=1}^{K} correlati[k]
  point.estimator.variances[j] <- 1/K*(K−1)
        ∑_{k=1}^{sampling.size} (correlati[k] − point. estimator [j])²
end-for
ranking.MD <- sort MDs by decreasing value of their point estimators
for correlations, if two MD have the same point estimator value, put
first in the ranking the MD with lower point estimator variance.
```

PHASE 2: ITERATIVE SUBSET REDUCTION

```
Delete from ranking.MD all MDs with correlation puntot estimator below
the min.corr threshold;
stop <- false;
size.subset <- length(ranking.MD)
For each r1 varying from 1 to (size.subset-1) do:
  While (r2 <= size.subset) do:
    MDA <- ranking.MD[r1]; # the molecular descriptor in the r1
                           # position of the ranking is assigned
    MDB <- ranking.MD[r2];   # the molecular descriptor in the r2
                           #position of the ranking is assigned
    Similar.q <- 0;
      For each i varying from 1 to m do:
        DDAi <- [i,MDA]; # the discrete distribution associated to
                           # MDA in the material i-th is assigned
        DDBi <- [i,MDB]; # the discrete distribution associated to
                           # MDB in the material i-th is assigned
        BDi <- BD(DDAi, DDBi)  # the Bhattacharyya distance between
                               # DDAi and DDBi is assigned
        If (BDi < θBD) then similar.q <- similar.q+1;
      end-for
      similarity.ratio <- similar.q/m;
      if (similarity.ratio > θSR) then remove MDB from ranking.MD;
  end- while
  r2 <- r2+1;
  size.subset <- length(ranking.MD);
end- for
if (size.subset > max.card) then SSF <- ranking.MD[1..max.card]
else SSF <- ranking.MD;
end-algorithm
```

7.2.3. EVALUACIÓN DEL RENDIMIENTO DE $FS4RV_{DD}$

El objetivo de esta sección es determinar la capacidad del método $FS4RV_{DD}$ para detectar los DMs que fueron utilizados (correlacionados) para la construcción del target en el proceso de generación de las bases de datos sintéticas. De esta manera, el enfoque para evaluar $FS4RV_{DD}$ puede analizarse como un problema de clasificación en el que el algoritmo categoriza los DMs en dos clases: *correlacionados* y *no correlacionados* con el target. Por ejemplo, si el target es generado a partir de 5 descriptores correlacionados, de entre los 100 descriptores disponibles, a los que llamaremos arbitrariamente: D_1, D_2, D_3, D_4 y D_5,, el objetivo del método de Selección de Características propuesto es asignar la clase *correlacionados* a estos 5 descriptores y la otra clase (*no correlacionados*) a los 95 restantes. Si se lograra esto, el $FS4RV_{DD}$ obtendría un 100% de precisión. Por otro lado, si el método devolviera como descriptores correlacionados a D_1, D_2, D_3, D_4 y D_{100}, la precisión alcanzada sería del 98% ya que solamente clasificaría mal 2 de los 100 descriptores (D_5: Falso Negativo y D_{100}: Falso Positivo). Por lo tanto, las métricas aplicadas para evaluar su rendimiento pueden ser las utilizadas típicamente en la clasificación binaria. En este caso, se eligieron: el porcentaje de datos correctamente clasificados (%CC) o precisión (*Accuracy*) y la Tasa de No Error, conocido por las siglas en inglés NER para *Non-Error Rate* o también llamada Precisión Equilibrada (*Balanced Accuracy*). El %CC representa el número de predicciones correctas obtenidas sobre el número total de muestras, mientras que el valor de NER es la media aritmética de la sensibilidad de clase, es decir, el promedio de los porcentajes de muestras que se clasificaron correctamente para cada clase. Los resultados de $FS4RV_{DD}$ para escenarios sin ruido pueden encontrarse en la Tabla 7.1 y para escenarios con ruido en Tabla 7.2. Conjuntamente a las métricas nombradas se presenta la media y el desvío estándar (SD) para la correlación entre la selección de descriptores correspondientes a Falsos Positivos (FP) vs. Falsos Negativos (FN). Cuando en las tablas se muestran valores NA (Not Answer) corresponden a la falta de FP, que no permite el cálculo de la correlación pertinente y por lo tanto no existe un valor para mostrar.

TABLA 7. 1. RESULTADOS OBTENIDOS POR $FS4RV_{DD}$ PARA ESCENARIOS SIN RUIDO EN TÉRMINOS DE: PORCENTAJE DE DATOS CORRECTAMENTE CLASIFICADOS (%CC), SENSITIVIDAD, ESPECIFICIDAD Y LA TASA DE NO ERROR (NER). ADEMÁS, SE PRESENTA LA MEDIA Y EL DESVÍO ESTÁNDAR (SD) PARA LA CORRELACIÓN ENTRE LA SELECCIÓN DE DESCRIPTORES.

Materiales			**400**			**800**			**1600**		
Descriptores			**5**	**10**	**20**	**5**	**10**	**20**	**5**	**10**	**20**
Escenario sin ruido	**Target Lineal**	%CC	94%	82%	72%	100%	90%	78%	100%	96%	82%
		Sensibilidad	40%	10%	30%	100%	50%	45%	100%	80%	55%
		Especificidad	96.84%	90%	82.50%	100%	94.44%	86.25%	100%	97.78%	88.75%
		NER	68.42%	50%	56.25%	100%	72.22%	65.63%	100%	88.89%	71.88%
		Mean*	25.06	24.96	25,15	NA	25.48	24.89	NA	25.20	24.98
		SD*	0,.	0.22	0.14	NA	0.29	0.13	NA	0.90	0.10
	Target no Lineal	%CC	96%	86%	68%	98%	92%	78%	98%	96%	90%
		Sensibilidad	60%	30%	20%	80%	60%	45%	80%	80%	75%
		Especificidad	97.89%	92.22%	80%	98.95%	95.56%	86.25%	98.95%	97.78%	93.75%
		NER	78.95%	61.11%	50%	89.48%	77.78%	65.63%	89.48%	88.89%	84.38%
		Mean*	24.81	24.88	25.04	25.88	25.05	24.98	25.19	25.17	25.29
		SD*	0.68	0.31	0.14	0	0.38	0.12	0	1.00	0.25

TABLA 7. 2 RESULTADOS OBTENIDOS POR $FS4RV_{DD}$ PARA ESCENARIOS CON RUIDO EN TÉRMINOS DE: PORCENTAJE DE DATOS CORRECTAMENTE CLASIFICADOS (%CC), SENSITIVIDAD, ESPECIFICIDAD Y LA TASA DE NO ERROR (NER). ADEMÁS, SE PRESENTA LA MEDIA Y EL DESVÍO ESTÁNDAR (SD) PARA LA CORRELACIÓN ENTRE LA SELECCIÓN DE DESCRIPTORES.

Materiales			**400**			**800**			**1600**		
Descriptores			**5**	**10**	**20**	**5**	**10**	**20**	**5**	**10**	**20**
Escenario con Ruido	**Target Lineal**	%CC	96%	90%	80%	100%	94%	78%	100%	98%	84%
		Sensibilidad	60%	50%	50%	100%	70%	45%	100%	90%	60%
		Especificidad	97.89%	94.44%	87.50%	100%	96,67%	86,25%	100%	98,89%	90%
		NER	78.95%	72.22%	68.75%	100%	83.34%	65.63%	100%	94.45%	75%
		Media*	24.31	25.56	25,40	NA	24.68	25.18	NA	24.63	24.85
		SD*	1.36	0.46	0,20	NA	0.60	0.13	NA	0	0.14
	Target no Lineal	%CC	98%	90%	74%	98%	92%	78%	98%	98%	90%
		Sensibilidad	80%	50%	35%	80%	60%	45%	80%	90%	75%
		Especificidad	98.95%	94.44%	83.75%	98.95%	95.56%	86.25%	98.95%	98.89%	93.75%
		NER	89.48%	72.22%	59.38%	89.48%	77.78%	65.63%	89.48%	94.45%	84.38%
		Media*	24.75	25.06	25.22	25.63	25.12	25.07	25.69	24.69	25.05
		SD*	0	0.49	0.15	0	0.34	0.14	0	0	0.15

Para realizar un análisis visual de los resultados obtenidos, a continuación, los presentamos en formato de gráfico de barra. En primer lugar se analiza el escenario sin ruido mostrado en la Figura 7.9, donde puede observarse el porcentaje de DMs correctamente clasificado (%CC). La precisión del algoritmo aumenta a medida que aumenta el tamaño de la base de datos (cantidad de materiales poliméricos), y disminuye cuando

aumenta el número de DMs utilizados para crear los targets sintéticos. Estos resultados se corresponden con el comportamiento esperado. Para las bases de datos de igual tamaño, las clasificaciones erróneas (*no correlacionados*) aumentaron cuando debían recuperarse más DMs, es decir, a medida que el número de DMs utilizados para la generación del target aumentaba. Por otro lado, a medida que las bases de datos se hacen más grandes, los rendimientos mejoran, lo cual es una consecuencia lógica del aumento en el número de polímeros disponibles para detectar los DMs correctamente correlacionados con el target. En otras palabras, cuando hay más datos disponibles, es más fácil detectar un patrón entre ellos (DMs y targets). Con respecto al tipo de correlación utilizada para la generación del target (lineal en la izquierda y no lineal a la derecha de la Figura 7.9), los comportamientos de desempeño del algoritmo fueron bastante similares y no existe un patrón disímil a destacar.

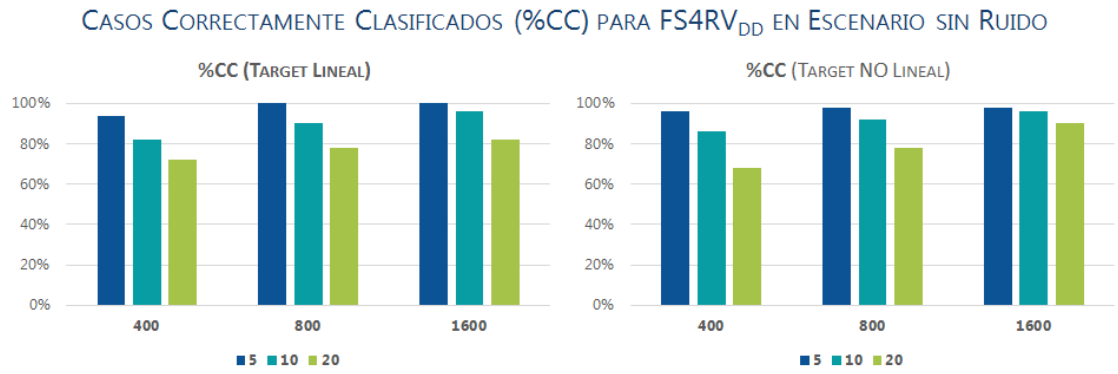


FIGURA 7.9. PORCENTAJE DE DMS CORRECTAMENTE CLASIFICADOS (%CC) POR FS4RV$_{DD}$ PARA EXPERIMENTOS REALIZADOS EN EL ESCENARIO SIN RUIDO.

La Figura 7.10 muestra que la tendencia de los resultados para %CC en el escenario con ruido, tiene tendencias similares a las reportadas en el anterior (sin ruido). La única diferencia, en términos generales, es una ligera mejora en el rendimiento. Los datos experimentales incorrectos tienden a dificultar la inferencia de modelos QSAR/QSPR con alto desempeño, contra intuitivamente, agregar ruido puede mejorar el rendimiento en algunos algoritmos adaptativos [Kay, 2000]. Como los datos originales son suficientemente similares, el agregar ruido podría generar diferencias en los DMs que hagan que el método pueda detectar mejor las diferentes características. Este fenómeno se conoce como resonancia estocástica y es frecuentemente usado para corregir, agregándole o aumentando el ruido blanco, a una señal que normalmente es demasiado débil para ser detectada por un sensor.

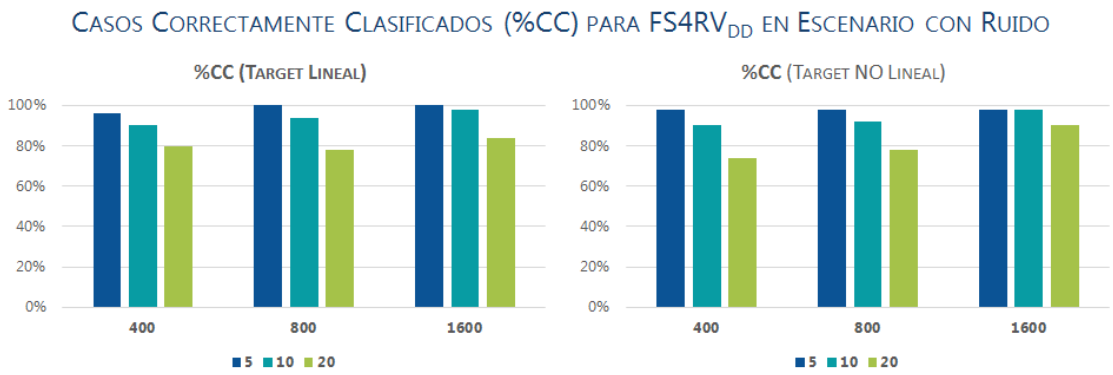


FIGURA 7. 10. PORCENTAJE DE DMS CORRECTAMENTE CLASIFICADOS (%CC) POR FS4RV$_{DD}$ PARA EXPERIMENTOS REALIZADOS EN EL ESCENARIO CON RUIDO.

El porcentaje de datos correctamente clasificados es una de las métricas más utilizadas para evaluar modelos de clasificación, sin embargo, no siempre es adecuada para evaluar un modelo cuando existen desequilibrios entre las clases, porque su valor estará sesgado a la clase más numerosa. Suele considerarse una proporción 1:3 como desbalanceada [Zakharov *et al.*, 2014]. Nuestros experimentos presentan en el escenario más proporcionado una relación 1:4, es decir, 100 DMs totales y 20 correlacionados con el target. Una medida recomendada para corregir o evitar el sesgo de estas situaciones es el uso de la Tasa de No Error (NER), que es considera una métrica imparcial incluso en bases de datos no balanceadas porque considera: falsos positivos (*false positive, FP*), falsos negativos (*false negative, FN*), verdaderos positivos (*true positive, TP*) y verdaderos negativos (*true negative, TN*) [Ballabio *et al.*, 2018].

Los resultados obtenidos para NER se muestran para el escenario sin ruido en la Figura 7.11 y agregando ruido, en la Figura 7.12. Para las bases de datos más pequeñas (400 materiales), en ambos escenarios, el rendimiento disminuye notablemente. En las demás, el rendimiento presenta una disminución menor. Esta reducción general en el rendimiento, en comparación con el presentado por %CC, es esperable y razonable porque NER realiza correcciones en presencia de muestras desbalanceadas y %CC, no lo hace. Como en nuestros experimentos, el desequilibrio es mayor en las bases de datos más pequeñas, es donde más se nota la diferencia. Tal como sucedió con los resultados de %CC, en los escenarios con ruido, el rendimiento mejora frente aquellos a los que no se les ha añadido ruido.

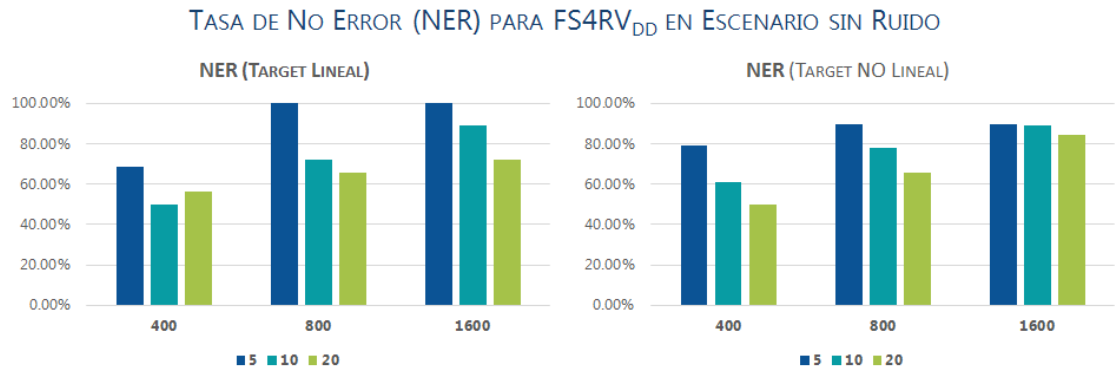


FIGURA 7. 11. MÉTRICAS DE LA TASA DE NO ERROR (NER) OBTENIDAS POR FS4RV$_{DD}$ PARA LOS EXPERIMENTOS REALIZADOS EN EL ESCENARIO SIN RUIDO.

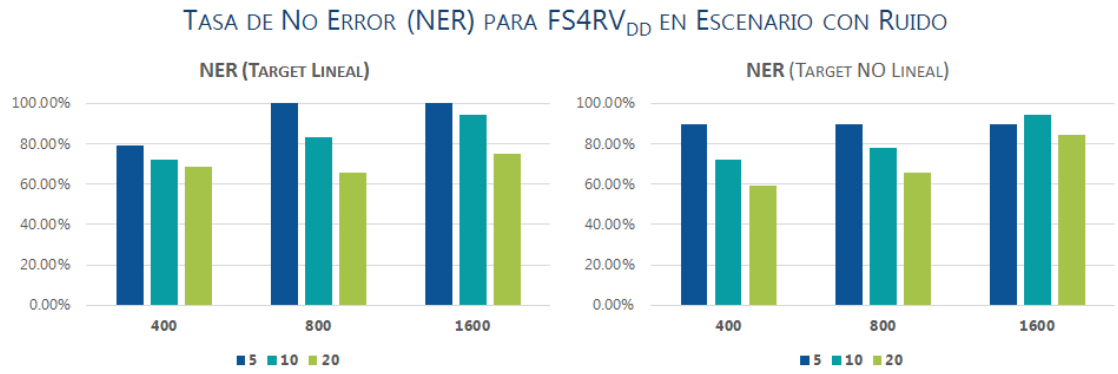


FIGURA 7. 12. MÉTRICAS DE LA TASA DE NO ERROR (NER) OBTENIDAS POR FS4RV$_{DD}$ PARA LOS EXPERIMENTOS REALIZADOS EN EL ESCENARIO CON RUIDO.

En este punto, es relevante descartar la posibilidad de que los DMs incorrectamente seleccionados (FP) por el algoritmo FS4RV$_{DD}$ estuvieran altamente correlacionados con aquellos que no fueron seleccionados por error (FN). Si los FP y FN corresponden a DMs bien correlacionados, esto significaría que el método no cometió un error grave durante el proceso de selección y, por lo tanto, solo reemplazaría las selecciones correctas con buenas alternativas, lo que invalidaría parte de los análisis de resultados realizados hasta aquí. El estudio de correlación realizado nos permitió descartar esta preocupación, porque en todos los casos las correlaciones entre FP y FN son bajas. En la tabla 7.1 se incluyen todos los resultados al respecto, donde la media y el desvío estándar (SD) corresponden a la correlación existente entre FP y FN, cuando aparece NA significa que no existen FP y por lo tanto no puede ser calculada la correlación.

7.2.3.a. COMPARACIÓN DEL DESEMPEÑO DE $FS4RV_{DD}$ VS ENFOQUES UNIVALUADOS

Hasta el momento se evaluó el rendimiento general de la robustez y la escalabilidad del algoritmo $FS4RV_{DD}$. Un desafío adicional es definir un marco experimental justo para realizar comparaciones de desempeño contra enfoques univaluados, ya que no existen métodos similares a $FS4RV_{DD}$ que funcionen con descriptores caracterizados mediante distribuciones. Por esta razón, para comparar el rendimiento del método propuesto con otros similares en la literatura, es necesario realizar algunas consideraciones previas.

En Informática de Polímeros, la representación basada en URE (o monómeros) es la más usada en el modelado QSPR, entonces, se decidió emplear el valor más pequeño de la curva de distribución de pesos, es decir, $start$ (Figura 7.4) como análogo al enfoque basado en la URE, porque es el mínimo valor al que podemos acceder por las decisiones de modelado asumidas; aun sabiendo que en un material polimérico real las cadenas más cortas nunca son tan pequeñas como para estar formadas por una única URE. De manera similar, se usa la media aritmética de la curva como análoga a la representación basada en pesos promedios para poder comparar con las representaciones propuestas recientemente (Mn y Mw) [Cravero *et al.*, 2019b], las cuales ya fueron descriptas en los capítulos anteriores. De esta manera, se podrían definir dos bases de datos adicionales con representaciones de valores univaludados de DMs: una considerando solo el primer valor de la distribución discreta como modelo molecular ($start$), y otra considerando solo el valor de la media aritmética de la distribución discreta ($Med.Art$). La Figura 7.13, muestra un esquema de cómo se proyecta para cada descriptor estas dos representaciones univaluadas a partir de su distribución discreta. Nótese además, que esta transformación no requiere un recálculo del target, ya que en la realidad el valor experimental de una propiedad es el resultado del aporte de toda la distribución de pesos. De este modo, las evaluaciones de desempeño para estas representaciones univaluadas sintéticas se basan en el mismo problema de clasificación planteado en la sección anterior solo que, en este caso, los descriptores se encuentran univaluados.

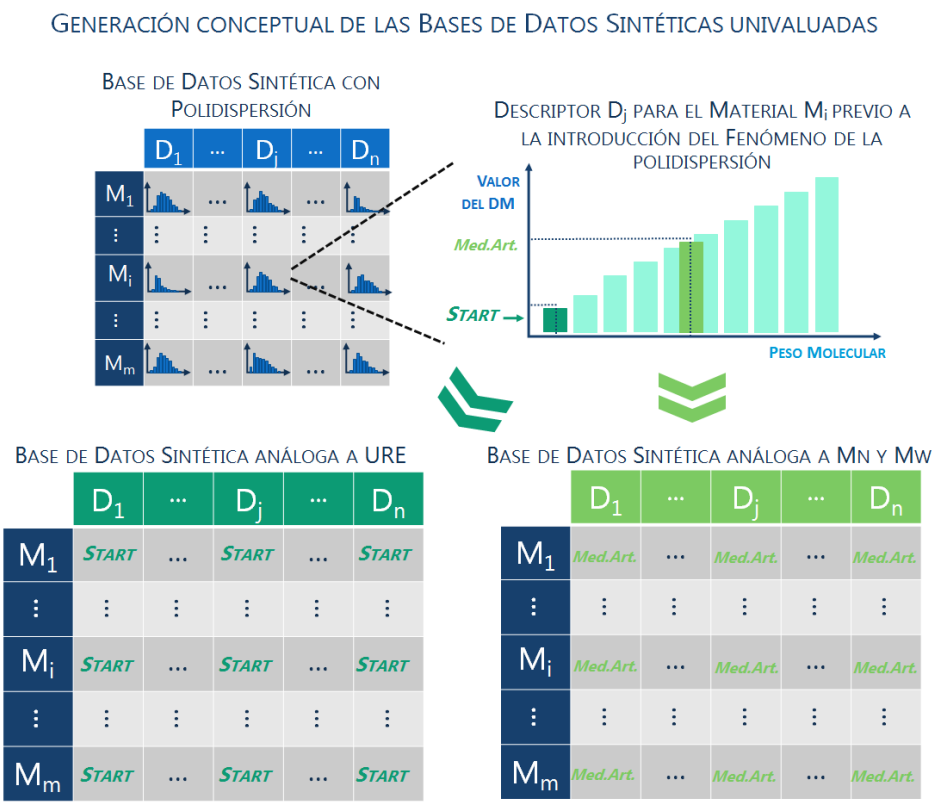


FIGURA 7. 13. ESQUEMA DE LA GENERACIÓN DE LAS BASES DE DATOS SINTÉTICAS UNIVALUADAS A PARTIR DEL PROCESO UTILIZADO PARA LA INTRODUCCIÓN DE LA POLIDISPERSIÓN PARA LA REPRESENTACIÓN MULTIVALUADA.

Con respecto al método de Selección de Características utilizado para estas representaciones de valores únicos, se utilizó el método *AttributesSelection* incluido en Weka [Hall et al., 2009]. Este proporciona una comparación justa y directa con respecto a $FS4RV_{DD}$ en términos de métricas de clasificación. En particular, utilizamos los siguientes parámetros: *CorrelationAttributeEval* como Evaluador de atributos (DMs) y *Ranker* como Método de búsqueda. *CorrelationAttributeEval* evalúa el valor de un atributo midiendo su correlación de Pearson con el target y *Ranker* construye un ranking entre los DMs considerando sus evaluaciones individuales.

La Figura 7.14 muestra gráficamente los resultados de NER para el valor más bajo (análogo a URE) para escenarios sin ruido y la Figura 7.15, los resultados para escenarios con ruido. En la Tabla A.7.1 en Anexos, están disponibles los resultados numéricos. El análisis de estos se convierte en una oportunidad para responder la Quinta Pregunta de Investigación de esta tesis: *¿Es posible identificar con más precisión los DMs más relevantes usando un algoritmo de Selección de Características multivaluadas que usando enfoques tradicionales sobre representaciones univaluadas*?

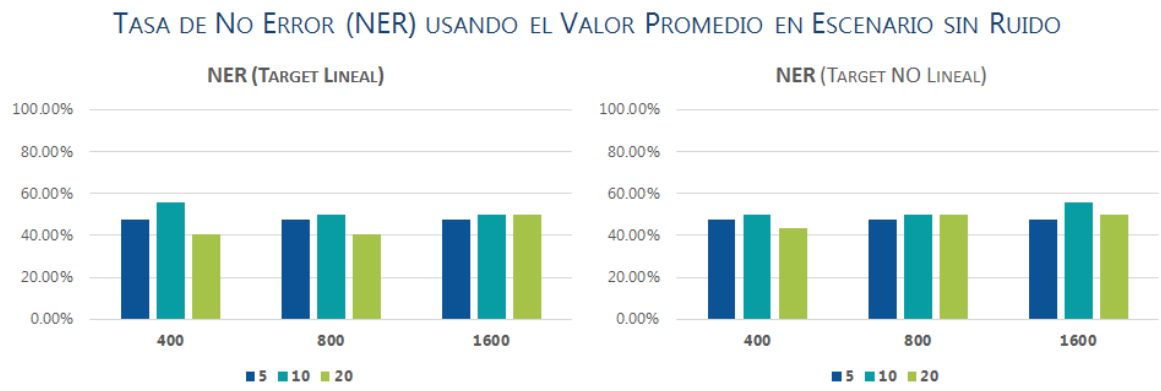


FIGURA 7. 14 TASA DE NO ERROR OBTENIDA MEDIANTE UN MÉTODO TRADICIONAL DE SELECCIÓN DE CARACTERÍSTICAS UTILIZANDO EL VALOR MÁS BAJO (ANÁLOGO A LA URE) PARA EXPERIMENTOS REALIZADOS EN ESCENARIOS SIN RUIDO.

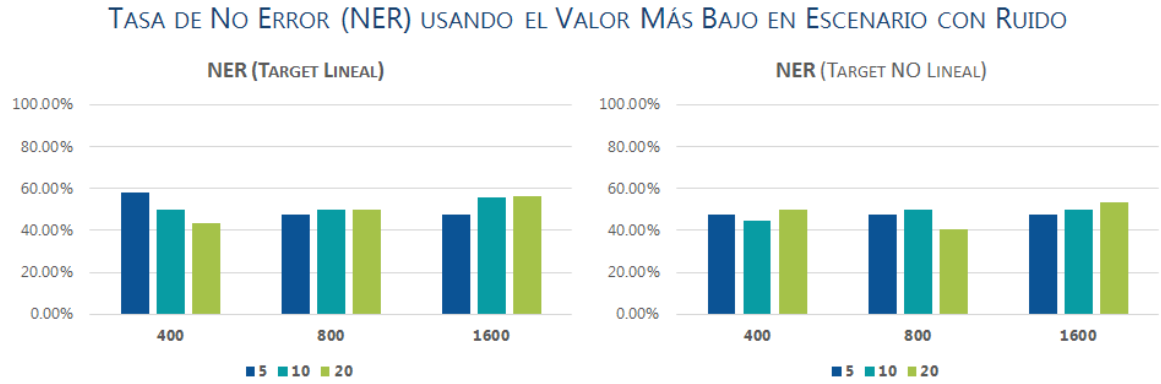


FIGURA 7. 15. TASA DE NO ERROR OBTENIDA MEDIANTE UN MÉTODO TRADICIONAL DE SELECCIÓN DE CARACTERÍSTICAS UTILIZANDO EL VALOR MÁS BAJO (ANÁLOGO A LA URE) PARA EXPERIMENTOS REALIZADOS EN ESCENARIOS CON RUIDO.

Para el escenario sin ruido, es evidente que el rendimiento disminuyó (rango para NER: 40.63-57.90; ver Figura 7.14) en contraste con los resultados logrados por $FS4RV_{DD}$. (rango para NER: 50-100; ver Figura 7.11). Además, no mostró una mejora notoria cuando se agregó ruido (rango para NER: 40.63-57.90; ver Figura 7.15), como si sucedió con la representación multivaluada (rango para NER: 59.38-100; ver Figura 7.12). Por esta razón, podemos concluir que la representación análoga a URE no contiene información suficiente para capturar la complejidad estructural de la polidispersión, lo que dificultaría la identificación de los DMs más relevantes para la variable objetivo en estudio.

Los valores de NER obtenidos utilizando los valores promedios de la curva (análogo a Mn o Mw), tanto para escenarios sin ruido (Figura 7.16) como para los escenarios con ruido (Figura 7.17) fueron, en general, relativamente mayores que los reportados para el mínimo valor (Figura 7.13 y Figura 7.14, respectivamente) y similares a los obtenidos por $FS4RV_{DD}$. Los valores numéricos para ambos escenarios están disponibles en la Tabla A.7.2. Los resultados alcanzados fueron

coherentes con lo esperado, ya que se sabía por trabajos previos [Cravero *et al.*, 2019b] que los pesos promedios son más informativos que la URE en términos de modelos moleculares para generar modelos QSPR.

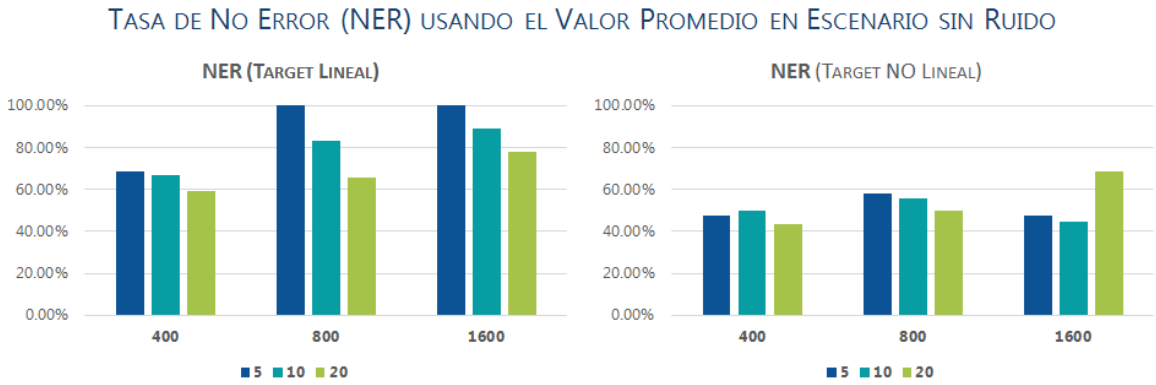


FIGURA 7. 16. TASA DE NO ERROR OBTENIDA MEDIANTE UN MÉTODO TRADICIONAL DE SELECCIÓN DE CARACTERÍSTICAS UTILIZANDO EL VALOR PROMEDIO (ANÁLOGO A MN O MW) PARA EXPERIMENTOS REALIZADOS EN ESCENARIOS SIN RUIDO.

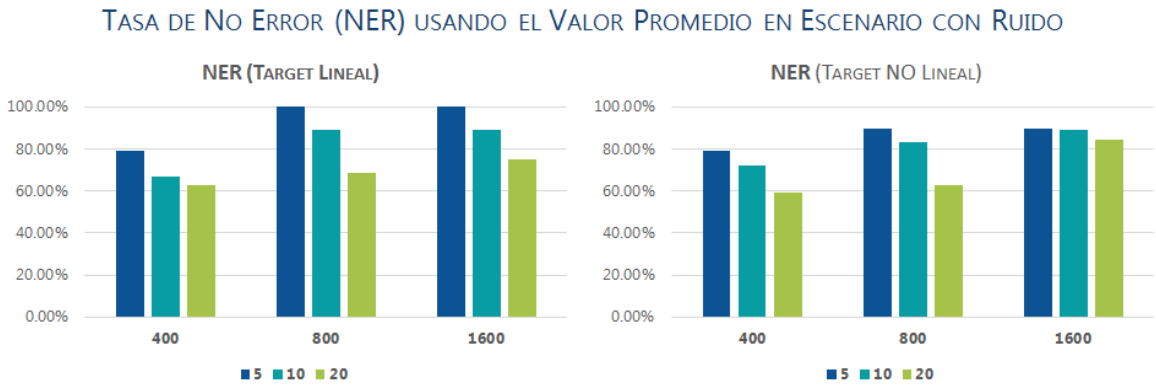


FIGURA 7. 17. TASA DE NO ERROR OBTENIDA MEDIANTE UN MÉTODO TRADICIONAL DE SELECCIÓN DE CARACTERÍSTICAS UTILIZANDO EL VALOR PROMEDIO (ANÁLOGO A MN O MW) PARA EXPERIMENTOS REALIZADOS EN ESCENARIOS CON RUIDO.

Sin embargo, en los escenarios sin ruido y con targets no lineales (parte derecha de la Figura 7.16), los valores de NER obtenidos fueron bajos, comparables con los obtenidos para representaciones basadas en el valor más bajo (parte derecha de la Figura 7.14). Una explicación posible de estos resultados podría ser que esta representación simplificada que resume las distribuciones de los DMs, podría estar requiriendo más esfuerzo para encontrar este tipo de targets más complejos que los presentes en los escenarios lineales. A partir de estos experimentos, podemos concluir que al comparar los rendimientos logrados por ambas representaciones univaluadas con los resultados obtenidos por el método $FS4RV_{DD}$, está claro que la caracterización de los DMs mediante distribuciones discretas es mejor representación para capturar información sobre la polidispersión de este tipo de materiales.

De modo similar a como se hizo para la representación multivaluada, se descartó la posibilidad de que los DMs incorrectamente seleccionados (FP) por el algoritmo de selección univaluada estuvieran altamente correlacionados con aquellos que no fueron seleccionados por error (FN). Los resultados están disponibles en las Tabla A.7.1 y Tabla A.7.2 del Anexo.

7.3. CONCLUSIONES SOBRE LA TERCERA PROPUESTA ALTERNATIVA

La Selección de Características es un problema conocido de optimización combinatorial ampliamente estudiado para diferentes entornos del mundo real. En particular, aquí se ha abordado un problema de selección de características en el área del modelado QSPR en Informática de Polímeros. La característica principal de este escenario es la incertidumbre introducida por el fenómeno de polidispersión de los materiales poliméricos, lo que hace necesario representar los valores de los DMs como distribuciones probabilísticas en vez de un único valor. Se avanzó en una representación probabilística de polímeros y se propuso un nuevo método de selección de características, llamado con el acrónimo $FS4RV_{DD}$. Una evaluación rigurosa del rendimiento del algoritmo $FS4RV_{DD}$ requirió la generación de bases de datos sintéticas para emular el fenómeno de polidispersión, que desempeña un papel central en el diseño de nuevos materiales poliméricos. Diferentes escenarios fueron considerandos para probar la escalabilidad y robustez del método, variando el número de materiales (400, 800 y 1600) y el de DMs seleccionados para construir el target (5, 10 y 20), así como el tipo de correlaciones con el target (lineal y no lineal) y la presencia (o no) de ruido en los datos.

El rendimiento logrado por $FS4RV_{DD}$ fue contrastado con los enfoques tradicionales de Selección de Características utilizados en la Informática de Polímeros, en los que los polímeros se caracterizan típicamente por representaciones univaluadas. En particular, el objetivo principal de este estudio fue responder a la quinta pregunta de investigación: ¿Es posible identificar con más precisión los DMs más relevantes usando un algoritmo de Selección de Características multivaluadas que usando enfoques tradicionales sobre representaciones univaluadas?

El análisis de los resultados de los experimentos proporciona evidencia para responder afirmativamente a esta pregunta, ya que en todos los escenarios el algoritmo $FS4RV_{DD}$ superó o al menos igualó a las otras alternativas.

Síntesis y Conclusiones del Capítulo 7

La distribución de pesos moleculares, distintiva de los materiales poliméricos, hace que una representación computacional basada en una o pocas instancias de peso, se aleje bastante de la realidad. En este capítulo se identifica la necesidad de recuperar teóricamente las curvas de distribución de pesos moleculares y se introduce un método para lograrlo. Además, se presenta un novedoso algoritmo de Selección de Características llamado $FS4RV_{DD}$ (*Feature Selection for Random Variables with Discrete Distribution*) especialmente diseñado para lidiar con la incertidumbre que produce el proceso de polimerización. La metodología propuesta supone que cada DM se caracteriza por una distribución probabilística de los valores asociados con la distribución de los pesos moleculares de los polímeros. El algoritmo $FS4RV_{DD}$ tiene dos fases secuenciales: 1) un rankeo de las características, generado por el análisis de correlación de las mismas con el target, y 2) una reducción iterativa del subconjunto de DMs, obtenida por análisis de redundancia (alta correlación entre par de DMs).

El rendimiento del método fue evaluado utilizando conjuntos de datos sintéticos de diferentes tamaños y variando la cardinalidad de los subconjuntos de DMs seleccionados para construir el target. El principal desafío fue resolver cómo capturar la información de polidispersión, desde la curva de distribución de pesos moleculares, para lograr una representación computacional más efectiva de los materiales poliméricos. Los resultados obtenidos permiten concluir que la representación matemática elegida y el método de selección propuesto son adecuados para manejar la incertidumbre inherente a la polimerización, ya que el algoritmo $FS4RV_{DD}$ superó o igualó a los métodos tradicionales en todos los escenarios propuestos. Esto sugiere la conveniencia de utilizar un algoritmo como $FS4RV_{DD}$ para atacar el problema de incertidumbre que la polidispersión introduce en el modelado QSPR de materiales poliméricos.

CONCLUSIONES GENERALES Y TRABAJOS FUTUROS

CAPÍTULO 8

En este capítulo se presentan las conclusiones generales de este trabajo de tesis. En primer lugar, se describen los aportes realizados en cuanto a la hibridación de características provenientes de procesos de Selección y Aprendizaje de Características. Luego, se analizan los aportes propios más originales de esta tesis en el campo de la Informática de Polímeros. Finalmente, se delinean los trabajos a futuro que derivan de las problemáticas que quedaron sin explorar.

8.1. EN CUANTO AL CUMPLIMIENTO DE OBJETIVOS

Esta tesis doctoral en Ciencias de la Computación titulada "Modelado predictivo de sistemas complejos para Informática Molecular: desarrollo de métodos de selección y aprendizaje de características en presencia de incertidumbre", está enmarcada en el campo de la Informática de Polímeros. El principal desafío es el desarrollo de nuevas representaciones de datos moleculares y algoritmia para resolver problemas vinculados al diseño *in silico* de materiales poliméricos, los cuales constituyen un caso particular de sistemas complejos. En particular, esta tesis presenta cinco hipótesis relacionadas a esta disciplina, en forma de cinco preguntas de investigación, y propone tres alternativas de representación computacional para polímeros de alto peso molecular superadoras a las que se usan, hasta el momento, en la literatura.

En este sentido, además de las nuevas representaciones propuestas, fue necesario abordar el desarrollo de algoritmos de Aprendizaje Maquinal (*Machine Learning*), especialmente ideados para atacar el problema de Selección (*Feature Selection*) de Descriptores caracterizados a través de distribuciones probabilísticas discretas que capturen la incertidumbre inherente a la polidispersión de estas macromoléculas.

En cuanto a los objetivos específicos, en una primera instancia se exploraron, aplicaron y testearon técnicas de Informática Molecular bien establecidas en el diseño racional de fármacos, como es el modelado QSAR/QSPR (Relación Cuantitativa de la Estructura-Actividad/Propiedad) a través de enfoques de Aprendizaje Maquinal, al campo del diseño de materiales poliméricos. Estas

técnicas, como se esperaba, resultaron ser insuficientes, porque los polímeros son moléculas significativamente más complejas que las típicas moléculas pequeñas, en términos de tratamiento computacional. Además, se diseñó una metodología híbrida que trata conjuntamente las características provenientes de técnicas de Selección (*Feature Selection*) y Aprendizaje de Características (*Feature Learning*). Esta metodología híbrida obtuvo resultados positivos al ser aplicada en fármacos, pero no en polímeros.

Al no ser suficientemente informativos los descriptores obtenidos mediante Selección y Aprendizaje de Características para predecir propiedades mecánicas de polímeros de alto peso molecular, se decidió modificar la forma en la que se representaba computacionalmente (modelo molecular) a estos materiales. Para eso se desarrolló una herramienta capaz de polimerizar *in silico* cadenas poliméricas a partir de la Unidad Repetitiva Estructural (URE). Se infirieron modelos QSPR predictivos para la menor instancia de peso (modelo molecular URE) de un polímero, y se los aplicó en instancias mayores de pesos (modelos moleculares Mn y Mw); y viceversa, para evaluar el desempeño de los modelos QSPR inferidos. Esto permitió estudiar los alcances y limitaciones de trabajar con el modelo molecular URE.

Sobre modelos QSPR para la inferencia de propiedades mecánicas de polímeros, se desarrollaron, usando enfoques de Aprendizaje Maquinal, diferentes estrategias que permitieron la entrada de descriptores moleculares trivaluados. Además, se logró representar computacionalmente la polidispersión de polímeros mediante la recuperación teórica de curvas de polidispersión a partir de pesos promedios (Mn y Mw) que definen al material. Todas estas estrategias fueron testeadas y validadas en la predicción de propiedades mecánicas de relevancia para el diseño de nuevos materiales poliméricos: Modulo de Tensión, Elongación a la Rotura y Resistencia a la Rotura.

Por último, se diseñó un algoritmo de selección de características que permiten como entrada variables caracterizadas como una distribución discreta de probabilidad. Para esto fue necesario desarrollar una base de datos sintética que emulara la incertidumbre que genera la polidispersión y que involucrará la caracterización estructural de un polímero por descriptores moleculares multivaluados.

8.2. CONTRIBUCIONES REALIZADAS

Las contribuciones principales de esta tesis se dan en el marco del desarrollo de metodologías computacionales basadas en aprendizaje automático para el modelado QSPR aplicado a la Informática de Polímeros. Sin embargo, es importante destacar que las tecnologías propuestas para representar materiales polidispersos, así como también la nueva algoritmia diseñada para la selección de características multivaluadas, pueden resultar ser fácilmente adaptables a otras aplicaciones que demanden la selección de características en contextos de modelado con incertidumbre. Además, dentro del modelado QSAR, también se efectuaron aportes en la inferencia de nuevos modelos para predecir propiedades de interés para pequeñas moléculas tales como fármacos y COVs (compuestos orgánicos volátiles).

En esta subsección se presentan las conclusiones generales alcanzadas durante el desarrollo de los capítulos 4, 5, 6 y 7, que corresponden al campo de la Informática de Polímeros. Este campo es relativamente nuevo y, por lo tanto, presenta una mayor oportunidad para realizar aportes por ser un nicho poco explorado en comparación con otras áreas de la Informática Molecular. Uno de los principales problemas que enfrenta el área es la baja cantidad de información disponible en bases de datos. Otro problema tecnológico relevante es la ausencia de métodos de cálculo de descriptores moleculares específicos para polímeros que logren trabajar con modelos moleculares de alto peso.

Actualmente, la mayoría de las publicaciones en el ámbito del modelado QSPR ligado a la Informática de Polímeros utilizan la Unidad Repetitiva Estructural (URE) o el monómero para las representaciones moleculares. Hasta donde nuestro grupo sabe, no existen antecedes de modelado QSPR de las propiedades mecánicas derivadas del ensayo de tensión que se trabajan en esta tesis: Modulo de Tensión, Elongación a la Rotura, Resistencia a la Rotura. Además, una cuarta propiedad fue propuesta de esta tesis, la cual recibió el nombre de Comportamiento Dúctil, y que surge de la relación entre dos de las propiedades anteriores.

IMPACTO DE LA UTILIZACIÓN DE LA METODOLOGÍA HÍBRIDA

En esta subsección se busca responder a lo que denominamos Primera Pregunta de Investigación: *¿Puede el aprendizaje de características empleadas en enfoques QSAR resultar de utilidad en el contexto de Informática de Polímeros? ¿Qué sucede si combinamos las ventajas del aprendizaje de características con las*

de la selección de características para la inferencia de modelos QSPR en polímeros de alto peso molecular?

En términos generales, y teniendo en cuenta las bases de datos que se utilizaron en esta tesis, la respuesta a la primera parte de la pregunta es negativa, ya que los rendimientos alcanzados estaban por debajo del azar. En cuanto a la segunda parte, es decir, la aplicación de la Metodología Híbrida en Informática de Polímeros, se ensayaron dos alternativas de hibridación, por un lado la combinación de la totalidad de los descriptores obtenidos por cada una de las técnicas (modelo combinado) y, por el otro, la utilización del conjunto de descriptores obtenidos por Aprendizaje de Características y los descriptores con mayor significado fisicoquímico para la propiedad en estudio (modelo enriquecido). De estos experimentos se pudo concluir que la información provista por ambas técnicas, Selección y Aprendizaje de Características, es complementaria y contribuye a la inferencia de modelos QSPR con razonable desempeño estadístico pero por debajo de los alcanzados en otros dominios de la Informática Molecular. Por lo que, en términos generales, los desempeños alcanzados estuvieron por debajo de lo esperado.

Representación computacional de polímeros polidispersos de alto peso molecular

Uno de los temas más complejos para el modelado computacional de materiales poliméricos está relacionado con la polidispersión que caracteriza a estas estructuras macromoleculares. En el modelado QSPR de polímeros, el modelo molecular ampliamente usado es la URE. Sin embargo, en esta tesis se puso en duda si es esta la representación adecuada para este tipo de polímeros en la predicción de propiedades mecánicas, si es posible encontrar una mejor o si es necesario contar con modelos multivaluados capturando el rasgo distintivo de este tipo de materiales, es decir, su naturaleza de polidispersión de pesos.

A) Representación univaluada en la URE

Con respecto al modelo molecular URE, se persiguió el objetivo de determinar si es una representación adecuada y para esto se formuló la Segunda Pregunta de Investigación: *¿Son los modelos QSPR basados en el modelo molecular URE efectivos cuando se testean sobre modelos moleculares de alto peso? Los descriptores moleculares que fueron seleccionados en los modelos QSPR basados en el modelo molecular URE ¿pueden resultar de utilidad para inferir nuevos modelos QSPR a partir de base de datos de otras instancias univaluadas de representación de mayor peso que URE (Mn y Mw)?*

Para lograr las representaciones basadas en Mn y Mw se desarrolló un algoritmo llamado PolyMaS (*Polymer Maker Smile-based*), capaz de imitar una polimerización del tipo cabeza-cola a partir del código SMILES de un polímero. PolyMaS es útil para obtener a partir de la URE cualquier instancia de peso mayor que se desee realizando una polimerización *in silico* hasta llegar al peso deseado.

Los modelos QSPR inferidos a partir de modelos moleculares URE demostraron no ser precisos al ser aplicados sobre modelos moleculares de alto peso molecular. Además, tampoco resulta recomendable emplear Descriptores Moleculares (DMs) seleccionados desde modelos moleculares URE en el modelado QSPR de modelos moleculares de alto peso molecular. De este modo, se pudieron establecer las limitaciones de la representación computacional basada en URE.

B) Representación univaluada en Mn o Mw

La Tercera Pregunta de Investigación apunta a explorar la Primera Propuesta Alternativa de representación computacional de polímeros de esta tesis, la cual se basada en los modelos moleculares univaluados Mn y Mw. Esta pregunta se compone de cuatro interrogantes o subpreguntas, las dos primeras hacen referencia al modelo molecular Mn y las otras dos, al modelo molecular Mw: *¿Son los modelos QSPR basados en el modelo molecular Mn efectivos cuando se testean sobre modelos moleculares de otro peso (URE y Mw)? Los descriptores moleculares que fueron seleccionados en los modelos QSPR basados en el modelo molecular Mn ¿pueden resultar de utilidad para inferir nuevos modelos QSPR a partir de bases de datos de otras instancias univaluadas de representación (URE y Mw)? ¿Son los modelos QSPR basados en el modelo molecular Mw efectivos cuando se testean sobre modelos moleculares de otro peso (URE y Mn)? Los descriptores moleculares que fueron seleccionados en los modelos QSPR basados en el modelo molecular Mw ¿pueden resultar de utilidad para inferir nuevos modelos QSPR a partir de bases de datos de otras instancias univaluadas de representación (URE y Mn)?*

Los modelos QSPR inferidos a partir de los modelos moleculares Mn y Mw tienen mejor desempeño que los obtenidos con la representación basada en URE. Además, como se esperaba, los modelos QSPR entrenados y evaluados en la misma instancia de peso tienen rendimiento superior que cuando se utiliza el modelo QSPR para testear en otros pesos. Es notorio que los modelos QSPR generados con instancias de alto peso, funcionan mejor cuando se testean con el otro alto peso (Mn testeado con Mw y viceversa) que cuando lo hacen con URE. Esto podría deberse a que estas instancias de peso (Mn y Mw) están numérica y fisicoquímicamente más relacionadas entre sí, que lo que lo están con la URE. Todo

esto confirma la hipótesis de trabajo, acerca de la necesidad de emplear representaciones más realistas para materiales poliméricos.

C) Representación Trivaluada

El modelado parcial de la polidispersión, a través de la Segunda Propuesta Alternativa de Representación computacional de polímeros basada en representaciones trivaluadas formadas por los modelos moleculares URE, Mn y Mw inspiraron la Cuarta Pregunta de Investigación: *¿Existen modelos moleculares basados en sus pesos moleculares promedios, de los materiales, que den como resultado modelos QSPR predictivos más precisos que los obtenidos por modelos moleculares URE? ¿Es aconsejable integrar en una única base de datos los descriptores moleculares correspondientes a modelos moleculares de los diferentes pesos característicos relacionados con las curvas de distribución de peso molecular de los materiales?*

La hipótesis planteada, en forma de preguntas, fue efectivamente verificada y para poder demostrarla se generó una base de datos que contiene información trivaluada de los valores de los Descriptores Moleculares. Puede concluirse que atravesar un proceso de Selección de Características multivaluadas permite la inferencia de modelos QSPR con mejor precisión estadística y habilidades de generalizabilidad que aquellos que se infieren a partir de unir los DMs previamente seleccionados en cada una de las bases de datos univaluadas. Cabe destacar que el subconjunto de DMs seleccionados a partir de la base de datos trivaluada, contiene algunos DMs valuados en más de una instancia. Esto permite acercarse, al menos parcialmente, a representaciones polidispersas para el modelado QSPR de polímeros polidispersos de alto peso molecular. Finalmente, se comprobó que modelos QSPR trivaluados, frente a los modelos QSPR univaluados tomados de forma agregada, logran mejores rendimientos estadísticos para todas las propiedades mecánicas evaluadas.

D) Representación Multivaluada

La naturaleza macromolecular y polidispersa distingue a los polímeros de los demás materiales. La polidispersión hace que no sea siempre una buena decisión de modelado la representación computacional basada en una o pocas instancias de peso para polímeros. En este sentido, se generó un algoritmo capaz de recuperar teóricamente las curvas de distribución de pesos moleculares de un polímero a partir de sus pesos promedios, para avanzar en la Tercera Propuesta Alternativa de modelo molecular mediante una representación multivaluada que permita

trasladar el fenómeno de polidispersión a las distribuciones probabilísticas asociadas a los DMs. Uno de los principales problemas de esta representación es el cálculo de los DMs para las instancias de pesos más altas (2.2×10^5 [g/mol]). Como esto aún se encuentra en etapa de resolución, debió generarse una base de datos sintética, donde cada descriptor está formulado matemáticamente por una distribución discreta de valores de dicho descriptor molecular.

Al tener datos polidispersos, los métodos tradicionales de Selección de Características no pueden ser aplicados a estas nuevas representaciones multivaluadas, ya que solo operan sobre representaciones univaluadas. Como consecuencia, se desarrolló un algoritmo de Selección de Características capaz de tratar el fenómeno de la polidispersión caracterizado por distribuciones probabilísticas, llamado $FS4RV_{DD}$ (*Feature Selection for Random Variables with Discreet Distribution*), el cual es único en su tipo hasta donde sabemos. Diferentes escenarios fueron considerandos para probar el rendimiento general de la escalabilidad y la robustez de este algoritmo.

Para poder evaluar la Quinta Pregunta de Investigación: *¿Es posible identificar con más precisión los DMs más relevantes usando un algoritmo de Selección de Características multivaluadas que usando enfoques tradicionales sobre representaciones univaluadas?*, se preparó un diseño experimental adecuado y amplio, que permitió también evaluar la escalabilidad y robustez frente al ruido del método en diferentes escenarios de correlación lineal y no lineal entre los DMs y el *target*. El rendimiento logrado por $FS4RV_{DD}$ fue contrastado con los enfoques tradicionales de Selección de Características utilizados en la Informática de Polímeros, en los que los polímeros se caracterizan típicamente por representaciones univaluadas. El análisis de los resultados de los experimentos proporciona evidencia para responder afirmativamente a esta quinta pregunta, ya que en todos los escenarios el algoritmo $FS4RV_{DD}$ superó o igualó a los otros enfoques.

8.3. TRABAJOS FUTUROS

Se pretende continuar trabajando, en Informática de Polímeros, en engrosar las bases de datos y en resolver el cálculo de descriptores de cadenas de alto peso de la curva de distribución de pesos moleculares. También, con respecto a modelos univaluados de polímeros, se proyecta implementar en la Herramienta PolyPP un módulo de identificación del Dominio de Aplicabilidad para poder ofrecerla a la comunidad interesada en la predicción de propiedades mecánicas de polímeros basados en la URE. Con respecto a la Metodología Híbrida, ya que esta no tuvo el

desempeño esperado para el modelo molecular URE, se estudiará su comportamiento en otros modelos moleculares univaluados.

Solo dos buenas opciones (Mn y Mw) fueron consideradas para la representación alternativa a la URE en la representación de polímeros, en este sentido podría avanzarse en averiguar si existen otros modelos moleculares mejores, ya sea basados en distintos momentos de la curva de distribución de pesos moleculares, o a través de transformaciones logarítmicas o reducciones porcentuales de los pesos moleculares. Estas representaciones podrían reducir el costo computacional que implica el cálculo de descriptores para modelos moleculares de muy alto peso.

Para la generación de modelos no-univaluados solo se usaron tres modelos moleculares. Esto puede aumentarse con más representantes de peso, para caracterizar de mejor manera a la polidispersión. Si se consiguen generar bases de datos lo suficientemente grandes, podría ensayarse una estrategia de Aprendizaje Profundo para el modelado QSPR.

En cuanto a representaciones multivaluadas, es necesario generar un método de modelado QSPR que permita trabajar con los descriptores seleccionados por $FS4RV_{DD}$. Una idea preliminar para este punto es trabajar con modelos QSPR por consensos. Es decir, si se tiene en cuenta los descriptores por instancia de peso, de a una a la vez, pueden inferirse tantos modelos QSPR como instancias de peso haya, y entre estos modelos podría realizarse un consenso de modelos QSPR obteniendo así un único valor de predicción y no una distribución. Además, será necesario desarrollar un método de Dominio de Aplicabilidad capaz de trabajar con esta representación multivaluada para determinar el espacio dentro del cual el modelo será confiable. Finalmente, una alternativa de trabajo es representar a los polímeros como números difusos, y por ende diseñar e implementar los algoritmos de selección de características, modelado QSPR y dominio de aplicabilidad correspondientes para lidiar con entradas caracterizadas por estos números.

REFERENCIAS

Abraham, M. H., Ibrahim, A., & Acree Jr, 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.

Adams, N. (2010). *Polymer informatics*. In Polymer Libraries (pp. 107-149). Springer, Berlin, Heidelberg.

Adams, N., & Murray-Rust, P. (2008). Engineering Polymer Informatics: Towards the Computer-Aided Design of Polymers. *Macromolecular Rapid Communications*, 29(8), 615-632.

Afantitis, A., Melagraki, G., Makridima, K., Alexandridis, A., Sarimveis, H., & Iglessi-Markopoulou, O. (2005). Prediction of high weight polymers glass transition temperature using RBF neural networks. *Journal of molecular structure*: THEOCHEM, 716(1-3), 193-198.

Aha, D. W., Kibler, D., & Albert, M. K. (1991). Instance-based learning algorithms. *Machine learning*, 6(1), 37-66.

Aines M., Vazquez N., Marin Chinas E., Castro F. M. (2006). *Algunos conceptos Básicos de la química computacional.* Universidad Nacional Autonoma de Mexico. Facultad de Estudios Superiores. Cuautitlaán, México. Comité editorial.

Al-Jarrah, O. Y., Yoo, P. D., Muhaidat, S., Karagiannidis, G. K., & Taha, K. (2015). Efficient machine learning for big data: A review. *Big Data Research*, 2(3), 87-93.

Amaldi, E., & Kann, V. (1998). On the approximability of minimizing nonzero variables or unsatisfied relations in linear systems. *Theoretical Computer Science*, 209(1-2), 237-260.

Arvidson, K. B., Chanderbhan, R., Muldoon-Jacobs, K., Mayer, J., & Ogungbesan, A. (2010). Regulatory use of computational toxicology tools and databases at the United States Food and Drug Administration's Office of Food Additive Safety. *Expert opinion on drug metabolism & toxicology*, 6(7), 793-796.

Askeland, D. R., Fulay, P. P., & Wright, W. J. (2011). *Ciencia e ingeniería de materiales* (p. 952). EUA: Cengage learning.

Audus, D. J., & de Pablo, J. J. (2017). Polymer informatics: opportunities and challenges. *ACS Macro Letters*, vol. 6, pp. 1078-1082, 2017.

Balachandran, P. V., Xue, D., Theiler, J., Hogden, J., Gubernatis, J. E., & Lookman, T. (2018). Importance of Feature Selection in Machine Learning and Adaptive Design for Materials. *In Materials Discovery and Design* (pp. 59-79). Springer, Cham.

Ballabio, D., Grisoni, F., & Todeschini, R. (2018). Multivariate comparison of classification performance measures. *Chemometrics and Intelligent Laboratory Systems*, 174, 33-44.

Barton, D. H. R. (1950). The conformation of the steroid nucleus. *Experientia*, 6(8), 316-320.

Baumann, K., Ecker, G. F., Mestres, J., & Schneider, G. (2011). Molecular Informatics–The First Year. *Molecular informatics*, 30(1), 3-3.

Bellman, R., & Kalaba, R. E. (1965). *Dynamic programming and modern control theory* (Vol. 81). New York: Academic Press.

Bender, A., & Glen, R. C. (2004). Molecular similarity: a key technique in molecular informatics. *Organic & biomolecular chemistry*, 2(22), 3204-3218.

Bhattacharyya A., (1943). On a measure of divergence between two statistical populations defined by probability distributions. *Bulletin of Calcutta Mathematical Society*, vol. 35, pp. 99–109.

Bicerano, J. (2002). *Prediction of polymer properties*. cRc Press.

Bishop, C. M. (1995). Neural networks for pattern recognition. *Oxford university press.*

Block, J. H. (2000). *Topological Indices and Related Descriptors in QSAR and QSPR*. Edited by James Devillers and Alexandru T. Balaban. Gordon and Breach Science Publishers: Amsterdam. ISBN 90-5699-239-2.

Boyán I.B. (2010). *Feature Selection based on Information Theory*. PhD thesis, University of Alicante, Alicante, España.

Brandolin, A., Lacunza, M. H., Ugrin, P. E., Capiati, N. J. (1996). High-pressure polymerization of ethylene. An improved mathematical model for industrial tubular reactors. *Polymer Reaction Engineering*, 4(4), 193-241.

Breiman, L. (1996). Bagging predictors. *Machine learning*, 24(2), 123-140.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1), 5-32.

Breiman, L. (2017). Classification and regression trees. *Routledge*.

Brinkhoff T. (2009) *Real and Synthetic Test Datasets*. In: LIU L., ÖZSU M.T. (eds) Encyclopedia of Database Systems. Springer, Boston, MA.

Brownlee, J. (2016). Master Machine Learning Algorithms: discover how they work and implement them from scratch. *Machine Learning Mastery*.

Buchanan, B. G. (2005). A (very) brief history of artificial intelligence. *Ai Magazine*, 26(4), 53-53.

Cai, J., Luo, J., Wang, S., & Yang, S. (2018). Feature selection in machine learning: A new perspective. *Neurocomputing*, 300, 70-79.

Callister, W. D., & Rethwisch, D. G. (2007). *Materials science and engineering: an introduction* (Vol. 7, pp. 665-715). New York: John wiley & sons.

Cao, C., & Lin, Y. (2003). Correlation between the glass transition temperatures and repeating unit structure for high molecular weight polymers. *Journal of chemical information and computer sciences*, 43(2), 643-650.

Carrasco-Velar R (2003). *Nuevos descriptores atómicos y moleculares para estudios de estructura-actividad: Aplicaciones*. PhD Thesis, La Habana, Cuba.

Chen, M., Jabeen, F., Rasulev, B., Ossowski, M., & Boudjouk, P. (2018). A computational structure–property relationship study of glass transition temperatures for a diverse set of polymers. *Journal of Polymer Science Part B: Polymer Physics*, 56(11), 877-885.

Cherkasov, A., Muratov, E. N., Fourches, D., Varnek, A., Baskin, I. I., Cronin, M., Dearden, J., Gramatica. P., Martin, Y. C., Todeschini, R., Consonni, V., Kuz'min, V. E., Cramer, R., Benigni, R., Yang, C., Rathman, J., Terfloth, L., Gasteiger, J., Richard, A. & Tropsha, A. (2014). QSAR modeling: where have you been? Where are you going to?. *Journal of medicinal chemistry*, 57(12), 4977-5010.

Chok, N. S. (2010). *Pearson's versus Spearman's and Kendall's correlation. Coefficients for continuous data*. Master's Thesis, University of Pittsburgh. Pittsburgh, Pensilvania, Estados Unidos.

Cleary, J. G., & Trigg, L. E. (1995). K*: An instance-based learner using an entropic distance measure. *In Machine Learning Proceedings* 1995 (pp. 108-114

Consonni, V., & Todeschini, R. (2001). *Rational approaches to drug design*. Prous Science: Barcelona.

Consonni, V., & Todeschini, R. (2010). Molecular descriptors. In *Recent advances in QSAR studies* (pp. 29-102). Springer, Dordrecht.

Consonni, V., Todeschini, R., & Pavan, M. (2002a). Structure/response correlations and similarity/diversity analysis by GETAWAY descriptors. 1. Theory of the novel 3D molecular descriptors. *Journal of Chemical Information and Computer Sciences*, 42(3), 682-692.

Consonni, V., Todeschini, R., Pavan, M., & Gramatica, P. (2002b). Structure/response correlations and similarity/diversity analysis by GETAWAY descriptors. 2. Application of the novel 3D molecular descriptors to QSAR/QSPR studies. *Journal of chemical information and computer sciences*, 42(3), 693-705.

Cook, K. A., & Thomas, J. J. (2005). *Illuminating the path: The research and development agenda for visual analytics* (No. PNNL-SA-45230). Pacific Northwest National Lab. (PNNL), Richland, WA (United States).

Craig A. J. (2009). An introduction to the Computer Science and Chemistry of Chemical Information Systems. *eMolecules, Inc*.

Cravero, F; Díaz, M. F.; Ponzoni, I. (2015a) *Modelado Predictivo: Nuevos Descriptores Moleculares para Polímeros*. Congreso Internacional Científico y Tecnológico de la Provincia de Buenos Aires (2 ConCyT2015). La Plata, Argentina.

Cravero, F.; Martínez, M. J., Ponzoni, I.; M.; Vázquez, G. E., Díaz, M. F. (2015b) *Desarrollo de Modelos QSPR Asistido por Técnicas de Analítica Visual para la Predicción de Propiedades Mecánicas de Polímeros Lineales*. Simposio Argentino de Polímeros (SAP 2015). Santa Fe, Argentina.

Cravero, F.; Martínez, M. J., Ponzoni, I.; M.; Vázquez, G. E., Díaz, Mónica F. (2015c) *Predicción del Módulo Elástico Para Polímeros Lineales aplicando Analítica Visual y Aprendizaje Automático*. Simposio Argentino de Polímeros (SAP 2015). Santa Fe, Argentina.

Cravero, F.; Martínez, M. J..; Vázquez, G. E., Díaz, M. F., & Ponzoni, I. (2015d) *Modelado QSPR de Propiedades Mecánicas de Materiales Poliméricos Empleando Técnicas de Reducción de Variables Basadas en Algoritmos de Aprendizaje Automático*. VIII Congreso Argentino de Ingeniería Química (CAIQ 2015). Ciudad Autónoma de Buenos Aires, Argentina.

Cravero, F.; Martínez, M. J.; Díaz, M. F.; Vázquez, G. E.; Ponzoni, I. (2015e). *An integral framework for QSAR Modelling using Computational Intelligence and Visual Analytics*. Fiorella Cravero, M. Jimena Martínez, Mónica F. Díaz, Gustavo E. Vazquez, Ignacio Ponzoni. VI Congreso Argentino de Bioinformática y Biología Computacional (6CAB2C). Bahía Blanca, Argentina.

Cravero, F; Martínez, M. J.; Vázquez, G. E., Díaz, M. F., & Ponzoni, I. (2016a). Feature learning applied to the estimation of tensile strength at break in polymeric material design. *Journal of integrative bioinformatics*, 13(2), 15-29.

Cravero, F.; Martínez, M. J.; Vázquez, G. E., Díaz, M. F., & Ponzoni, I. (2016b) Intelligent Systems for Predictive Modelling in Cheminformatics: QSPR Models for Material Design using Machine Learning and Visual Analytics Tools. 10th International Conference on Practical Applications of Computational Biology & Bioinformatics (PACBB 2016). *Advances in Intelligent Systems and Computing*, Vol. 477, pp. 3-11 (2016). Springer-Verlag.

Cravero, F., Vázquez, G. E., Ponzoni, I., Díaz, M. F. (2016c). *Modelado molecular de materiales poliméricos en quimioinformatica*. Simposio Argentino de Materiales (SAM-CONAMET 2016). Córdoba, Argentina.

Cravero, F., Martínez, M. J., Vázquez, G. E., Ponzoni, I., Díaz, M. F. (2016d). *Representación de la Estructura Molecular de Polímeros Sintéticos de Alto Peso*. 31° Congreso Argentino de Química, Buenos Aires, Argentina.

Cravero, F., Martínez, M. J., Vázquez, G. E., Ponzoni, I., Díaz, M. F. (2016e). *Predicción de curvas teóricas de distribución de peso molecular de resinas poliméricas*. 31° Congreso Argentino de Química. Buenos Aires, Argentina.

Cravero, F., Ponzoni, I., Diaz, M. F. (2016f). *Informática Molecular: Polímeros modelados por Distribución de Pesos*. III Congreso Internacional de Ciencia y Tecnología de la Provincia de Buenos Aires. La Plata, Argentina.

Cravero, F., Martínez, M. J., Vázquez, G. E., Díaz, M. F., Ponzoni, I. (2017a). QSAR Classification Models for Predicting Affinity to Blood or Liver of Volatile Organic Compounds in e-Health. 5th International Work-Conference on Bioinformatics and Biomedical Engineering. IWBBIO 2017. *Lecture Notes in Computer Science*, Vol. 10209, pp. 424-433. Springer-Verlag. ISSN: 0302-9743.

Cravero, Fiorella; Schustik, Santiago; Martínez, Jimena; Ponzoni, Ignacio & Díaz, Mónica F. (2017b) *Herramienta Computacional para Testeo Virtual de Propiedades Mecánicas de Polímeros. Etapa de Ensamble y Prueba*. Simposio Argentino de Polímeros (SAP 2017). Los Cocos, Córdoba.

Cravero, F., Schustik, S. A., Martínez, M. J., Ponzoni, I., Díaz, M. F. (2017c). *Macro Approach to Molecular Modelling of Linear Polymers Applied to Estimation of Tensile Modulus for New Materials Development*. VIII International Symposium on Materials (Materias 2017). Aveiros, Portugal.

Cravero, F., Schustik, S., Martínez, M. J., Barranco, C. D., Díaz, M. F., Ponzoni, I. (2018a). *Feature Selection and Polydispersity Characterization for QSPR Modelling: Predicting a Tensile Property*. In International Conference on Practical Applications of Computational Biology & Bioinformatics (pp. 43-51). Springer, Cham.

Cravero, F., Schustik, S., Martínez, M. J., Díaz, M. F., & Ponzoni, I. (2018b). $FS4RV_{DD}$: A feature selection algorithm for random variables with discrete distribution. In International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems. *Communications in Computer and Information Science*, Vol. 855, pp. 211-222. Springer-Verlag.

Cravero, F., Martínez, M. J., Ponzoni, I., & Díaz, M. F. (2019a). Computational modelling of mechanical properties for new polymeric materials with high molecular weight. *Chemometrics and Intelligent Laboratory Systems*, 193, 103851.

Cravero, F., Schustik, S. A., Martínez, M. J., Barranco, C. D., Díaz, M. F., Ponzoni, I. (2019b). Computer-aided design of polymeric materials: Computational study for characterization of databases for prediction of mechanical properties under polydispersity. *Chemometrics and Intelligent Laboratory Systems*, 191, 65-72.

Cravero F., Schustik S., Martínez M.J., Vázquez G., Díaz M., Ponzoni I. (2020). Feature Selection for Polymer Informatics: Evaluating Scalability and Robustness of

the FS4RV$_{DD}$ algorithm using Synthetic Polydisperse Datasets. *Journal of Chemical Information and Modeling*, 60 (2), 592-603.

Cypcar, C. C., Camelio, P., Lazzeri, V., Mathias, L. J., & Waegell, B. (1996). Prediction of the glass transition temperature of multicyclic and bulky substituted acrylate and methacrylate polymers using the energy, volume, mass (EVM) QSPR model. *Macromolecules*, 29(27), 8954-8959.

Dashtbozorgi, Z., & 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.

de Pablo, J. J., Jackson, N. E., Webb, M. A., Chen, L. Q., Moore, J. E., Morgan, D., ... & Analytis, J. (2019). New frontiers for the materials genome initiative. npj *Computational Materials*, 5(1), 41.

Devinyak, O., Havrylyuk, D., & Lesyk, R. (2014). 3D-MoRSE descriptors explained. *Journal of Molecular Graphics and Modelling*, 54, 194-203.

Domingos, P. (1999). The role of Occam's razor in knowledge discovery. *Data mining and knowledge discovery*, 3(4), 409-425.

Dorronsoro, I., Chana, A., Abasolo, M. I., Castro, A., Gil, C., Stud, M., & Martinez, 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. *QSAR & Combinatorial Science*, 23(2-3), 89-98.

DRAGON for Windows (Software for Molecular Descriptor Calculations), Version 5.5; Talete srl: Milan, Italy, 2007.

Duce, C., Micheli, A., Starita, A., Tiné, M. R., Solaro, R. (2006). Prediction of polymer properties from their structure by recursive neural networks. *Macromolecular rapid communications*, 27(9), 711-715.

Ebrahimpour, M. K., Zare, M., Eftekhari, M., & Aghamolaei, G. (2017). Occam's razor in dimension reduction: Using reduced row Echelon form for finding linear independent features in high dimensional microarray datasets. *Engineering Applications of Artificial Intelligence*, 62, 214-221.

Faulon, J. L., & Bender, A. (2010). *Handbook of chemoinformatics algorithms*. CRC press.

Frank, E., Hall, M., Trigg, L., Holmes, G., & Witten, I. H. (2004). Data mining in bioinformatics using Weka. *Bioinformatics*, 20(15), 2479-2481

Friedman, J., Hastie, T., & Tibshirani, R. (2000). *Additive logistic regression: a statistical view of boosting*. The annals of statistics, 28(2), 337-407.

García-Domenech, R., & de Julián-Ortiz, J.V. (2002). Prediction of indices of refraction and glass transition temperatures of linear polymers by using graph theoretical indices. *The Journal of Physical Chemistry B*, 106(6), 1501-1507.

García-Domenech, R., Gálvez, J., de Julián-Ortiz, J. V., & Pogliani, L. (2008). Some new trends in chemical graph theory. *Chemical Reviews*, 108(3), 1127-1169.

Glen, R., & Aldridge, S. (2002). Developing tools and standards in molecular informatics. *Chemical Communications*, 2002(23), 2745-2747.

Gola, J., Obrezanova, O., Champness, E., Segall, M. (2006) ADMET property prediction: the state of the art and current challenges. *QSAR & Comb. Science*, 25, pp. 1172-1180, Wiley, 2006.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT press.

Grisoni, F., Consonni, V., & Todeschini, R. (2018). Impact of molecular descriptors on computational models. In *Computational Chemogenomics* (pp. 171-209). Humana Press, New York, NY.

Guha, R. (2007). Chemical informatics functionality in R. *Journal of Statistical Software*, 18(5), 1-16.

Gupta, K. (2014). Qsar Studies on Gallic Acid Derivatives and Molecular Docking Studies of Bace1 Enzyme–A Potent Target of Alzheimer Disease. *Bioscience & Engineering: An International Journal*, 1 (1), 11-27.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & 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. & Fujita, T. (1964). p-σ-π analysis: a method for the correlation of biological activity and chemical structure. *J. Am. Chem*. Soc. 86, 1616–1626.

Hansch, C., Maloney, P. P., Fujita, T., & Muir, R. M. (1962). Correlation of biological activity of phenoxyacetic acids with Hammett substituent constants and partition coefficients. *Nature*, 194(4824), 178.

Hansch, C., Unger, S. H., & Forsythe, A. B. (1973). Strategy in drug design. Cluster anlysis as an aid in the selection of substituents. *Journal of medicinal chemistry*, 16(11), 1217-1222.

Hill, J., Mulholland, G., Persson, K., Seshadri, R., Wolverton, C., & Meredig, B. (2016). *Materials science with large-scale data and informatics: unlocking new opportunities*. Mrs Bulletin, 41(5), 399-409.

Hira, Z. M., & Gillies, D. F. (2015). A review of feature selection and feature extraction methods applied on microarray data. *Advances in bioinformatics*, 2015..

Holdren, J. P. (2011). *Materials genome initiative for global competitiveness. National Science and technology council*. Executive Office of the President, National Science and Technology Council. Washington, USA.

Horvath, A.L. (1992). *Molecular Design, Chemical Structure Generation from the Properties of Pure Organic Compounds*, Number 75 in Studies in Physical and Theoretical Chemistry, 1st edn., Elsevier, Amsterdam.

Hughes, K.; Paterson, J.; Meek, M. E. (2009). Tools for the Prioritization of Substances on the Domestic Substances List in Canada on the Basis of Hazard Regul. *Toxicol. Pharmacol*, 55, 382– 393

Hulten, G., Spencer, L., & Domingos, P. (2001, August). *Mining time-changing data streams*. In Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 97-106). ACM.

Imbalzano, G., Anelli, A., Giofré, D., Klees, S., Behler, J., & Ceriotti, M. (2018). Automatic selection of atomic fingerprints and reference configurations for machine-learning potentials. *The Journal of chemical physics*, 148(24), 241730.

Jabeen, F., Chen, M., Rasulev, B., Ossowski, M., Boudjouk, P. (2017). Refractive indices of diverse data set of polymers: A computational QSPR based study. *Computational Materials Science*, 137, 215-224.

Jennings, P. C., Lysgaard, S., Hummelshøj, J. S., Vegge, T., Bligaard, T. (2019). Genetic algorithms for computational materials discovery accelerated by machine learning. npj *Computational Materials*, 5(1), 46.

John, G. H., & Langley, P. (1995, August). *Estimating continuous distributions in Bayesian classifiers*. In Proceedings of the Eleventh conference on Uncertainty in artificial intelligence (pp. 338-345). Morgan Kaufmann Publishers Inc.

Johnson, M. A., & Maggiora, G. M. (1990). *Concepts and applications of molecular similarity*. Wiley.

Kalousis, A., Prados, J., & Hilario, M. (2007). Stability of feature selection algorithms: a study on high-dimensional spaces. *Knowledge and information systems*, 12(1), 95-116.

Karelson, M. (2000). *Molecular descriptors in QSAR/QSPR*. Wiley-Interscience.

Katritzky, A. R., Rachwal, P., Law, K. W., Karelson, M., & Lobanov, V. S. (1996). Prediction of polymer glass transition temperatures using a general quantitative structure− property relationship treatment. *Journal of chemical information and computer sciences*, 36(4), 879-884.

Katritzky, A.R., Sild, S., Lobanov, V.,Karelson, M.(1998). Quantitative structure-property relationship (QSPR) correlation of glass transition temperatures of high molecular weight polymers. *J. Chem Inf and Comp Sci*, 38(2), 300-304.

Kay, S. (2000). Can detectability be improved by adding noise?. *IEEE signal processing letters*, 7(1), 8-10.

Keim, D. A. (2001). Visual exploration of large data sets. *Communications of the ACM*, 44(8), 38-44.

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

Khaire, U. M., & Dhanalakshmi, R. (2019). Stability of feature selection algorithm: A review. *Journal of King Saud University-Computer and Information Sciences*

Kier, L. B., & Hall, L. H. (1999). *Molecular structure description*. Academic.

Klopman, G. (1984). Artificial intelligence approach to structure-activity studies. Computer automated structure evaluation of biological activity of organic molecules. *Journal of the American Chemical Society*, 106(24), 7315-7321.

Kubinyi, H. (1977). Quantitative structure-activity relations. 7. The bilinear model, a new model for nonlinear dependence of biological activity on hydrophobic character. *Journal of medicinal chemistry*, 20(5), 625-629.

Kunal R., Supratik K., Rudra N. D. (2015). *Understanding the Basics of QSAR for Applications in Pharmaceutical Sciences and Risk Assessment*. Chapter 2 - Chemical Information and Descriptors, Pages 47-80. Academic press.

Kutz, Myer. (2002). *Handbook of Materials Selection*. first ed., John Wiley & Sons, New York, United States

Laxmi, D., & Priyadarshy, S. (2002). HyperChem 6.03. *Biotech Software & Internet Report: The Computer Software Journal for Scientists*, 3(1), 5-9.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). *Deep learning*. nature, 521(7553), 436-444,

Leo, A., Jow, P. Y. C., Silipo, C., & Hansch, C. (1975). Calculation of hydrophobic constant (log P) from. pi. and f constants. *Journal of medicinal chemistry*, 18(9), 865-868.

Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R. P., Tang, J., & Liu, H. (2018). Feature selection: A data perspective. *ACM Computing Surveys* (CSUR), 50(6), 94.

Liu, W., & Cao, C. (2009). Artificial neural network prediction of glass transition temperature of polymers. *Colloid and Polymer Science*, 287(7), 811-818.

Livingstone, D. J. (2000). The characterization of chemical structures using molecular properties. A survey. *Journal of chemical information and computer sciences*, 40(2), 195-209.

Livingstone, D. J., Hesketh, G., & 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.

Lorenzini, P., Pons, M., & Villermaux, J. (1992). Free-radical polymerization engineering—III. Modelling homogeneous polymerization of ethylene:

mathematical model and new method for obtaining molecular-weight distribution. *Chemical engineering science*, 47(15-16), 3969-3980.

Ma H, Melillo G, Oliva L et al. (2005) Aluminum alkyl complexes supported by [OSSO] type bisphenolato ligands: synthesis, characterization and living polymerization of rac-lactide. *Dalton Trans* 721–727.

Ma, R., Liu, Z., Zhang, Q., Liu, Z., Luo, T. (2019). Evaluating Polymer Representations via Quantifying Structure-Property Relationships. *Journal of Chemical Information and Modeling*, 59, 7, 3110-3119.

Mannodi-Kanakkithodi, A., Chandrasekaran, A., Kim, C., Huan, T. D., Pilania, G., Botu, V., & Ramprasad, R. (2018). Scoping the polymer genome: A roadmap for rational polymer dielectrics design and beyond. *Materials Today*, 21(7), 785-796.

Martin, J. R., Johnson, J. F., & Cooper, A. R. (1972). Mechanical properties of polymers: the influence of molecular weight and molecular weight distribution. *Journal of Macromolecular Science-Reviews in Macromolecular Chemistry*, 8(1), 57-199.

Martínez, M. J.; Cravero, F; Vázquez, G. E.; Díaz, M. F.; Soto, A. J.; Ponzoni, I. (2014a). *Interactive Visual Analysis Methodology for Improving Descriptor Selection in QSPR: First Steps* (3 pág.) V Congreso Argentino de Bioinformática y Biología Computacional (5CAB2C). San Carlos de Bariloche, Argentina.

Martínez, M. J; Cravero, F.; Palomba, D.; Soto, A. J.; Díaz, M. F. Vázquez, G. E.; Ponzoni, I. (2014b). *Feature Selection in Molecular Informatics: Improving QSAR/QSPR Modeling by Computational Intelligence Approaches and Interactive Visual Analysis*. Simposio Argentino de Inteligencia Artificial (43JAIIO). Ciudad Autónoma de Buenos Aires, Argentina.

Martínez, M. J., Ponzoni, I., Díaz, M. F., Vazquez, G. E., & Soto, A. J. (2015). Visual analytics in cheminformatics: user-supervised descriptor selection for QSAR methods. *Journal of cheminformatics*, 7(1), 39.

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

McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5(4), 115-133.

McLeod, A.I. (2015). *Package 'Kendall'*. Retrieved from [https://cran.r-project.org/web/packages/Kendall/Kendall.pdf]

Meier, M.A.R. & Webster, D.C. (2009). *Polymer Libraries*. Springer-Verla, Berlin Heidelberg g. e-ISBN 978-3-642-00170-3

Meijer, H. E., & Govaert, L. E. (2005). Mechanical performance of polymer systems: The relation between structure and properties. *Progress in polymer science*, 30(8-9), 915-938.

Mierswa, I., & Klinkenberg, R. (2018). *RapidMiner Studio* (9.1) [Data science, machine learning, predictive analytics]. Retrieved from [https://rapidminer.com/]

Miljković, D. (2017). *Brief review of self-organizing maps*. In 2017 40th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO) (pp. 1061-1066). IEEE.

Mitchell, T. M. (1997). Does machine learning really work?. *AI magazine*, 18(3), 11-11.

Monteiro, M. J. (2015). Fitting molecular weight distributions using a log-normal distribution model. *European Polymer Journal*, 65, 197-201.

Moriwaki, H., Tian, Y. S., Kawashita, N., & Takagi, T. (2018). Mordred: a molecular descriptor calculator. *Journal of cheminformatics*, 10(1), 4.

Mukaka, M. M. (2012). A guide to appropriate use of correlation coefficient in medical research. *Malawi Medical Journal*, 24(3), 69-71.

Nettles, J. H., Jenkins, J. L., Bender, A., Deng, Z., Davies, J. W., & Glick, M. (2006). Bridging chemical and biological space: "target fishing" using 2D and 3D molecular descriptors. *Journal of medicinal chemistry*, 49(23), 6802-6810.

Newson, R. (2002). Parameters behind "nonparametric" statistics: Kendall's tau, Somers' D and median differences. *The Stata Journal*, 2(1), 45-64.

Nosengo, N. (2016). Can artificial intelligence create the next wonder material?. *Nature News*, 533(7601), 22.

O'Boyle, N. M., Banck, M., James, C. A., Morley, C., Vandermeersch, T., & Hutchison, G. R. (2011). Open Babel: An open chemical toolbox. *Journal of cheminformatics*, 3(1), 33.

OECD Principles (2013). *Organisation for Economic Co-Operation and Development OECD Principles for the Validation, for Regulatory Purposes, of (Quantitative) Structure -Activity Relationship Models*. Disponible en: http://www.oecd.org/chemicalsafety/risk-assessment/37849783.pdf

Otsuka, S., Kuwajima, I., Hosoya, J., Xu, Y., & Yamazaki, M. (2011). *PoLyInfo: Polymer database for polymeric materials design*. In 2011 International Conference on Emerging Intelligent Data and Web Technologies (pp. 22-29). IEEE.

Palomba D. (2014). *Predicción de propiedades de sustancias y materiales de interés en la industria química a través del desarrollo de métodos computacionales*. PhD thesis, UNS, Bahía Blanca, Argentina.

Palomba, D., Martínez, M.J., Ponzoni, I., Díaz, M.F., Vazquez, G.E., Soto, A.J. (2012a). QSPR models for predicting log pliver values for volatile organic

compounds combining statistical methods and domain knowledge. *Molecules* 17(12), 14937–14953

Palomba, D., Vazquez, G. E., & Díaz, M. F. (2012b). Novel descriptors from main and side chains of high-molecular-weight polymers applied to prediction of glass transition temperatures. *Journal of Molecular Graphics and Modelling*, 38, 137-147.

Palomba, D., Vazquez, G. E., & Díaz, M. F. (2014a). Prediction of elongation at break for linear polymers. *Chemometrics and Intelligent Laboratory Systems*, 139, 121-131.

Palomba; D; Cravero; F.; Vázquez; G. E.; Díaz; M. F. (2014b) *Prediction of tensile strength at break for linear polymers applied to new materials development*. Congreso Internacional de Metalurgia y Materiales (SAM-CONAMET) / Iberomat / Simposio Materia 2014.Santa Fe, Argentina.

Palomba; D.; Cravero; F.; Vázquez; G. E.; Díaz; M. F. (2014c) *Prediction of tensile modulus for linear polymers applied to new materials development*. Congreso Internacional de Metalurgia y Materiales (SAM-CONAMET) / Iberomat / Simposio Materia 2014. Santa Fe, Argentina.

Palomba, D; Martínez M. J.; Cravero F.; Soto A., Vazquez, G. E.; Ponzoni, I., Díaz, M. F. (2014d) Prediction of mechanical properties of tensile test for linear polymers. QSPR modeling with computational intelligence and interactive visual analysis. *Journal of the Argentine Chemical Society*, Vol. 101(1-2), pp. 137-147. Asociación Argentina de Química.

Pantano, I. A. G., Díaz, M. F., Brandolin, A., Sarmoria, C. (2009). Mathematical modeling of the catalytic degradation of polystyrene in the presence of aluminum chloride. *Polymer Degradation and Stability*, 94(4), 566-574.

Peerless, J. S., Milliken, N. J., Oweida, T. J., Manning, M. D., & Yingling, Y. G. (2019). Soft matter informatics: current progress and challenges. *Advanced Theory and Simulations*, 2(1), 1800129.

Petrosyan, L. S., Sizochenko, N., Leszczynski, J., & Rasulev, B. (2019). Modeling of Glass Transition Temperatures for Polymeric Coating Materials: Application of QSPR Mixture-based Approach. *Molecular informatics*.

Piramuthu, S. (2004). Evaluating feature selection methods for learning in data mining applications. *European journal of operational research*, 156(2), 483-494.

Ponzoni, I., Sebastián-Pérez, V., Martínez, M. J., Roca, C., De la Cruz Pérez, C., Cravero, F., Vazquez, G. E., Páez, J. A., Díaz, M. F., Campillo, N. E. (2019). QSAR Classification Models for Predicting the Activity of Inhibitors of Beta-Secretase (BACE1) Associated with Alzheimer's disease. *Scientific Reports*, 9(1), 9102.

Ponzoni, I., Sebastián-Pérez, V., Requena-Triguero, C., Roca, C., Martínez, M. J., Cravero, F., Díaz, M. F., Páez, J. A., Gómez Arrayás, R., Adrio, J. & Campillo, N. E.

(2017). Hybridizing feature selection and feature learning approaches in QSAR modeling for drug discovery. *Scientific Reports*, 7(1), 2403.

Rester, U. (2008). From virtuality to reality-Virtual screening in lead discovery and lead optimization: a medicinal chemistry perspective. *Current opinion in drug discovery & development*, 11(4), 559-568.

Richon, A. B. (2008). An early history of the molecular modeling industry. *Drug discovery today*, 13(15-16), 659-664.

Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6), 386.

Rosi, P. (2010). *Introducción a la representación molecular*. "Colección Encuentro Inet" Ministerio de Educación de la Nación. Instituto Nacional de Educación Tecnológica, ed. Kirschenbaum J.M. Buenos Aires, Argentina

Russell, S. J., & Norvig, P. (2004). *Artificial intelligence: a modern approach*. Malaysia; Pearson Education S.A., 2nd Edition.

Sahoo, S., Adhikari, C., Kuanar, M., & K Mishra, B. (2016). A short review of the generation of molecular descriptors and their applications in quantitative structure property/activity relationships. *Current computer-aided drug design*, 12(3), 181-205.

Schustik, S. A.; Cravero, F.; Martínez, M. J.; Ponzoni, I., Díaz, M. F. (2019a) *Informática de Polímeros Aplicada a la Estimación de la Ductilidad en un Nuevo Material a Partir de la Relación de Propiedades Derivadas del Ensayo de Tensión*. 32° Congreso Argentino de Química, Buenos Aires, Argentina

Schustik, S. A., Cravero, F., Martínez, M. J., Ponzoni, I., Díaz, M. F. (2019b). *PolyMaS: A new tool for computational polymerization of structural repetitive units*. (Artículo actualmente en revisión en una revista indexada del área de Quimioinformática).

Schweizer, K. S., & Curro, J. G. (1994). *PRISM theory of the structure, thermodynamics, and phase transitions of polymer liquids and alloys*. In Atomistic Modeling of Physical Properties (pp. 319-377). Springer, Berlin, Heidelberg.

Seymour, R.B. & Carraher, C.E. (1998). *Introducción a la química de los polímetros*. Editorial Reverté, 3ra ed. Barcelona, España.

Soto, A. J., Cecchini, R. L., Vazquez, G. E., & Ponzoni, I. (2009a). Multi-Objective Feature Selection in QSAR Using a Machine Learning Approach. *Molecular Informatics*, 28(11-12), 1509-1523.

Soto A.J., Ponzoni I., Vazquez G.E. (2009b) Segregating Confident Predictions of Chemicals' Properties for Virtual Screening of Drugs. In: Omatu S. et al. (eds) Distributed Computing, Artificial Intelligence, Bioinformatics, Soft Computing, and

Ambient Assisted Living. IWANN 2009. *Lecture Notes in Computer Science*, vol 5518. Springer, Berlin, Heidelberg

Soto, A. J., Martínez, M. J., Cecchini, R. L., Vazquez, G. E. & Ponzoni, I. (2010). *DELPHOS: Computational Tool for Selection of Relevant Descriptor Subsets in ADMET Prediction*. First International Meeting of Pharmaceutical Sciences (RICiFA 2010), Córdoba, Argentina; page 79.

STATISTICA, Version 8.0, StatSoft Inc., Tulsa, USA, 2007.

Tabor, D., & Winterton, R. H. S. (1969). *The direct measurement of normal and retarded van der Waals forces*. Proceedings of the Royal Society of London. A. Mathematical and Physical Sciences, 312(1511), 435-450.

Timofeev, R. (2004). *Classification and regression trees (CART) theory and applications*. Humboldt University, Berlin.

Todeschini, R., & Consonni, V. (2008). *Handbook of molecular descriptors* (Vol. 11). John Wiley & Sons.

Todeschini, R., & Consonni, V. (2009). *Molecular descriptors for chemoinformatics: volume I: alphabetical listing/volume II: appendices, references* (Vol. 41). John Wiley & Sons.

Todeschini, R., Consonni, V., & Gramatica, P. (2009). *Chemometrics in QSAR.* In S. Brown, R. Tauler, & R. Walczak (a cura di), Comprehensive Chemometrics (pp. 129-172). Oxford: Elsevier.

Toropova, A. P., Toropov, A. A., Kudyshkin, V. O., Leszczynska, D., Leszczynski, J. (2014). Optimal descriptors as a tool to predict the thermal decomposition of polymers. *Journal of Mathematical Chemistry*, 52(5), 1171-1181.

Tresp, V. (2001). *Committee machines, book chapter in: Handbook for Neural Network Signal Processing*, Yu Hen Hu and Jenq-Neng Hwang (eds.), CRC Press.

Ulmer II, C. W., Smith, D. A., Sumpter, B. G., & Noid, D. I. (1998). Computational neural networks and the rational design of polymeric materials: the next generation polycarbonates. *Computational and theoretical polymer science*, 8(3-4), 311-321.

Van Krevelen D.W., *Properties of Polymers* (2009), fourth ed. Elsevier, Amsterdam, The Netherlands.

Van Krevelen, D. W., & Te Nijenhuis, K. (2009). *Properties of polymers: their correlation with chemical structure; their numerical estimation and prediction from additive group contributions*. Elsevier.

Voulodimos, A., Doulamis, N., Doulamis, A., & Protopapadakis, E. (2018). Deep learning for computer vision: A brief review. *Computational intelligence and neuroscience*.

Walters, W. P., Stahl, M. T., & Murcko, M. A. (1998). Virtual screening-an overview. *Drug discovery today*, 3(4), 160-178.

Wang, B. (2006). *Computer applications in pharmaceutical research and development*. S. Ekins (Ed.). Wiley-Interscience.

Ward M. and Sweeney J. (2004) *Yielding and Instability in Polymers*. In: M. Ward and J. Sweeney (ed) An Introduction to the Mechanical Properties of Solid Polymers, 2nd ed. John Wiley & Sons Ltd, England, 241-272 pp.

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.

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

Wong, P. C., & Thomas, J. (2004). Visual analytics. *IEEE Computer Graphics and Applications*, 24(5), 20-21.

Wu, K., Sukumar, N., Lanzillo, N. A., Wang, C., Ramprasad, R., Ma, R., Baldwin, A. F., Sotzing, G., Breneman, C. (2016). Prediction of polymer properties using infinite chain descriptors (ICD) and machine learning: Toward optimized dielectric polymeric materials. *Journal of Polymer Science Part B: Polymer Physics*, 54(20), 2082-2091.

Xue, L., & Bajorath, J. (2000). Molecular descriptors in chemoinformatics, computational combinatorial chemistry, and virtual screening. *Combinatorial chemistry & high throughput screening*, 3(5), 363-372.

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

Yousefinejad, S., & Hemmateenejad, B. (2015). Chemometrics tools in QSAR/QSPR studies: A historical perspective. *Chemometrics and Intelligent Laboratory Systems*, 149, 177-204.

Yu, X.L., Yi, B., & Wang, X.Y. (2008). Prediction of the glass transition temperatures for polymers with artificial neural network. *Journal of Theoretical and Computational Chemistry*, 7(05), 953-963.

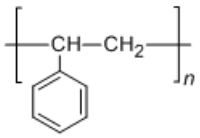
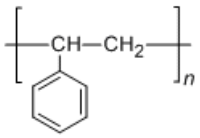
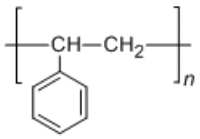
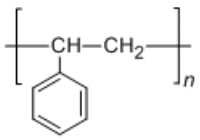
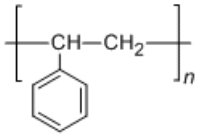
Zakharov, A. V., Peach, M. L., Sitzmann, M., & Nicklaus, M. C. (2014). QSAR modeling of imbalanced high-throughput screening data in PubChem. *Journal of chemical information and modeling*, 54(3), 705-712.

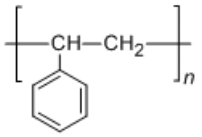
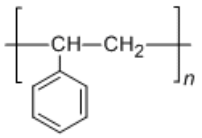
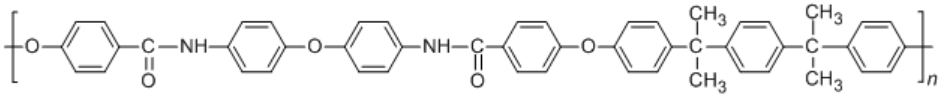
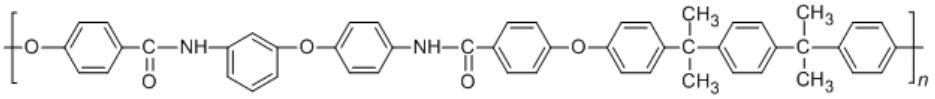
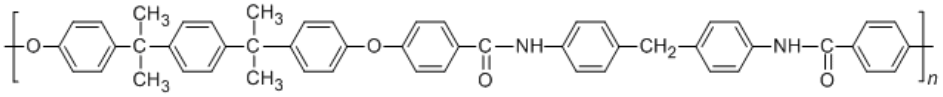
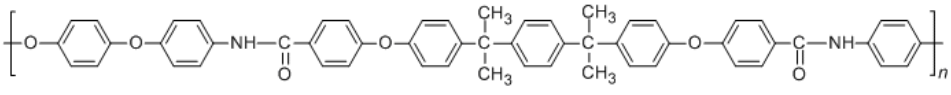
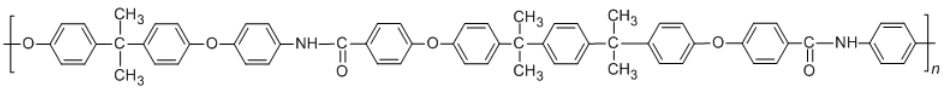
Zhao, Y. H., Le, J., Abraham, M. H., Hersey, A., Eddershaw, P. J., Luscombe, C. N., ... & Platts, J. A. (2001). Evaluation of human intestinal absorption data and subsequent derivation of a quantitative structure–activity relationship (QSAR) with the Abraham descriptors. *Journal of pharmaceutical sciences*, 90(6), 749-784.

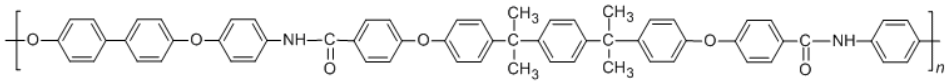
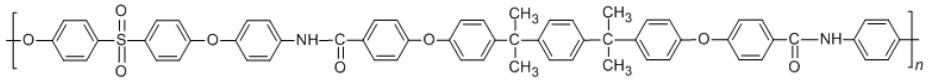
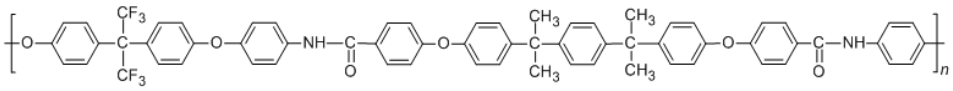
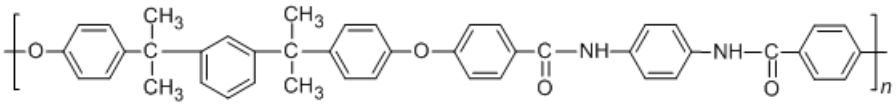
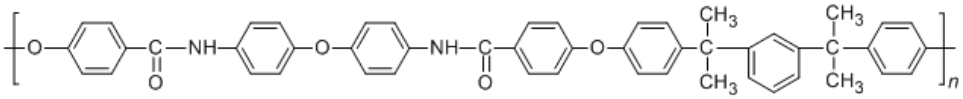
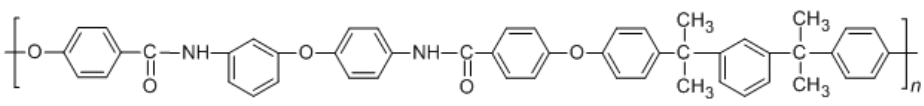
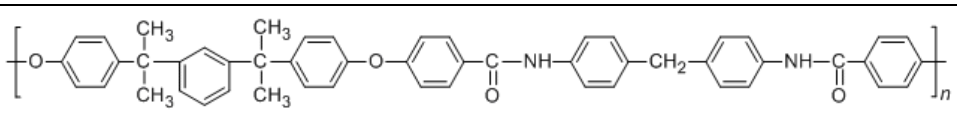
ANEXOS

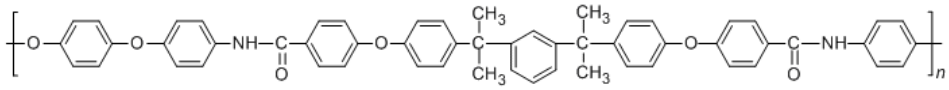
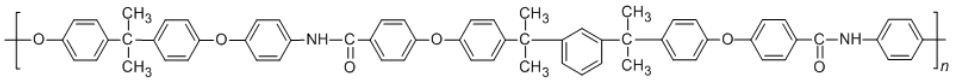
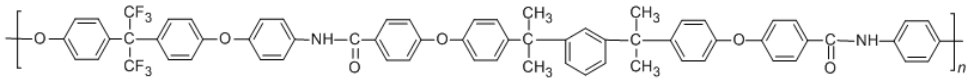
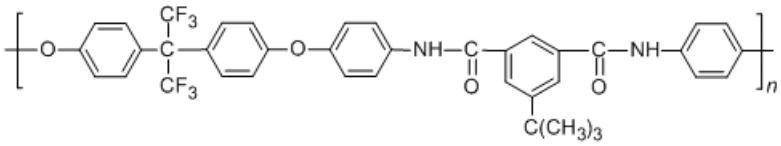
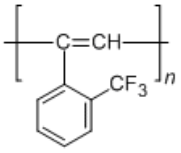
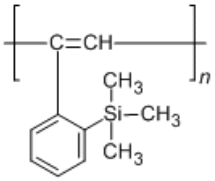
CAPÍTULO 4: INFORMÁTICA DE POLÍMEROS

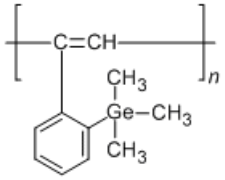
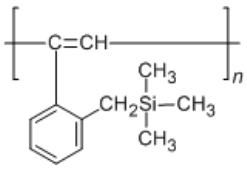
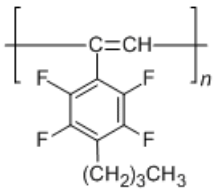
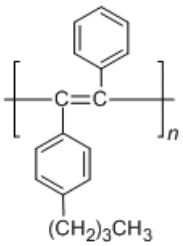
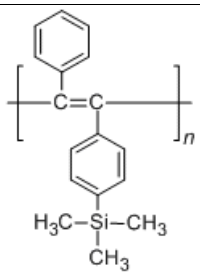
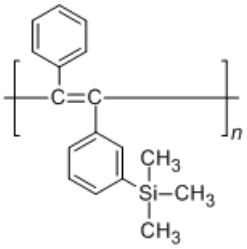
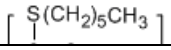
TABLA A4.1. INFORMACIÓN SOBRE NOMBRES DE LOS 77 POLÍMEROS SEGÚN DATOS EXTRAÍDOS DE POLYINFO, JUNTO CON EL IDENTIFICADOR NUMÉRICO ÚNICO (ID) QUE LOS IDENTIFICA EN LA BASE DE DATOS INTERNA UTILIZADA, SUS VALORES DE MN (PESO PROMEDIO EN NÚMERO), MW (PESO PROMEDIO EN PESO) Y CHS (CROSS-HEAD SPEED, VELOCIDAD DEL ENSAYO DE TENSIÓN) Y LAS IMÁGENES DE ESTRUCTURA MOLECULAR 2D CORRESPONDIENTES A LA URE DE CADA POLÍMERO.

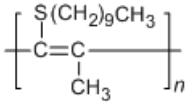
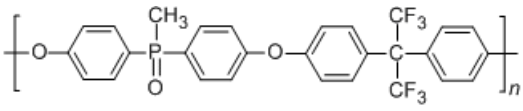
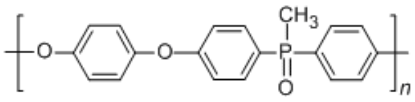
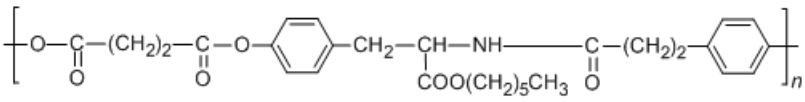
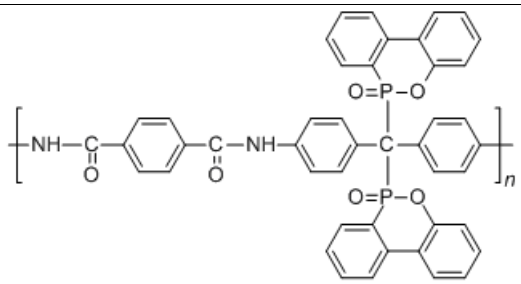
ID: 1	Polystyrene
Mn: 123000 Mw: 258000 CHS: 5	
ID: 2	Polystyrene
Mn: 146000 Mw: 168000 CHS: 1.27	
ID: 3	Polystyrene
Mn: 765000 Mw: 880000 CHS: 1.27	
ID: 4	Polystyrene
Mn: 231000 Mw: 266000 CHS: 1.27	
ID: 5	Polystyrene
Mn: 13500 Mw: 27000 CHS: 5	
ID: 6	Polystyrene
Mn:	

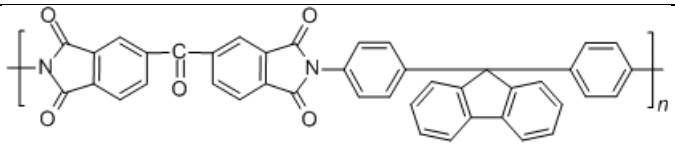
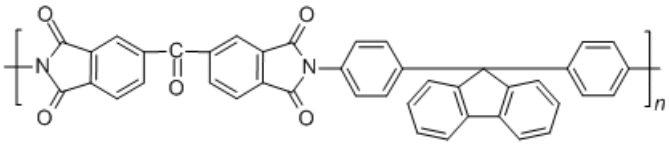
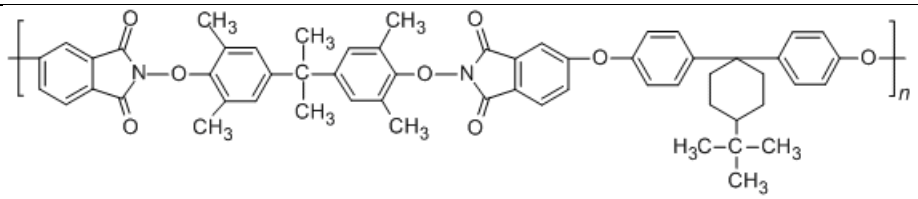
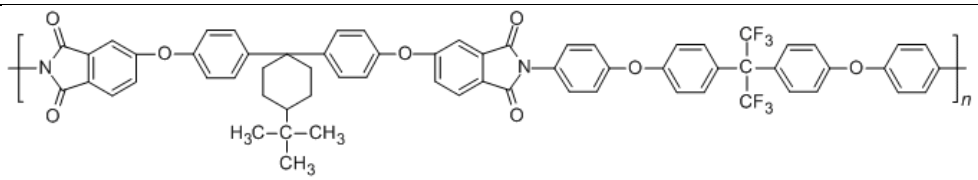
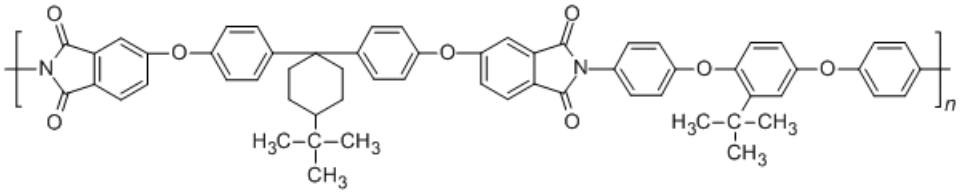
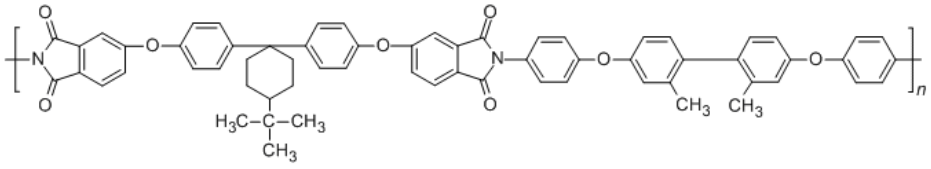
13500 Mw: 27000 CHS: 5	
ID: 7	Polystyrene
Mn: 13500 Mw: 27000 CHS: 5	
ID: 8	Poly((4,4'-oxydianiline)-alt-{alpha,alpha'-bis[4-(4-carboxyphenoxy)phenyl]-1,4-diisopropylbenzene})
Mn: 41000 Mw: 79000 CHS: 50	
ID: 9	Poly((3,4'-oxydianiline)-alt-{alpha,alpha'-bis[4-(4-carboxyphenoxy)phenyl]-1,4-diisopropylbenzene})
Mn: 33000 Mw: 51000 CHS: 50	
ID: 10	Poly((4,4'-methylenedianiline)-alt-{alpha,alpha'-bis[4-(4-carboxyphenoxy)phenyl]-1,4-diisopropylbenzene})
Mn: 42000 Mw: 67000 CHS: 50	
ID: 11	Poly([4,4'-(1,4-phenylenedioxy)dianiline]-alt-{alpha,alpha'-bis[4-(4-carboxyphenoxy)phenyl]-1,4-diisopropylbenzene})
Mn: 44000 Mw: 69000 CHS: 50	
ID: 12	Poly({4,4'-[1-methylethane-1,1-diylbis(4,1-phenyleneoxy)]dianiline}-alt-{alpha,alpha'-bis[4-(4-carboxyphenoxy)phenyl]-1,4-diisopropylbenzene})
Mn: 41000 Mw: 71000 CHS: 50	
ID: 13	Poly([4,4'-(biphenyl-4,4'-diyldioxy)dianiline]-alt-{alpha,alpha'-bis[4-(4-

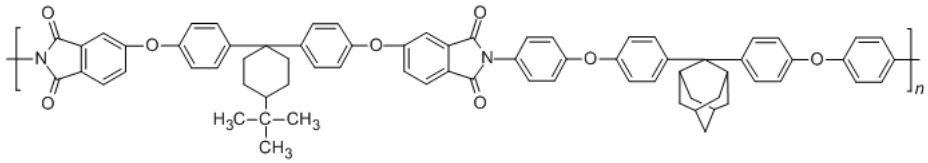
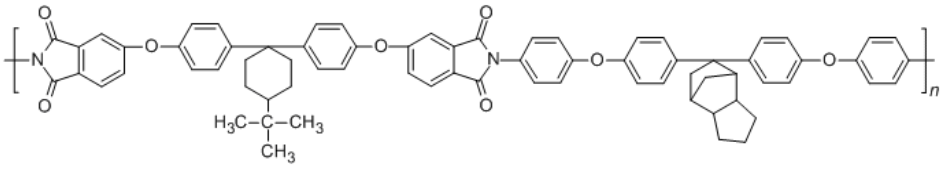
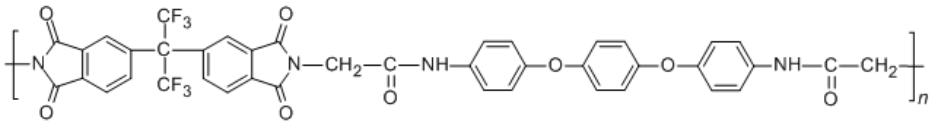
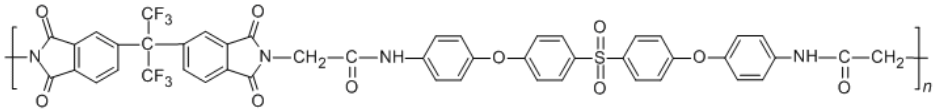
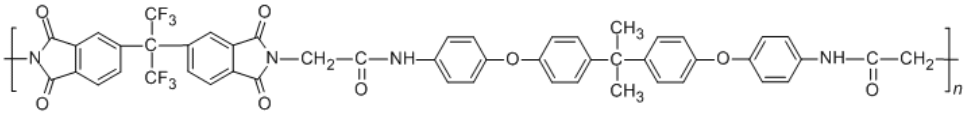
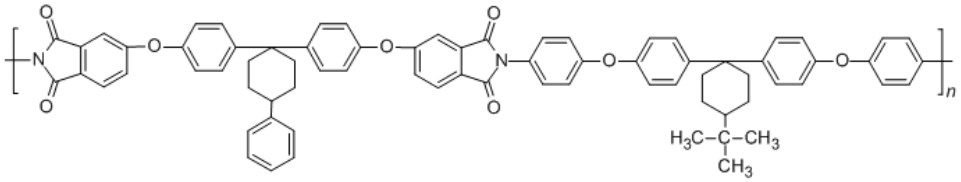
	carboxyphenoxy)phenyl]-1,4-diisopropylbenzene})
Mn: 44000 Mw: 74000 CHS: 50	
ID: 14	Poly({4,4'-[sulfonylbis(4,1-phenyleneoxy)]dianiline}-alt-{alpha,alpha'-bis[4-(4-carboxyphenoxy)phenyl]-1,4-diisopropylbenzene})
Mn: 41000 Mw: 73000 CHS: 50	
ID: 15	Poly({4,4'-[1-(trifluoromethyl)-2,2,2-trifluoroethane-1,1-diylbis(4,1-phenyleneoxy)]dianiline}-alt-{alpha,alpha'-bis[4-(4-carboxyphenoxy)phenyl]-1,4-diisopropylbenzene})
Mn: 38000 Mw: 60000 CHS: 50	
ID: 16	Poly((m-phenylenediamine)-alt-{alpha,alpha'-bis[4-(4-carboxyphenoxy)phenyl]-1,3-diisopropylbenzene})
Mn: 65000 Mw: 89000 CHS: 50	
ID: 17	Poly((4,4'-oxydianiline)-alt-{alpha,alpha'-bis[4-(4-carboxyphenoxy)phenyl]-1,3-diisopropylbenzene})
Mn: 55000 Mw: 92000 CHS: 50	
ID: 18	Poly((3,4'-oxydianiline)-alt-{alpha,alpha'-bis[4-(4-carboxyphenoxy)phenyl]-1,3-diisopropylbenzene})
Mn: 64000 Mw: 163000 CHS: 50	
ID: 19	Poly((4,4'-methylenedianiline)-alt-{alpha,alpha'-bis[4-(4-carboxyphenoxy)phenyl]-1,3-diisopropylbenzene})
Mn: 58000 Mw:	

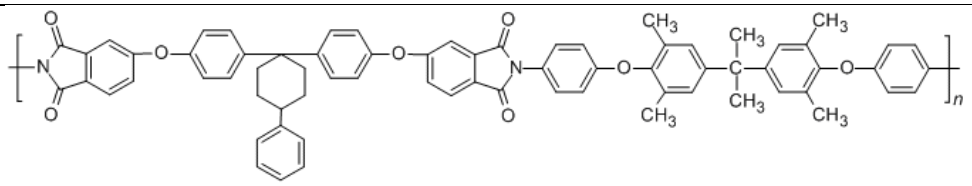
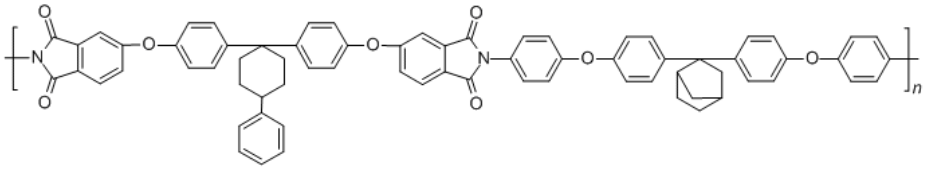
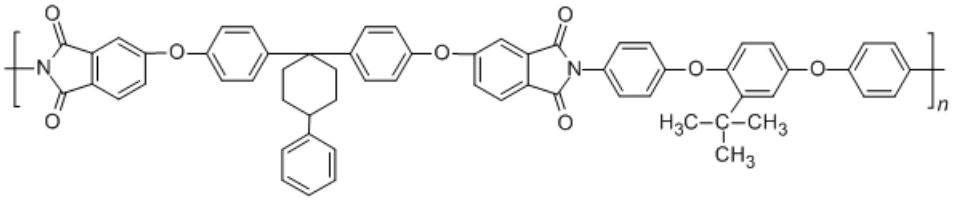
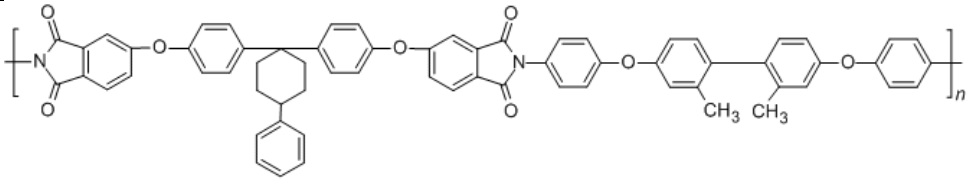
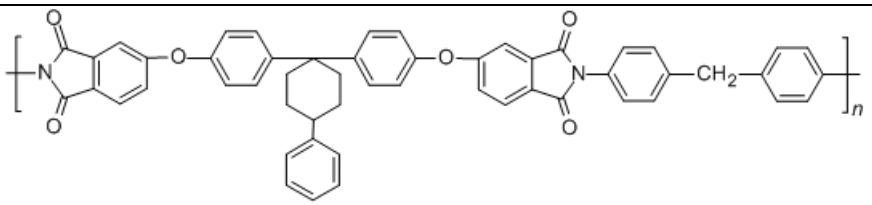
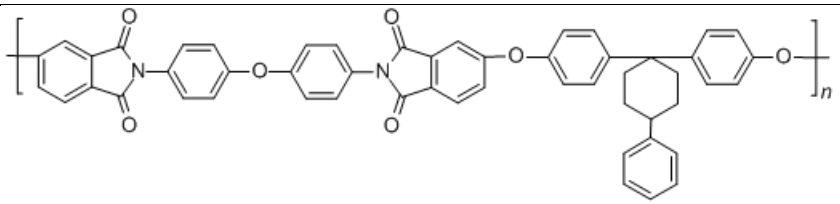
107000 CHS: 50	
ID: 20	Poly([4,4'-(1,4-phenylenedioxy)dianiline]-alt-{alpha,alpha'-bis[4-(4-carboxyphenoxy)phenyl]-1,3-diisopropylbenzene})
Mn: 61000 Mw: 120000 CHS: 50	
ID: 21	Poly({4,4'-[1-methylethane-1,1-diylbis(4,1-phenyleneoxy)]dianiline}-alt-{alpha,alpha'-bis[4-(4-carboxyphenoxy)phenyl]-1,3-diisopropylbenzene})
Mn: 71000 Mw: 122000 CHS: 50	
ID: 22	Poly({4,4'-[1-(trifluoromethyl)-2,2,2-trifluoroethane-1,1-diylbis(4,1-phenyleneoxy)]dianiline}-alt-{alpha,alpha'-bis[4-(4-carboxyphenoxy)phenyl]-1,3-diisopropylbenzene})
Mn: 81000 Mw: 126000 CHS: 50	
ID: 23	Poly({4,4'-[1-(trifluoromethyl)-2,2,2-trifluoroethane-1,1-diylbis(4,1-phenyleneoxy)]dianiline}-alt-(5-tert-butylisophthalic acid))
Mn: 69000 Mw: 202000 CHS: 5	
ID: 24	Poly{[2-(trifluoromethyl)phenyl]acetylene}
Mn: 190000 Mw: 690000 CHS: 30.1	
ID: 25	Poly{[o-(trimethylsilyl)phenyl]acetylene}
Mn: 500000 Mw: 1900000 CHS: 30.1	
ID: 26	Poly[(2-ethynylphenyl)(trimethyl)germane]

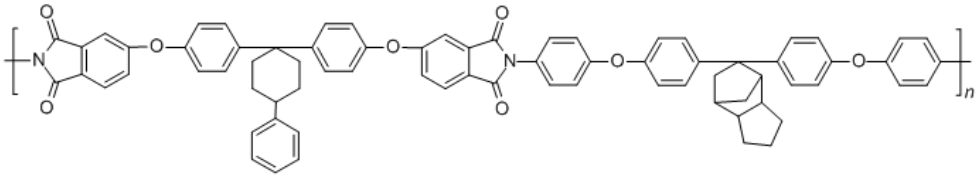
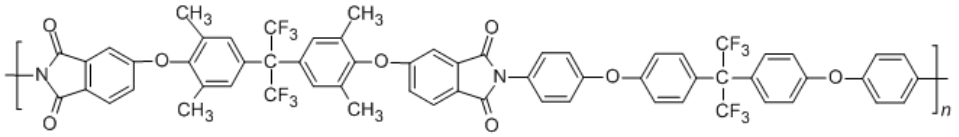
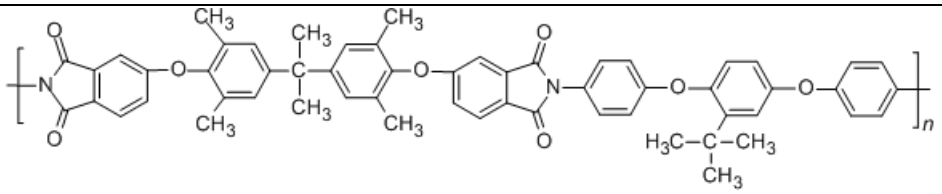
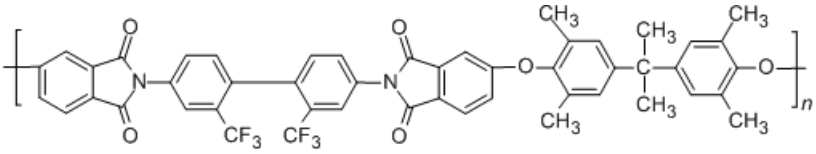
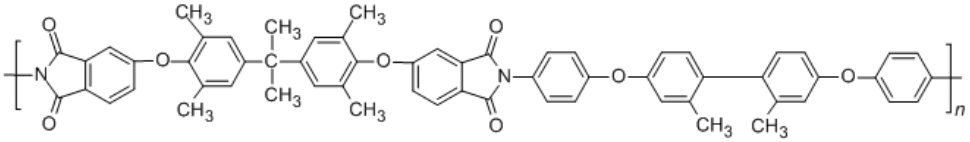
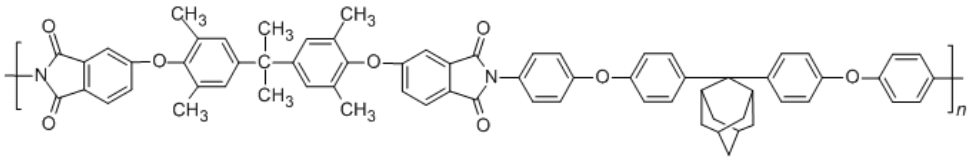
Mn: 190000 Mw: 690000 CHS: 30.1	
ID: 27	Poly(1-{2-[(trimethylsilyl)methyl]phenyl}ethene-1,2-diyl)
Mn: 450000 Mw: 700000 CHS: 30.1	
ID: 28	Poly[(4-butyl-2,3,5,6-tetrafluorophenyl)acetylene]
Mn: 470000 Mw: 850000 CHS: 30.1	
ID: 29	Poly[1-(4-butylphenyl)-2-phenylacetylene]
Mn: 460000 Mw: 1300000 CHS: 30.1	
ID: 30	Poly{1-phenyl-2-[4-(trimethylsilyl)phenyl]acetylene}
Mn: 750000 Mw: 2200000 CHS: 30.1	
ID: 31	Poly{1-phenyl-2-[3-(trimethylsilyl)phenyl]acetylene}
Mn: 250000 Mw: 1400000 CHS: 30.1	
ID: 32	Poly[1-(hexylsulfanyl)prop-1-yne]
Mn:	

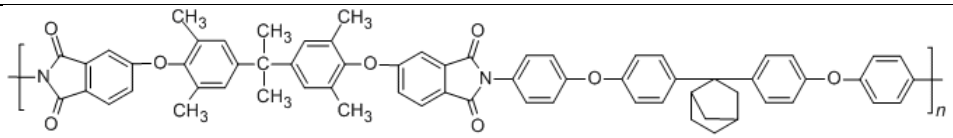
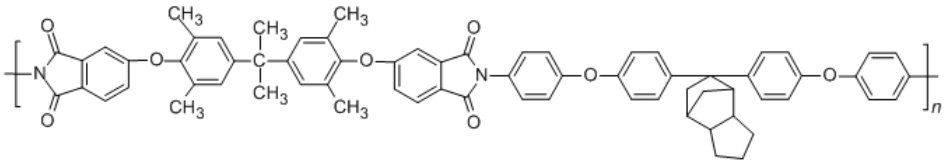
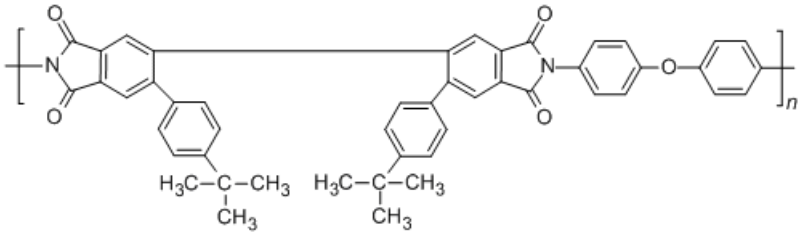
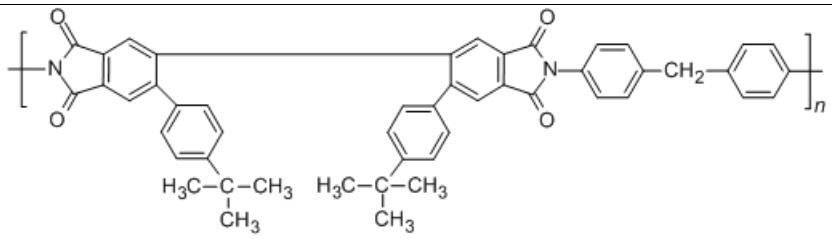
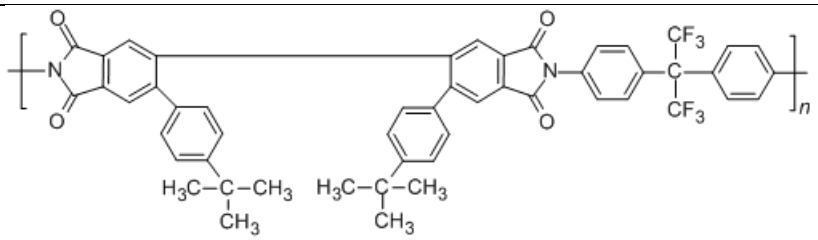
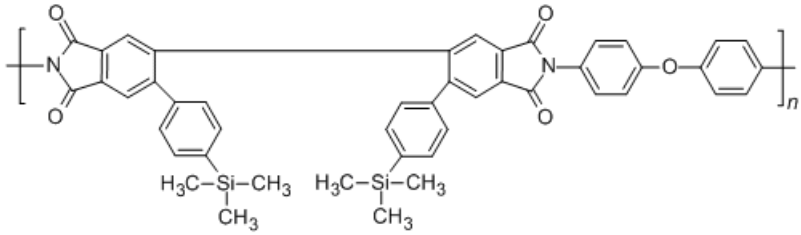
34000 Mw: 110000 CHS: 30.1	
ID: 33	Poly[1-(decylsulfanyl)prop-1-yne]
Mn: 39000 Mw: 170000 CHS: 30.1	
ID: 34	Poly{(1,1,1,3,3,3-hexafluoro-2,2-diphenyl-propane)-alt-[bis(4-fluorophenyl)methylphosphine oxide]}
Mn: 21000 Mw: 33000 CHS: 12.7	
ID: 35	Poly{hydroquinone-alt-[bis(4-fluorophenyl)methylphosphine oxide]}
Mn: 44000 Mw: 72000 CHS: 12.7	
ID: 36	Poly[(desaminotyrosyl-L-tyrosine hexyl ester)-alt-(succinic acid)]
Mn: 68000 Mw: 102000 CHS: 100	
ID: 37	Poly({4,4'-[bis(2-oxodibenzo[c,e][1,2]oxaphoshinin-2-yl)methylene]dianiline}-alt-(terephthalic acid))
Mn: 53000 Mw: 115000 CHS: 50	
ID: 38	Poly{[4,4'-(9H-fluorene-9,9-diyl)dianiline]-alt-[5,5'-carbonylbis(isobenzofuran-1,3-dione)]}
Mn: 11000	

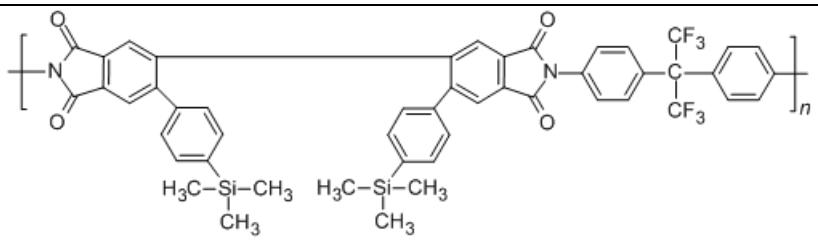
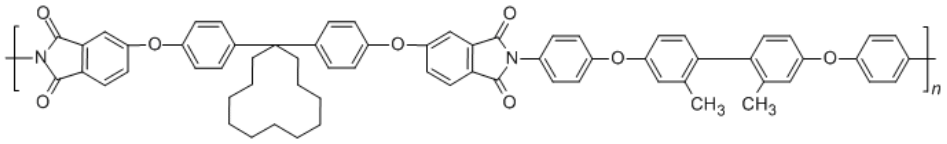
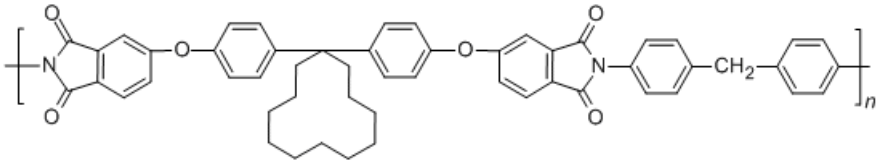
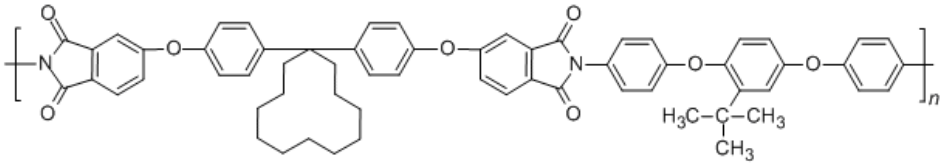
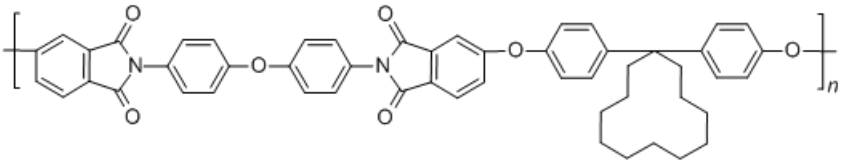
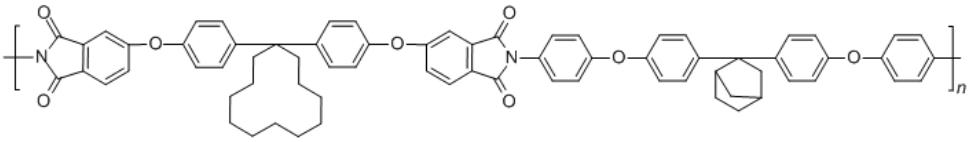
Mw: 26500 CHS: 1	
ID: 39	Poly{[4,4'-(9H-fluorene-9,9-diyl)dianiline]-alt-[5,5'-carbonylbis(isobenzofuran-1,3-dione)]}
Mn: 9500 Mw: 22500 CHS: 1	
ID: 40	Poly({O,O'-[1-methylethane-1,1-diylbis(2,6-dimethyl-4,1-phenylene)]dihydroxylamine}-alt-{5,5'-[4-tert-butylcyclohexane-1,1-diylbis(4,1-phenyleneoxy)]diisobenzofuran-1,3-dione})
Mn: 64000 Mw: 142000 CHS: 20	
ID: 41	Poly({4,4'-[1-(trifluoromethyl)-2,2,2-trifluoroethane-1,1-diylbis(4,1-phenyleneoxy)]dianiline}-alt-{5,5'-[4-tert-butylcyclohexane-1,1-diylbis(4,1-phenyleneoxy)]diisobenzofuran-1,3-dione})
Mn: 42000 Mw: 79000 CHS: 20	
ID: 42	Poly([4,4'-(2-tert-butyl-1,4-phenylenedioxy)dianiline]-alt-{5,5'-[4-tert-butylcyclohexane-1,1-diylbis(4,1-phenyleneoxy)]diisobenzofuran-1,3-dione})
Mn: 37000 Mw: 85000 CHS: 20	
ID: 43	Poly([4,4'-(2,2'-dimethylbiphenyl-4,4'-diyldioxy)dianiline]-alt-{5,5'-[4-tert-butylcyclohexane-1,1-diylbis(4,1-phenyleneoxy)]diisobenzofuran-1,3-dione})
Mn: 43000 Mw: 76000 CHS: 20	
ID: 44	Poly({4,4'-[adamantane-2,2-diylbis(4,1-phenyleneoxy)]dianiline}-alt-{5,5'-[4-tert-butylcyclohexane-1,1-diylbis(4,1-phenyleneoxy)]diisobenzofuran-1,3-dione})
Mn: 31000 Mw: 60000	

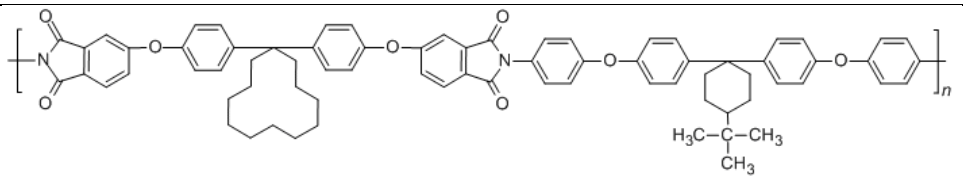
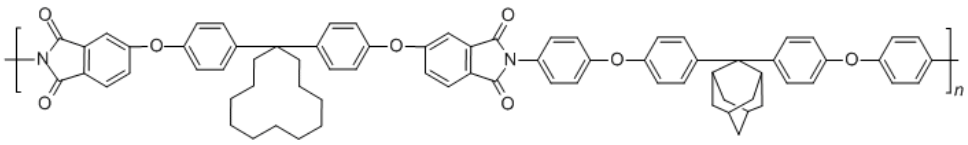
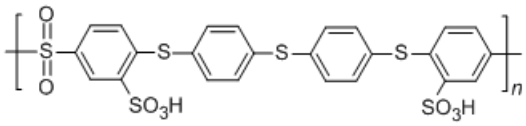
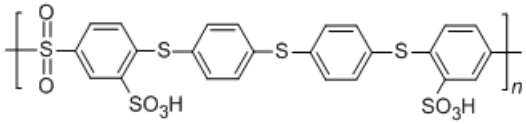
CHS: 20	
ID: 45	Poly({4,4'-[tricyclo[5.2.1.0^2,6^^]decane-8,8-diylbis(4,1-phenyleneoxy)]dianiline}-alt-{5,5'-[4-tert-butylcyclohexane-1,1-diylbis(4,1-phenyleneoxy)]diisobenzofuran-1,3-dione})
Mn: 50000 Mw: 96000 CHS: 20	
ID: 46	Poly([4,4'-(1,4-phenylenedioxy)dianiline]-alt-{N,N'-bis[(chloroformyl)methyl]-4,4'-[2,2,2-trifluoro-1-(trifluoromethyl)ethane-1,1-diyl]diphthalimide})
Mn: 8200 Mw: 19500 CHS: 9	
ID: 47	Poly({4,4'-[sulfonylbis(4,1-phenyleneoxy)]dianiline}-alt-{N,N'-bis[(chloroformyl)methyl]-4,4'-[2,2,2-trifluoro-1-(trifluoromethyl)ethane-1,1-diyl]diphthalimide})
Mn: 6300 Mw: 26500 CHS: 9	
ID: 48	Poly({4,4'-[1-methylethane-1,1-diylbis(4,1-phenyleneoxy)]diphenol}-alt-{N,N'-bis[(chloroformyl)methyl]-4,4'-[2,2,2-trifluoro-1-(trifluoromethyl)ethane-1,1-diyl]diphthalimide})
Mn: 4700 Mw: 24500 CHS: 9	
ID: 49	Poly({4,4'-[4-(tert-butyl)cyclohexane-1,1-diylbis(4,1-phenyleneoxy)]dianiline}-alt-{5,5'-[4-phenylcyclohexane-1,1-diylbis(4,1-phenyleneoxy)]diisobenzofuran-1,3-dione})
Mn: 30000 Mw: 65000 CHS: 20	
ID: 50	Poly({4,4'-[1-methylethane-1,1-diylbis(2,6-dimethyl-4,1-phenyleneoxy)]dianiline}-alt-{5,5'-[4-phenylcyclohexane-1,1-diylbis(4,1-phenyleneoxy)]diisobenzofuran-1,3-dione})
Mn: 17000	

Mw: 41000 CHS: 20	
ID: 51	Poly({4,4'-[(bicyclo[2.2.1]heptane-2,2-diyl)bis(4,1-phenyleneoxy)]dianiline}-alt-{5,5'-[4-phenylcyclohexane-1,1-diylbis(4,1-phenyleneoxy)]diisobenzofuran-1,3-dione})
Mn: 23000 Mw: 51000 CHS: 20	
ID: 52	Poly([4,4'-(2-tert-butyl-1,4-phenylenedioxy)dianiline]-alt-{5,5'-[4-phenylcyclohexane-1,1-diylbis(4,1-phenyleneoxy)]diisobenzofuran-1,3-dione})
Mn: 57000 Mw: 130000 CHS: 20	
ID: 53	Poly([4,4'-(2,2'-dimethylbiphenyl-4,4'-diyldioxy)dianiline]-alt-{5,5'-[4-phenylcyclohexane-1,1-diylbis(4,1-phenyleneoxy)]diisobenzofuran-1,3-dione})
Mn: Mw: CHS:	
ID: 54	Poly((4,4'-methylenedianiline)-alt-{5,5'-[4-phenylcyclohexane-1,1-diylbis(4,1-phenyleneoxy)]diisobenzofuran-1,3-dione})
Mn: 25000 Mw: 57000 CHS: 20	
ID: 55	Poly((4,4'-oxydianiline)-alt-{5,5'-[4-phenylcyclohexane-1,1-diylbis(4,1-phenyleneoxy)]diisobenzofuran-1,3-dione})
Mn: 28000 Mw: 56000 CHS: 20	
ID: 56	Poly({4,4'-[tricyclo[5.2.1.0^2,6^^]decane-8,8-diylbis(4,1-phenyleneoxy)]dianiline}-alt-{5,5'-[4-phenylcyclohexane-1,1-diylbis(4,1-phenyleneoxy)]diisobenzofuran-1,3-dione})
Mn: 36000 Mw: 71000 CHS: 20	

	
ID: 57	Poly({4,4'-[1-(trifluoromethyl)-2,2,2-trifluoroethane-1,1-diylbis(4,1-phenyleneoxy)]dianiline}-alt-{5,5'-[1-methylethane-1,1-diylbis(2,6-dimethyl-4,1-phenyleneoxy)]bis(isobenzofuran-1,3-dione)})
Mn: 50000 Mw: 108000 CHS: 20	
ID: 58	Poly([4,4'-(2-tert-butyl-1,4-phenylenedioxy)dianiline]-alt-{5,5'-[1-methylethane-1,1-diylbis(2,6-dimethyl-4,1-phenyleneoxy)]bis(isobenzofuran-1,3-dione)})
Mn: 33000 Mw: 61000 CHS: 20	
ID: 59	Poly([2,2'-bis(trifluoromethyl)benzidine]-alt-{5,5'-[1-methylethane-1,1-diylbis(2,6-dimethyl-4,1-phenyleneoxy)]bis(isobenzofuran-1,3-dione)})
Mn: 27000 Mw: 51000 CHS: 20	
ID: 60	Poly([4,4'-(2,2'-dimethylbiphenyl-4,4'-diyldioxy)dianiline]-alt-{5,5'-[1-methylethane-1,1-diylbis(2,6-dimethyl-4,1-phenyleneoxy)]bis(isobenzofuran-1,3-dione)})
Mn: 46000 Mw: 94000 CHS: 20	
ID: 61	Poly({4,4'-[adamantane-2,2-diylbis(4,1-phenyleneoxy)]dianiline}-alt-{5,5'-[1-methylethane-1,1-diylbis(2,6-dimethyl-4,1-phenyleneoxy)]bis(isobenzofuran-1,3-dione)})
Mn: 39000 Mw: 73000 CHS: 20	
ID: 62	Poly({4,4'-[(bicyclo[2.2.1]heptane-2,2-diyl)bis(4,1-phenyleneoxy)]dianiline}-alt-{5,5'-[1-methylethane-1,1-diylbis(2,6-dimethyl-4,1-phenyleneoxy)]bis(isobenzofuran-1,3-dione)})
Mn: 54000 Mw:	

124000 CHS: 20	
ID: 63	Poly({4,4'-[tricyclo[5.2.1.0^2,6^^]decane-8,8-diylbis(4,1-phenyleneoxy)]dianiline}-alt-{5,5'-[1-methylethane-1,1-diylbis(2,6-dimethyl-4,1-phenyleneoxy)]bis(isobenzofuran-1,3-dione)})
Mn: 41000 Mw: 72000 CHS: 20	
ID: 64	Poly{(4,4'-oxydianiline)-alt-[6,6'-bis(4-tertbutylphenyl)biphenyl-3,3',4,4'-tetracarboxylic anhydride]}
Mn: 154000 Mw: 332500 CHS: 5	
ID: 65	Poly{(4,4'-methylenedianiline)-alt-[6,6'-bis(4-tertbutylphenyl)biphenyl-3,3',4,4'-tetracarboxylic anhydride]}
Mn: 61000 Mw: 141500 CHS: 5	
ID: 66	Poly({4,4'-[2,2,2-trifluoro-1-(trifluoromethyl)ethane-1,1-diyl]dianiline}-alt-[6,6'-bis(4-tertbutylphenyl)biphenyl-3,3',4,4'-tetracarboxylic anhydride])
Mn: 74000 Mw: 138000 CHS: 5	
ID: 67	Poly((4,4'-oxydianiline)-alt-{6,6'-bis[4-(trimethylsilyl)phenyl]biphenyl-3,3',4,4'-tetracarboxylic anhydride})
Mn: 31500 Mw: 110500 CHS: 5	
ID: 68	Poly({4,4'-[2,2,2-trifluoro-1-(trifluoromethyl)ethane-1,1-diyl]dianiline}-alt-{6,6'-bis[4-(trimethylsilyl)phenyl]biphenyl-3,3',4,4'-tetracarboxylic anhydride})
Mn: 31500	

Mw: 60500 CHS: 5	
ID: 69	Poly([4,4'-(2,2'-dimethylbiphenyl-4,4'-diyldioxy)dianiline]-alt-{5,5'-[cyclododecane-1,1-diylbis(4,1-phenylene)]bis(isobenzofuran-1,3-dione)})
Mn: 16000 Mw: 32000 CHS: 20	
ID: 70	Poly((4,4'-methylenedianiline)-alt-{5,5'-[cyclododecane-1,1-diylbis(4,1-phenylene)]bis(isobenzofuran-1,3-dione)})
Mn: 36000 Mw: 62000 CHS: 20	
ID: 71	Poly([4,4'-(2-tert-butyl-1,4-phenylenedioxy)dianiline]-alt-{5,5'-[cyclododecane-1,1-diylbis(4,1-phenylene)]bis(isobenzofuran-1,3-dione)})
Mn: 33000 Mw: 78000 CHS: 20	
ID: 72	Poly((4,4'-oxydianiline)-alt-{5,5'-[cyclododecane-1,1-diylbis(4,1-phenylene)]bis(isobenzofuran-1,3-dione)})
Mn: 63000 Mw: 115000 CHS: 20	
ID: 73	Poly({4,4'-[(bicyclo[2.2.1]heptane-2,2-diyl)bis(4,1-phenyleneoxy)]dianiline}-alt-{5,5'-[cyclododecane-1,1-diylbis(4,1-phenylene)]bis(isobenzofuran-1,3-dione)})
Mn: 51000 Mw: 94000 CHS: 20	
ID: 74	Poly({4,4'-[4-(tert-butyl)cyclohexane-1,1-diylbis(4,1-phenyleneoxy)]dianiline}-alt-{5,5'-[cyclododecane-1,1-diylbis(4,1-phenylene)]bis(isobenzofuran-1,3-dione)})
Mn: 36000 Mw: 81000	

CHS: 20	
ID: 75	Poly({4,4'-[adamantane-2,2-diylbis(4,1-phenyleneoxy)]dianiline}-alt-{5,5'-[cyclododecane-1,1-diylbis(4,1-phenylene)]bis(isobenzofuran-1,3-dione)})
Mn: 26000 Mw: 44000 CHS: 20	
ID: 76	Poly[sulfonyl(3-sulfo-1,4-phenylene)sulfanediyl-1,4-phenylenesulfanediyl-1,4-phenylenesulfanediyl(2-sulfo-1,4-phenylene)]
Mn: 41000 Mw: 94000 CHS: 5	
ID: 77	Poly[sulfonyl(3-sulfo-1,4-phenylene)sulfanediyl-1,4-phenylenesulfanediyl-1,4-phenylenesulfanediyl(2-sulfo-1,4-phenylene)]
Mn: 43000 Mw: 158000 CHS: 5	

CAPÍTULO 5: MODELADO QSPR CON DMS UNIVALUADOS

TABLA A.5. 1. MÉTRICAS ESTADÍSTICAS PARA EVALUAR EL COMPORTAMIENTO DE LAS CARACTERÍSTICAS Y LOS MODELOS PARA MÓDULO DE TENSIÓN: PREGUNTAS a Y b.

MÓDULO DE TENSIÓN																
	PRENGUNTA a								**PREGUNTA b**							
	UR vs Mn				URE vs Mw				FS:SRU - Mn vs Mn				FS:URE - Mw vs Mw			
	CFS															
	LR	NN	RF	RC	LR	NN	RF	RC	LR	NN	RF	RC	LR	NN	RF	RC
R²	0.37	-0.18	0.74	0.56	0.30	-0.18	0.74	0.56	0.36	0.96	0.97	0.98	0.30	0.55	0.96	0.92
MAE	13.17	4.78	0.49	0.46	30.65	4.78	0.49	0.46	0.31	0.22	0.16	0.14	0.48	0.48	0.22	0.26
RMSE	25.60	4.87	0.70	0.75	63.90	4.87	0.70	0.75	0.64	0.28	0.21	0.16	0.84	0.73	0.31	0.33
	W-LR															
	LR	NN	RF	RC	LR	NN	RF	RC	LR	NN	RF	RC	LR	NN	RF	RC
R²	0.46	0.39	0.96	0.92	0.36	0.39	0.96	0.92	0.36	0.90	0.89	0.93	0.34	0.86	0.84	0.90
MAE	62.02	1.83	0.86	1.08	135.16	1.83	0.86	1.08	0.41	0.30	0.24	0.19	0.81	0.83	0.27	0.25
RMSE	80.09	2.19	1.01	1.21	189.95	2.19	1.01	1.21	0.68	0.38	0.32	0.27	1.18	1.32	0.51	0.40
	W-NN															
	LR	NN	RF	RC	LR	NN	RF	RC	LR	NN	RF	RC	LR	NN	RF	RC
R²	0.12	0.39	0.43	0.63	0.14	0.39	0.43	0.63	0.32	0.76	0.97	0.98	0.29	0.24	0.93	0.95
MAE	39.42	3.53	0.55	0.49	81.92	3.53	0.55	0.49	0.33	0.30	0.17	0.13	0.58	1.28	0.27	0.23
RMSE	47.28	3.86	0.82	0.74	95.23	3.87	0.82	0.74	0.66	0.45	0.22	0.16	0.95	2.80	0.47	0.37
	W-RF															
	LR	NN	RF	RC	LR	NN	RF	RC	LR	NN	RF	RC	LR	NN	RF	RC
R²	0.44	0.74	0.50	0.06	0.34	0.74	0.50	0.06	0.00	0.50	0.85	0.93	0.00	0.39	0.80	0.86
MAE	62.79	3.97	1.38	1.47	133.64	3.97	1.38	1.47	0.36	0.62	0.27	0.20	0.36	1.11	0.38	0.28
RMSE	72.78	4.07	1.50	1.64	164.35	4.07	1.50	1.64	0.70	0.86	0.35	0.25	0.70	1.64	0.53	0.36
	W-RC															
	LR	NN	RF	RC	LR	NN	RF	RC	LR	NN	RF	RC	LR	NN	RF	RC
R²	0.37	0.00	0.95	0.95	0.29	0.00	0.95	0.95	0.41	0.39	0.82	0.82	0.34	0.57	0.85	0.86
MAE	69.31	0.47	1.24	1.06	148.62	0.47	1.24	1.06	0.44	0.46	0.29	0.27	0.79	0.69	0.38	0.33
RMSE	83.37	0.79	1.34	1.19	192.42	0.79	1.34	1.19	0.65	0.73	0.39	0.40	1.22	0.88	0.49	0.43

TABLA A.5. 2. MÉTRICAS ESTADÍSTICAS PARA EVALUAR EL COMPORTAMIENTO DE LAS CARACTERÍSTICAS Y LOS MODELOS PARA ELONGACIÓN A LA ROTURA: PREGUNTAS a Y b.

ELONGACIÓN A LA ROTURA																
	PRENGUNTA a								**PREGUNTA b**							
	URE vs Mn				URE vs Mw				FS:URE - Mn vs Mn				FS:URE - Mw vs Mw			
	CFS															
	LR	NN	RF	RC	LR	NN	RF	RC	LR	NN	RF	RC	LR	NN	RF	RC
R²	0.06	0.50	0.49	0.48	-0.03	0.54	0.49	0.48	0.67	0.37	0.70	0.85	0.40	0.39	0.72	0.74
MAE	161.18	7.89	2.21	3.02	348.06	9.68	2.20	3.02	1.96	2.89	1.75	0.83	2.67	2.85	1.55	1.87
RMSE	206.35	12.98	3.58	5.68	465.55	13.54	3.58	5.68	2.35	3.70	2.39	1.09	2.98	3.69	2.26	3.21
	W-LR															
	LR	NN	RF	RC	LR	NN	RF	RC	LR	NN	RF	RC	LR	NN	RF	RC
R^2	0.10	0.35	0.81	0.79	0.14	0.34	0.81	0.79	0.39	0.56	0.59	0.43	0.58	0.42	0.46	0.55
MAE	214.76	31.59	10.35	9.74	474.51	31.56	10.35	9.75	2.47	1.46	1.33	1.62	2.03	1.70	1.57	1.82
RMSE	263.36	34.92	10.74	10.23	650.84	34.95	10.74	10.24	2.93	2.08	1.67	2.52	2.43	2.43	2.07	2.23
	W-NN															
	LR	NN	RF	RC	LR	NN	RF	RC	LR	NN	RF	RC	LR	NN	RF	RC
R²	0.71	0.32	0.71	0.70	0.72	0.32	0.71	0.70	-0.51	0.53	0.67	0.72	-0.61	0.46	0.48	0.60
MAE	1009.84	36.89	11.66	13.54	2090.33	36.89	11.66	13.54	3.08	2.64	1.37	1.18	3.05	1.83	1.55	1.68
RMSE	1302.81	39.51	12.24	14.32	2629.86	39.51	12.24	14.32	3.99	4.93	1.70	1.49	3.97	2.41	2.19	2.19
	W-RF															
	LR	NN	RF	RC	LR	NN	RF	RC	LR	NN	RF	RC	LR	NN	RF	RC
R²	-0.79	-0.81	0.81	0.81	-0.71	-0.81	0.81	0.81	-0.37	-0.27	0.47	0.27	-0.14	-0.21	0.64	0.68
MAE	197.96	2.44	3.41	2.48	418.25	2.44	3.39	2.48	4.12	2.37	1.55	1.66	3.98	2.45	1.21	1.27
RMSE	231.15	2.73	3.69	2.85	505.63	2.73	3.66	2.85	5.00	3.01	1.89	2.55	5.00	2.96	1.66	1.77
	W-RC															
	LR	NN	RF	RC	LR	NN	RF	RC	LR	NN	RF	RC	LR	NN	RF	RC

R^2	0.88	0.81	0.81	0.80	0.88	0.81	0.81	0.80	-0.37	0.48	0.71	0.78	-0.14	0.53	0.51	0.57
MAE	1.50	35.11	6.30	5.25	1.49	35.10	6.31	5.29	4.12	1.53	1.20	1.27	3.98	1.60	1.48	1.56
RMSE	1.67	35.41	6.62	5.64	1.67	35.41	6.64	5.69	5.00	2.27	1.43	1.93	5.00	2.08	1.94	2.24

TABLA A.5. 3 MÉTRICAS ESTADÍSTICAS PARA EVALUAR EL COMPORTAMIENTO DE LAS CARACTERÍSTICAS Y LOS MODELOS PARA RESISTENCIA A LA ROTURA: PREGUNTAS a Y b.

RESISTENCIA A LA ROTURA																
	PREGUNTA a								**PREGUNTA b**							
	URE vs Mn				URE vs Mw				FS:URE - Mn vs Mn				FS:URE - Mw vs Mw			
	CFS															
	LR	NN	RF	RC	LR	NN	RF	RC	LR	NN	RF	RC	LR	NN	RF	RC
R^2	-0.79	0.08	0.78	0.80	-0.79	0.08	0.78	0.80	0.34	0.89	0.95	0.90	0.28	0.85	0.92	0.89
MAE	1570.56	25.03	11.98	11.20	3242.45	25.02	11.98	11.20	14.17	7.55	5.28	7.06	15.88	11.93	6.49	7.47
RMSE	1887.03	31.24	14.44	13.21	3838.05	31.24	14.44	13.21	18.45	9.97	6.16	7.99	19.62	14.40	7.67	9.05
	W-LR															
	LR	NN	RF	RC	LR	NN	RF	RC	LR	NN	RF	RC	LR	NN	RF	RC
R^2	0.13	0.33	0.53	0.49	0.03	0.48	0.53	0.49	0.72	0.77	0.90	0.92	0.49	0.83	0.91	0.92
MAE	1185.11	22.07	12.90	13.66	2432.71	21.04	12.90	13.66	9.89	11.92	7.17	7.21	13.93	11.08	7.51	8.14
RMSE	1274.48	25.72	17.55	19.09	2602.37	24.00	17.55	19.09	13.41	13.64	8.78	8.05	17.89	12.29	8.83	9.13
	W-NN															
	LR	NN	RF	RC	LR	NN	RF	RC	LR	NN	RF	RC	LR	NN	RF	RC
R^2	0.11	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.22	0.55	0.42	0.29	0.52	0.68	0.71
MAE	1762.02	17.23	16.51	16.59	3686.72	17.23	16.51	16.59	18.11	13.70	12.52	16.51	14.97	13.20	12.07	11.24
RMSE	1898.68	22.98	22.08	22.18	4061.46	22.98	22.08	22.18	20.00	19.90	15.98	20.06	18.52	16.10	15.53	14.93
	W-RF															
	LR	NN	RF	RC	LR	NN	RF	RC	LR	NN	RF	RC	LR	NN	RF	RC
R^2	0.76	0.29	0.88	0.84	0.73	0.29	0.88	0.84	0.80	0.89	0.94	0.94	0.73	0.86	0.94	0.94
MAE	597.51	14.48	10.81	10.61	1187.57	14.48	10.81	10.61	9.22	9.95	5.59	5.60	10.79	11.45	6.03	6.26
RMSE	700.27	19.20	12.87	12.64	1374.49	19.20	12.87	12.64	11.84	10.97	6.39	6.46	13.24	13.49	6.87	6.79
	W-RC															
	LR	NN	RF	RC	LR	NN	RF	RC	LR	NN	RF	RC	LR	NN	RF	RC
R^2	0.78	0.66	0.85	0.86	0.81	0.66	0.85	0.86	0.77	0.94	0.85	0.78	0.81	0.89	0.92	0.92
MAE	704.63	33.84	17.71	13.88	1387.04	33.84	17.71	13.88	10.66	6.29	8.56	11.14	9.76	7.33	6.83	7.05
RMSE	1003.52	55.24	20.47	16.50	1932.94	55.24	20.47	16.50	12.00	7.41	10.14	12.68	11.33	10.06	8.08	8.85

TABLA A.5. 4. MÉTRICAS ESTADÍSTICAS PARA EVALUAR EL COMPORTAMIENTO DE LAS CARACTERÍSTICAS Y LOS MODELOS PARA MÓDULO DE TENSIÓN: PREGUNTAS c Y d.

MÓDULO DE TENSIÓN																
MN	**PREGUNTA c**								**PREGUNTA d**							
	Mn vs URE				Mn vs Mw				FS:Mn - URE vs URE				FS:Mn - Mw vs Mw			
	CFS															
	LR	NN	RF	RC	LR	NN	RF	RC	LR	NN	RF	RC	LR	NN	RF	RC
R^2	0.79	0.83	0.90	0.75	0.74	0.85	0.94	0.91	0.84	0.83	0.92	0.91	0.73	0.87	0.93	0.93
MAE	0.42	0.56	0.34	0.44	0.37	0.44	0.23	0.24	0.33	0.49	0.23	0.24	0.36	0.45	0.22	0.22
RMSE	0.51	0.65	0.38	0.48	0.49	0.54	0.32	0.33	0.42	0.58	0.33	0.34	0.51	0.53	0.32	0.32
	W-LR															
	LR	NN	RF	RC	LR	NN	RF	RC	LR	NN	RF	RC	LR	NN	RF	RC
R^2	0.89	0.81	0.95	0.96	0.51	0.51	0.98	0.97	0.86	0.93	0.96	0.96	0.74	0.81	0.95	0.96
MAE	0.28	0.36	0.64	0.89	0.73	0.55	0.19	0.22	0.29	0.24	0.20	0.20	0.33	0.40	0.18	0.14
RMSE	0.35	0.46	0.76	0.98	1.23	0.83	0.29	0.37	0.46	0.31	0.29	0.28	0.51	0.54	0.25	0.19
	W-NN															
	LR	NN	RF	RC	LR	NN	RF	RC	LR	NN	RF	RC	LR	NN	RF	RC
R^2	0.00	0.81	0.96	0.79	0.00	0.70	0.97	0.97	0.91	0.93	0.93	0.90	0.82	0.87	0.94	0.85
MAE	0.35	0.72	0.76	0.39	0.35	0.47	0.22	0.18	0.27	0.24	0.23	0.24	0.31	0.35	0.19	0.21
RMSE	0.69	0.81	0.86	0.53	0.69	0.75	0.35	0.27	0.34	0.31	0.32	0.33	0.42	0.48	0.26	0.36
	W-RF															
	LR	NN	RF	RC	LR	NN	RF	RC	LR	NN	RF	RC	LR	NN	RF	RC
R^2	0.00	-0.06	0.95	0.00	0.00	0.86	0.93	0.85	0.59	0.96	0.92	0.93	0.28	0.84	0.92	0.87
MAE	0.37	1.52	1.61	1.86	0.37	0.40	0.24	0.24	0.32	0.23	0.28	0.26	0.37	0.29	0.20	0.23
RMSE	0.71	1.67	1.73	1.98	0.71	0.56	0.41	0.46	0.60	0.32	0.35	0.31	0.74	0.39	0.28	0.33
	W-RC															
	LR	NN	RF	RC	LR	NN	RF	RC	LR	NN	RF	RC	LR	NN	RF	RC

R^2	0.78	0.60	0.76	0.96	0.39	0.97	0.94	0.97	0.85	0.90	0.96	0.95	0.38	0.97	0.97	0.96
MAE	0.56	0.42	0.42	0.42	0.81	0.25	0.21	0.13	0.30	0.23	0.21	0.21	0.45	0.16	0.16	0.15
RMSE	0.70	0.65	0.52	0.51	1.11	0.31	0.40	0.18	0.42	0.36	0.34	0.35	0.72	0.18	0.21	0.21

TABLA A.5. 5. MÉTRICAS ESTADÍSTICAS PARA EVALUAR EL COMPORTAMIENTO DE LAS CARACTERÍSTICAS Y LOS MODELOS PARA ELONGACIÓN A LA ROTURA: PREGUNTAS c Y d.

ELONGACIÓN A LA ROTURA																
	PREGUNTA c								**PREGUNTA d**							
	Mn vs URE				Mn vs Mw				FS:Mn - URE vs URE				FS:Mn - Mw vs Mw			
	CFS															
	LR	NN	RF	RC	LR	NN	RF	RC	LR	NN	RF	RC	LR	NN	RF	RC
R^2	-0.34	0.35	0.69	0.80	-0.19	0.78	0.40	0.59	0.82	0.87	0.59	0.58	-0.16	0.62	0.87	0.82
MAE	16.39	3.17	11.62	5.06	3.08	1.25	1.58	1.56	1.08	0.90	1.35	1.35	3.06	1.77	0.87	0.89
RMSE	16.66	4.09	11.75	5.21	3.50	1.45	1.93	1.96	1.44	1.19	1.74	1.91	3.52	2.32	1.04	1.20
	W-LR															
	LR	NN	RF	RC	LR	NN	RF	RC	LR	NN	RF	RC	LR	NN	RF	RC
R^2	-0.33	-0.26	0.00	0.00	-0.48	-0.41	-0.81	-0.72	-0.33	-0.10	0.59	0.56	-0.48	-0.58	0.79	0.68
MAE	5.76	7.77	8.63	8.63	4.07	4.40	2.09	1.96	2.09	1.33	1.50	1.36	3.11	2.38	1.20	1.51
RMSE	6.11	8.04	8.86	8.86	5.20	5.76	3.55	3.40	3.13	2.24	1.97	2.09	3.88	3.07	1.40	2.18
	W-NN															
	LR	NN	RF	RC	LR	NN	RF	RC	LR	NN	RF	RC	LR	NN	RF	RC
R^2	0.79	0.72	0.74	0.84	-0.40	0.84	0.58	0.57	0.84	0.68	0.64	0.57	0.35	0.64	0.81	0.81
MAE	8.83	10.81	10.95	11.77	6.35	1.62	1.43	1.56	2.15	1.37	1.25	1.36	2.42	1.51	0.98	0.91
RMSE	9.53	11.97	11.38	11.93	8.09	1.79	1.66	1.94	2.57	1.64	1.60	1.87	3.44	2.13	1.22	1.32
	W-RF															
	LR	NN	RF	RC	LR	NN	RF	RC	LR	NN	RF	RC	LR	NN	RF	RC
R^2	-0.33	-0.16	0.03	-0.26	-0.48	-0.21	-0.01	0.73	-0.33	0.87	0.75	0.78	-0.48	0.59	0.67	0.74
MAE	8.01	23.74	23.03	29.77	5.27	7.04	1.79	1.13	2.51	1.47	1.18	1.36	4.11	2.13	1.26	1.13
RMSE	8.27	23.84	23.23	29.93	6.73	16.93	2.35	1.45	3.78	1.72	1.42	1.63	4.99	2.61	1.52	1.49
	W-RC															
	LR	NN	RF	RC	LR	NN	RF	RC	LR	NN	RF	RC	LR	NN	RF	RC
R^2	0.59	0.58	0.46	0.45	-0.43	-0.68	0.65	0.70	-0.33	-0.22	0.50	0.31	-0.41	-0.65	0.62	0.66
MAE	5.72	11.89	19.33	18.44	4.30	1.84	1.28	1.19	2.84	1.62	1.40	1.61	3.26	2.57	1.22	1.18
RMSE	6.05	12.04	20.15	19.43	5.47	2.44	1.62	1.77	4.28	3.10	1.82	2.48	4.03	3.54	1.64	1.72

TABLA A.5. 6. MÉTRICAS ESTADÍSTICAS PARA EVALUAR EL COMPORTAMIENTO DE LAS CARACTERÍSTICAS Y LOS MODELOS PARA RESISTENCIA A LA ROTURA: PREGUNTAS c Y d.

RESISTENCIA A LA ROTURA																
	PREGUNTA c								**PREGUNTA d**							
	Mn vs URE				Mn vs Mw				FS:Mn - URE vs URE				FS:Mn - Mw vs Mw			
	CFS															
	LR	NN	RF	RC	LR	NN	RF	RC	LR	NN	RF	RC	LR	NN	RF	RC
R^2	-0.51	0.80	0.00	0.00	0.75	0.82	0.70	0.72	0.81	0.85	0.83	0.83	0.75	0.83	0.87	0.84
MAE	17.22	19.08	24.75	20.60	18.92	12.65	11.06	11.95	8.92	8.35	9.19	9.57	10.92	8.28	8.55	9.78
RMSE	19.34	20.91	27.75	22.67	25.60	14.52	13.40	14.67	11.38	9.82	10.86	11.45	12.65	10.52	10.29	11.98
	W-LR															
	LR	NN	RF	RC	LR	NN	RF	RC	LR	NN	RF	RC	LR	NN	RF	RC
R^2	0.38	0.46	0.72	0.73	0.80	0.81	0.78	0.74	0.63	0.53	0.57	0.51	0.83	0.82	0.92	0.88
MAE	19.30	22.08	37.91	44.46	16.44	9.53	10.19	11.20	13.11	15.51	12.30	14.44	8.74	12.25	7.08	8.32
RMSE	21.15	24.74	41.46	47.68	21.17	11.34	11.75	13.07	14.88	17.47	16.98	19.55	10.57	13.63	7.93	9.43
	W-NN															
	LR	NN	RF	RC	LR	NN	RF	RC	LR	NN	RF	RC	LR	NN	RF	RC
R^2	0.15	0.25	0.32	-0.02	0.78	0.80	0.86	0.87	0.79	0.89	0.87	0.84	0.75	0.93	0.94	0.96
MAE	20.73	40.52	28.60	25.68	13.50	15.86	8.22	7.90	9.68	8.55	7.54	8.36	14.83	12.74	5.64	4.84
RMSE	22.79	44.37	33.02	29.58	20.82	20.33	10.50	11.07	11.71	9.79	9.38	10.43	17.63	14.64	6.83	6.15
	W-RF															
	LR	NN	RF	RC	LR	NN	RF	RC	LR	NN	RF	RC	LR	NN	RF	RC
R^2	0.68	0.78	-0.49	0.54	0.76	0.56	0.85	0.89	0.76	0.84	0.83	0.81	0.70	0.89	0.93	0.92
MAE	20.05	40.12	34.78	43.13	13.89	27.82	8.62	8.12	10.26	9.66	8.98	9.80	12.86	10.70	6.30	7.02
RMSE	22.05	43.76	39.61	46.62	19.29	37.24	10.39	9.70	12.80	11.52	10.79	11.76	17.09	14.04	7.29	7.88

	W-RC															
	LR	NN	RF	RC	LR	NN	RF	RC	LR	NN	RF	RC	LR	NN	RF	RC
R^2	0.45	0.61	0.51	-0.25	0.72	0.78	0.82	0.67	0.72	0.83	0.81	0.80	0.72	0.91	0.92	0.91
MAE	21.61	45.17	27.62	27.90	16.54	15.26	10.28	11.82	13.89	10.68	10.36	9.99	12.22	10.76	7.54	7.26
RMSE	24.14	48.64	31.85	31.90	25.39	19.60	11.61	14.80	16.72	14.44	11.86	11.72	15.92	13.35	9.79	8.91

TABLA A.5. 7. MÉTRICAS ESTADÍSTICAS PARA EVALUAR EL COMPORTAMIENTO DE LAS CARACTERÍSTICAS Y LOS MODELOS PARA MÓDULO DE TENSIÓN: PREGUNTAS e Y f.

MÓDULO DE TENSIÓN																
	PREGUNTA e								**PREGUNTA f**							
	Mw vs URE				Mw vs Mn				FS:Mw - URE vs URE				FS:Mw - Mn vs Mn			
	CFS															
	LR	NN	RF	RC	LR	NN	RF	RC	LR	NN	RF	RC	LR	NN	RF	RC
R^2	0.88	0.87	0.86	0.84	0.83	0.88	0.98	0.98	0.87	0.92	0.95	0.95	0.71	0.96	0.98	0.99
MAE	0.30	0.59	0.37	0.40	0.32	0.49	0.17	0.15	0.29	0.33	0.21	0.20	0.34	0.27	0.13	0.11
RMSE	0.49	0.66	0.41	0.45	0.48	0.58	0.26	0.17	0.45	0.39	0.29	0.28	0.48	0.36	0.19	0.15
	W-LR															
	LR	NN	RF	RC	LR	NN	RF	RC	LR	NN	RF	RC	LR	NN	RF	RC
R^2	0.75	0.66	0.96	0.94	0.80	0.88	0.37	0.73	0.89	0.93	0.96	0.91	0.79	0.94	0.96	0.93
MAE	0.34	0.81	0.52	0.59	0.32	0.45	0.37	0.38	0.23	0.24	0.20	0.23	0.32	0.35	0.19	0.18
RMSE	0.46	0.91	0.69	0.65	0.42	0.53	0.62	0.49	0.35	0.31	0.31	0.32	0.47	0.42	0.23	0.25
	W-NN															
	LR	NN	RF	RC	LR	NN	RF	RC	LR	NN	RF	RC	LR	NN	RF	RC
R^2	0.92	-0.44	0.95	0.00	0.51	0.61	-0.05	-0.27	0.79	0.92	0.91	0.89	0.45	0.96	0.96	0.96
MAE	0.51	1.61	1.34	1.44	0.38	0.44	0.66	0.85	0.33	0.24	0.27	0.28	0.33	0.20	0.18	0.17
RMSE	0.69	1.74	1.45	1.59	0.59	0.68	0.99	1.35	0.45	0.36	0.35	0.35	0.62	0.25	0.23	0.21
	W-RF															
	LR	NN	RF	RC	LR	NN	RF	RC	LR	NN	RF	RC	LR	NN	RF	RC
R^2	0.88	0.90	0.58	0.88	0.87	0.89	0.98	0.96	0.88	0.82	0.96	0.95	0.86	0.90	0.98	0.98
MAE	0.33	0.54	0.40	0.30	0.31	0.44	0.20	0.21	0.30	0.39	0.19	0.22	0.30	0.39	0.16	0.16
RMSE	0.42	0.61	0.55	0.36	0.41	0.50	0.32	0.31	0.38	0.51	0.30	0.30	0.37	0.46	0.20	0.18
	W-RC															
	LR	NN	RF	RC	LR	NN	RF	RC	LR	NN	RF	RC	LR	NN	RF	RC
R^2	0.78	0.50	0.72	0.94	0.36	0.90	0.97	0.85	0.77	0.88	0.96	0.95	0.44	0.95	0.98	0.98
MAE	0.40	0.37	0.40	0.70	0.34	0.23	0.23	0.31	0.36	0.25	0.21	0.21	0.33	0.25	0.15	0.12
RMSE	0.67	0.66	0.50	0.75	0.63	0.37	0.37	0.49	0.50	0.37	0.34	0.35	0.61	0.33	0.20	0.15

TABLA A.5. 8. MÉTRICAS ESTADÍSTICAS PARA EVALUAR EL COMPORTAMIENTO DE LAS CARACTERÍSTICAS Y LOS MODELOS PARA ELONGACIÓN A LA ROTURA: PREGUNTAS e Y f.

ELONGACIÓN A LA ROTURA																
	PRENGUNTA e								**PREGUNTA f**							
	Mw vs URE				Mw vs Mn				FS:Mw - URE vs URE				FS:Mw - Mn vs Mn			
	CFS															
	LR	NN	RF	RC	LR	NN	RF	RC	LR	NN	RF	RC	LR	NN	RF	RC
R^2	0.88	0.88	0.61	0.52	0.63	-0.11	-0.46	-0.50	0.78	0.41	0.56	0.53	0.52	0.76	0.75	0.71
MAE	4.63	22.18	17.15	19.70	2.77	3.89	5.93	8.06	1.78	2.10	1.40	1.32	2.01	1.28	1.04	1.08
RMSE	4.73	22.52	17.23	19.78	3.20	4.99	8.34	11.42	2.21	2.96	1.71	1.94	2.33	1.55	1.36	1.55
	W-LR															
	LR	NN	RF	RC	LR	NN	RF	RC	LR	NN	RF	RC	LR	NN	RF	RC
R^2	0.88	0.88	0.84	0.79	0.87	0.59	0.56	0.61	0.85	0.81	0.51	0.61	0.77	0.62	0.70	0.71
MAE	2.35	2.57	4.10	11.57	1.75	1.53	1.41	1.27	2.02	1.31	1.56	1.31	2.51	1.34	1.24	1.09
RMSE	2.60	2.82	4.65	13.37	2.12	1.99	1.78	1.73	2.48	1.59	1.83	1.76	2.87	1.68	1.47	1.53
	W-NN															
	LR	NN	RF	RC	LR	NN	RF	RC	LR	NN	RF	RC	LR	NN	RF	RC
R^2	-0.23	0.85	0.62	0.56	0.21	0.01	-0.49	0.05	0.74	0.53	0.67	0.62	0.73	0.70	0.62	0.61
MAE	36.76	14.21	18.03	16.63	3.48	4.01	4.77	3.59	3.08	1.75	1.31	1.36	1.52	1.40	1.47	1.43
RMSE	37.03	14.33	18.10	17.60	4.60	6.51	7.36	4.68	3.55	2.18	1.61	1.98	1.84	1.75	2.05	2.28
	W-RF															

	LR	NN	RF	RC	LR	NN	RF	RC	LR	NN	RF	RC	LR	NN	RF	RC
R^2	0.58	0.58	0.45	0.45	-0.34	-0.51	-0.33	-0.30	-0.33	-0.22	0.50	0.31	-0.37	-0.56	0.74	0.79
MAE	5.50	11.53	19.71	19.38	4.01	4.36	8.54	8.68	2.84	1.62	1.40	1.61	3.32	2.78	1.12	0.81
RMSE	5.85	11.69	20.56	20.50	4.72	5.62	12.47	12.90	4.28	3.10	1.82	2.48	4.13	3.72	1.40	1.40
								W-RC								
	LR	NN	RF	RC	LR	NN	RF	RC	LR	NN	RF	RC	LR	NN	RF	RC
R^2	0.58	0.22	0.45	0.45	-0.34	-0.39	-0.45	-0.37	0.17	0.87	0.74	0.67	-0.37	-0.03	0.57	0.72
MAE	5.50	28.68	22.57	23.79	4.01	11.00	7.49	8.01	3.50	1.84	1.28	1.60	3.32	2.48	1.66	0.99
RMSE	5.85	28.75	22.65	23.86	4.72	14.44	11.02	11.60	4.81	2.40	1.53	1.96	4.13	3.63	2.36	1.70

TABLA A.5. 9. MÉTRICAS ESTADÍSTICAS PARA EVALUAR EL COMPORTAMIENTO DE LAS CARACTERÍSTICAS Y LOS MODELOS PARA RESISTENCIA A LA ROTURA: PREGUNTAS e Y f.

RESISTENCIA A LA ROTURA																
	PRENGUNTA e								**PREGUNTA f**							
	Mw vs URE				Mw vs Mn				FS:Mw - URE vs URE				FS:Mw - Mn vs Mn			
								CFS								
	LR	NN	RF	RC	LR	NN	RF	RC	LR	NN	RF	RC	LR	NN	RF	RC
R^2	-0.65	0.35	-0.74	-0.74	0.73	0.89	0.79	0.70	0.81	0.81	0.83	0.82	0.72	0.89	0.88	0.88
MAE	19.82	20.31	24.18	20.87	12.94	8.25	13.16	14.20	9.83	9.85	8.07	8.40	12.08	8.23	7.15	7.41
RMSE	21.72	22.31	26.76	22.90	14.50	9.46	14.32	15.49	12.35	11.69	10.61	10.83	13.67	9.85	8.74	8.87
								W-LR								
	LR	NN	RF	RC	LR	NN	RF	RC	LR	NN	RF	RC	LR	NN	RF	RC
R^2	0.16	0.81	0.00	0.00	0.73	0.84	0.29	0.12	0.73	0.79	0.69	0.47	0.72	0.85	0.93	0.92
MAE	18.61	19.59	19.23	17.25	11.78	9.87	16.50	20.20	11.50	9.71	11.17	15.32	11.59	8.72	6.69	6.58
RMSE	20.44	21.46	21.07	19.35	13.61	11.84	19.84	23.60	14.09	11.38	14.66	20.46	13.75	10.75	7.38	8.18
								W-NN								
	LR	NN	RF	RC	LR	NN	RF	RC	LR	NN	RF	RC	LR	NN	RF	RC
R^2	0.09	0.10	-0.09	-0.06	0.70	0.81	0.77	0.58	0.75	0.81	0.80	0.81	0.81	0.91	0.90	0.85
MAE	21.90	21.62	32.90	31.78	17.38	10.54	12.40	16.60	10.35	9.58	9.81	9.14	12.25	6.04	7.68	9.11
RMSE	24.27	25.14	37.34	35.78	18.74	11.96	14.24	18.60	12.40	11.29	11.41	11.21	13.36	7.68	8.18	10.45
								W-RF								
	LR	NN	RF	RC	LR	NN	RF	RC	LR	NN	RF	RC	LR	NN	RF	RC
R^2	0.73	0.51	-0.67	-0.49	0.89	0.91	0.72	0.77	0.80	0.79	0.85	0.84	0.86	0.91	0.89	0.89
MAE	19.95	16.72	29.45	31.45	15.66	10.08	12.15	10.77	10.37	10.00	8.12	8.48	10.74	7.69	7.20	7.26
RMSE	21.87	19.00	33.86	36.33	16.99	12.23	14.31	13.13	11.46	12.13	9.90	10.05	12.29	9.29	8.65	8.74
								W-RC								
	LR	NN	RF	RC	LR	NN	RF	RC	LR	NN	RF	RC	LR	NN	RF	RC
R^2	0.46	0.63	-0.69	-0.64	0.61	0.91	0.71	0.53	0.80	0.75	0.83	0.79	0.40	0.85	0.89	0.89
MAE	16.23	16.55	22.74	20.78	13.30	8.97	12.37	13.35	10.37	11.53	8.03	9.44	13.44	8.40	7.02	7.17
RMSE	18.76	18.90	25.06	23.20	16.59	10.69	13.89	15.72	11.46	14.37	10.36	11.46	18.33	10.23	8.48	8.92

CAPÍTULO 7: SELECCIÓN DE CARACTERÍSTICAS MULTIVALUADAS

TABLA A.7. 1. VALORES DE LAS DIFERENTES MÉTRICAS UTILIZADAS PARA EVALUAR EL RENDIMIENTO CON RESPECTO A LA BASE DE DATOS SINTÉTICA UNIVALUADA BASADA EN LA REPRESENTACIÓN MÍNIMA.

Representación mínima (análoga a URE)										
Materiales			**400**			**800**			**1600**	
Descriptores			**5**	**10**	**20**	**5**	**10**	**20**	**5**	**10**	**20**
Escenario sin ruido	**Target Lineal**	%CC	90%	84%	62%	90%	82%	62%	90%	82%	68%
		Sensibilidad	0%	20%	5%	0%	10%	5%	0%	10%	20%
		Especificidad	94.74%	91.11%	76.25%	94.74%	90%	76.25%	94.74%	90%	80%
		NER	47.37%	55.56%	40.63%	47.37%	50%	40.63%	47.37%	50%	50%
		Mean*	26	25	25	25	25	25	25	25	25
		SD*	0.5	0	0	0	0	0	0	0	0
	Target no Lineal	%CC	90%	82%	64%	90%	82%	68%	90%	84%	68%
		Sensibilidad	0%	10%	10%	0%	10%	20%	0%	20%	20%
		Especificidad	94.74%	90%	77.50%	94.74%	90%	80%	94.74%	91.11%	80%
		NER	47.37%	50%	43.75%	47.37%	50%	50%	47.37%	55.56%	50%
		Mean*	25	25	25	25	25	25	25	25	25
		SD*	0	0	0	0	0	0	0	0	0
Escenario con ruido	**Target Lineal**	%CC	92%	82%	64%	90%	82%	68%	90%	84%	72%
		Sensibilidad	20%	10%	10%	0%	10%	20%	0%	20%	30%
		Especificidad	95.79%	90%	77.50%	94.74%	90%	80%	94.74%	91.11%	82.50%
		NER	57.90%	50%	43.75%	47.37%	50%	50%	47.37%	55.56%	56.25%
		Mean*	25	25	25	25	25	25	25	25	25
		SD*	0.5	0.5	0	0.5	0	0	0	0	0
	Target no Lineal	%CC	90%	80%	68%	90%	82%	62%	90%	82%	70%
		Sensibilidad	0%	0%	20%	0%	10%	5%	0%	10%	25%
		Especificidad	94.74%	88.89%	80%	94.74%	90%	76.25%	94.74%	90%	81.25%
		NER	47.37%	44.45%	50%	47.37%	50%	40.63%	47.37%	50%	53.13%
		Mean*	25	25	25	25	25	25	25	25	25
		SD*	0.5	0	0	0	0	0	0.5	0	0

TABLA A.7. 2. VALORES DE LAS DIFERENTES MÉTRICAS UTILIZADAS PARA EVALUAR EL RENDIMIENTO CON RESPECTO A LA BASE DE DATOS SINTÉTICA UNIVALUADA BASADA EN LA MEDIA ARITMÉTICA.

Representación basada en valor promedio (análoga a Mn o Mw)										
Materiales			**400**			**800**			**1600**	
Descriptores			**5**	**10**	**20**	**5**	**10**	**20**	**5**	**10**	**20**
Escenario sin ruido	**Target Lineal**	%CC	94%	88%	74%	100%	94%	78%	100%	96%	86%
		Sensibilidad	40%	40%	35%	100%	70%	45%	100%	80%	65%
		Especificidad	96.84%	93.33%	83.75%	100%	96.67%	86.25%	100%	97.78%	91.25%
		NER	68.42%	66.67%	59.38%	100%	83.34%	65.63%	100%	88.89%	78.13%
		Mean*	25%	25	25	NA	25	25	NA	25	25
		SD*	1%	0	0	NA	0.5	0	NA	1	0
	get no	%CC	90%	82%	64%	92%	84%	68%	90%	80%	80%

		Sensibilidad	0%	10%	10%	20%	20%	20%	0%	0%	50%
		Especificidad	94.74%	90%	77.50%	95.79%	91.11%	80%	94.74%	88.89%	87.50%
		NER	47.37%	50%	43.75%	57.90%	55.56%	50%	47.37%	44.45%	68.75%
		Mean*	25	25	25	25	25	25	25	25	25
		SD*	0.5	0	0	0.5	0	0	0	0	0
Escenario con ruido	**Target Lineal**	%CC	96%	88%	76%	100%	96%	80%	100%	96%	84%
		Sensibilidad	60%	40%	40%	100%	80%	50%	100%	80%	60%
		Especificidad	97.89%	93.33%	85%	100%	97.78%	87.50%	100%	97.78%	90%
		NER	78.95%	66.67%	62.50%	100%	88.89%	68.75%	100%	88.89%	75%
		Mean*	24	25	25	NA	25	25	NA	25	25
		SD*	1.5	0.5	0	NA	1.5	0	NA	1	0
	Target no Lineal	%CC	96%	90%	74%	98%	94%	76%	98%	96%	90%
		Sensibilidad	60%	50%	35%	80%	70%	40%	80%	80%	75%
		Especificidad	97.89%	94.44%	83.75%	98.95%	96.67%	85%	98.95%	97.78%	93.75%
		NER	78.95%	72.22%	59.38%	89.48%	83.34%	62.50%	89.48%	88.89%	84.38%
		Mean*	25.0	25.0	25.0	26.0	25.0	25.0%	26.0	25.0	25.0
		SD*	1	0.5	0	0	0.5	0%	0	1	0