ISBN: 978-628-95207-4-3. ISSN: 2414-6390. Digital Object Identifier: 10.18687/LACCEI2023.1.1.1572

Improved predictive modeling of polymeric materials through a hybrid approach of machine learning and expert intervention.

Fiorella Cravero, PhD[1], Ignacio Ponzoni, PhD[2], y Mónica F. Díaz, PhD[3]
[1,2]*Instituto de Ciencias e Ingeniería de la Computación (ICIC), y Departamento de Ciencias e Ingeniería de la Computación - Universidad Nacional del Sur (DCIC-UNS) - Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Argentina, fiorella.cravero@cs.uns.edu.ar, ip@cs.uns.edu.ar*
[3]*Planta Piloto de Ingeniería Química (PLAPIQUI UNS-CONICET), y Departamento de Ingeniería Química (DIQ-UNS) mdiaz@plapiqui.edu.ar*

Abstract– This work describes a hybrid methodology that combines machine learning and expert intervention to improve the predictive modeling of high-interest properties of polymeric materials. Although these materials have many advantages, developing a new material with specific properties, from a new molecular structure, is a great challenge and a costly and time-consuming task. The demand for materials with very specific properties continues to grow, so machine learning techniques have been applied for the prediction of these properties. The hybrid methodology was developed in an evolutionary way from expert intervention at the end of the machine learning process to a more determinative intervention throughout the cycle. This allows for more robust and reliable models for the design of new materials, which can help designers obtain property profiles for prototypes prior to the synthesis stage, saving time and resources.

Keywords. Polymeric materials, Machine learning, Expert-in-the-Loop, Mechanical properties, Refractive index.

Mejora del modelado predictivo de materiales poliméricos mediante un enfoque híbrido de aprendizaje automático y la intervención de expertos

Fiorella Cravero, PhD[1], Ignacio Ponzoni, PhD[2], y Mónica F. Díaz, PhD[3]
[1,2]*Instituto de Ciencias e Ingeniería de la Computación (ICIC), y Departamento de Ciencias e Ingeniería de la Computación - Universidad Nacional del Sur (DCIC-UNS) - Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Argentina, fiorella.cravero@cs.uns.edu.ar, ip@cs.uns.edu.ar*
[3]*Planta Piloto de Ingeniería Química (PLAPIQUI UNS-CONICET), y Departamento de Ingeniería Química (DIQ-UNS) mdiaz@plapiqui.edu.ar*

Resumen– ***Este trabajo describe una metodología híbrida que combina machine learning y la intervención de expertos para mejorar el modelado predictivo de propiedades de alto interés de materiales poliméricos. A pesar de que estos materiales tienen muchas ventajas, desarrollar un nuevo material con propiedades específicas, desde una nueva estructura molecular, constituye un gran desafío y es una tarea costosa que lleva mucho tiempo. La demanda de materiales con propiedades muy específicas sigue creciendo, por lo que las técnicas de machine learning se han estado aplicando para la predicción de estas propiedades. La metodología híbrida se desarrolló en forma evolutiva desde una intervención del experto al final del proceso de machine learning, hasta una intervención más determinante en todo el ciclo. Esto permite obtener modelos más robustos y confiables para el diseño de nuevos materiales, que pueden ayudar a los diseñadores a obtener perfiles de propiedades para prototipos previo a la etapa de síntesis, ahorrando tiempo y recursos.***

Palabras claves. Materiales poliméricos, Machine learning, Expert-in-the-Loop, Propiedades mecánicas, Índice de refracción.

Mejora del modelado predictivo de materiales poliméricos mediante un enfoque híbrido de aprendizaje automático y la intervención de expertos

Fiorella Cravero, PhD[1], Ignacio Ponzoni, PhD[2], y Mónica F. Díaz, PhD[3]
[1,2]*Instituto de Ciencias e Ingeniería de la Computación (ICIC), y Departamento de Ciencias e Ingeniería de la Computación - Universidad Nacional del Sur (DCIC-UNS) - Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Argentina, fiorella.cravero@cs.uns.edu.ar, ip@cs.uns.edu.ar*
[3]*Planta Piloto de Ingeniería Química (PLAPIQUI UNS-CONICET), y Departamento de Ingeniería Química (DIQ-UNS) mdiaz@plapiqui.edu.ar*

Resumen– Este trabajo describe una metodología híbrida que combina machine learning y la intervención de expertos para mejorar el modelado predictivo de propiedades de alto interés de materiales poliméricos. A pesar de que estos materiales tienen muchas ventajas, desarrollar un nuevo material con propiedades específicas, desde una nueva estructura molecular, constituye un gran desafío y es una tarea costosa que lleva mucho tiempo. La demanda de materiales con propiedades muy específicas sigue creciendo, por lo que las técnicas de machine learning se han estado aplicando para la predicción de estas propiedades. La metodología híbrida se desarrolló en forma evolutiva desde una intervención del experto al final del proceso de machine learning, hasta una intervención más determinante en todo el ciclo. Esto permite obtener modelos más robustos y confiables para el diseño de nuevos materiales, que pueden ayudar a los diseñadores a obtener perfiles de propiedades para prototipos previo a la etapa de síntesis, ahorrando tiempo y recursos.
Palabras claves. Materiales poliméricos, Machine learning Expert-in-the-Loop, Propiedades mecánicas, Índice de refracción.

I. Introducción

Dentro del amplio campo de los materiales, destacan los materiales poliméricos completamente sintéticos por sus múltiples ventajas y versatilidad. Se comenzaron a desarrollar a inicios del siglo 20 y han sido introducidos con éxito en industrias de todo tipo, creando nuevas maneras de construir, ensamblar y producir diferentes productos, tales como la industria automotriz, de alimentos, textil y electrónica, entre otras [1]. Al elegir un material polimérico para una aplicación específica, es muy importante conocer el desempeño mecánico y algunas otras propiedades específicas de uso como por ejemplo ópticas, eléctricas, etc. No obstante, desarrollar un material nuevo desde el diseño molecular hasta el producto final, es una tarea muy difícil y costosa, que además lleva mucho tiempo, ya que no todos los prototipos sintetizados llegan a tener el perfil de propiedades deseado y deben ser descartados y comenzar de nuevo. En este sentido, las técnicas de *machine learning* se han estado aplicando para la predicción de estas propiedades desde los años 90s [2]. Este campo, también denominado informática de polímeros, ha sido muy poco explorado debido a la complejidad que presentan los materiales poliméricos (macromoléculas polidispersas), y la poca cantidad de datos en la literatura (comparado a otros campos), a pesar de que la demanda de este mercado por materiales con propiedades muy específicas sigue creciendo [3]. Los avances en el modelado estructura-propiedad conocidos por sus siglas en inglés QSPR para *Quantitative Structure Property Relationship*, se han realizado sobre prototipos moleculares muy reducidos como monómeros, trímeros o similares [4]. Este abordaje tiene una perspectiva muy limitada respecto de lo que en realidad es un material polimérico (muy altos pesos y polidispersión) [5].

En la actualidad existe una necesidad creciente de guiar el descubrimiento *in silico* de nuevos polímeros industriales mediante enfoques de *machine learning* supervisado que identifiquen relaciones estructura-propiedad a partir de la información contenida en bases de datos. Estas bases de datos se crean a partir de datos medidos *in situ* con el material real, del cual se conoce su estructura molecular, pesos, historia térmica, etc. Con estas bases de datos se procede al desarrollo de modelos QSPR, que luego podrán ser empleados como un test virtual de una nueva estructura molecular, para estimar el valor de la propiedad estudiada.

El dogma central de la quimioinformática establece que la estructura de una molécula determina sus propiedades, y que, dada una estructura, en principio es posible predecir las propiedades resultantes de una molécula [6]. Esta relación puede modelarse a través de modelos QSPR que como se dijo antes, se obtienen aplicando *machine learning*, donde se busca encontrar relaciones empíricas entre la propiedad de interés y las variables que caracterizan las estructuras moleculares mediante descriptores moleculares (DMs) [7,8]. En nuestro grupo, tenemos desarrollos de este tipo de herramientas informáticas desde 2012, y nuestra propuesta se puede resumir en el esquema de la Figura 1.

Por lo tanto, los modelos QSPR pueden ser empleados para predecir propiedades de interés previo a la etapa de síntesis química, contribuyendo de este modo a acelerar el diseño de nuevos materiales y a reducir los costos asociados a su

Digital Object Identifier: (only for full papers, inserted by LACCEI).
ISSN, ISBN: (to be inserted by LACCEI).
DO NOT REMOVE

desarrollo [4,9,10]. Actualmente, los enfoques de aprendizaje automático han comenzado a reemplazar las heurísticas químicas tradicionales en el descubrimiento de nuevos materiales [11]. La combinación de ciencia de datos e inteligencia artificial se ha denominado el "cuarto paradigma de la ciencia", y el número de aplicaciones en el campo químico está creciendo aceleradamente [12,13]. Existe evidencia de la conveniencia de utilizar estrategias de ciencia de datos en combinación con la experiencia de un experto que mejora y acelera el lanzamiento de nuevos materiales al mercado. Una de las estrategias para abordar esto es la incorporación de un diseñador de materiales al ciclo de proceso de ML que guíe el diseño de nuevos materiales, uniendo la capacidad de ML para analizar y detectar patrones en los datos con la intuición humana. Esta colaboración puede ocurrir cuando el experto humano sugiere nuevos materiales candidatos que luego pueden refinarse aún más a través de la investigación computacional y experimental hasta que se encuentren polímeros que cumplan con las propiedades deseadas y/o los objetivos de rendimiento a través del testeo virtual [14]. Por otro lado, el humano experto puede intervenir guiando al algoritmo ML en diferentes etapas de su proceso. Estos enfoques se conocen como Expert-in-the-Loop y facilitan la introducción y el uso de la experiencia de un experto.

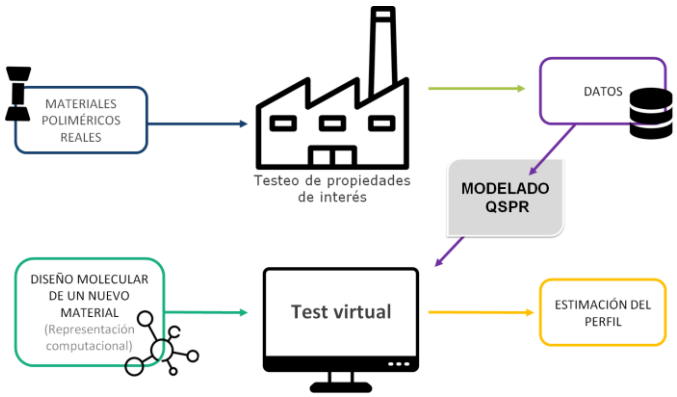

Fig. 1. Esquema del proceso de testeo virtual de propiedades para el diseño de nuevos materiales poliméricos.

El modelado 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, están formadas 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, lo que se denomina polidispersidad. Las variaciones en la frecuencia de aparición 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. Esta es la principal razón por la que muchas aproximaciones o soluciones informáticas desarrolladas, y que continúan desarrollándose, para el diseño racional de fármacos no sean directamente aplicables, ni lo suficientemente efectivas, en el ámbito de la informática de polímeros [6]. Casi la totalidad de los trabajos publicados en esta área utilizan modelos moleculares sintéticos (simplificados), es decir, caracterizan el polímero a través de una única URE. Si bien los resultados han sido prometedores, no se puede considerar que los DMs en los modelos computacionales sean los ideales, ni que los modelos obtenidos sirvan para todo el espectro de familias químicas [15]. En este sentido hay mucho trabajo por realizar en este tema, y no caben dudas de que es un trabajo interdisciplinario donde además de expertos en ciencias de la computación deberían trabajar expertos en materiales poliméricos. Por este motivo, consideramos que nuestra propuesta trae un abordaje más completo, al trabajar desde un inicio en forma híbrida, con aportes de ambas ciencias.

El objetivo del presente trabajo es describir una metodología híbrida para el desarrollo de modelos predictivos de propiedades de interés de materiales poliméricos, que incluye al experto en materiales, mostrando resultados para propiedades como índice refracción y mecánicas, derivadas del ensayo de tracción. El principal aporte de estas herramientas es colaborar en la primera etapa de desarrollo de nuevos materiales, cuando el diseñador está creando una nueva estructura molecular y necesita tener una estimación de las propiedades, previo a la síntesis. Este tipo de intervención puede representar un gran ahorro en tiempo de desarrollo de un nuevo material y en consecuencia también un menor costo.

II. Aspectos Teóricos

A. Propiedades Mecánicas

Hoy en día, la estructura molecular de los polímeros se diseña con precisión para cubrir las necesidades tecnológicas de la sociedad. En este sentido, las propiedades mecánicas deben conocerse ya que definen el perfil de aplicación del material. Si bien existen muchos tipos ensayos para propiedades mecánicas, el ensayo de tracción es uno de los más importantes por la completa información que brinda, por su simpleza y además está totalmente estandarizado y es relativamente económico. En este ensayo fundamental de la ciencia, una probeta se somete a tensión uniaxial hasta que se produce la falla. La Figura 2 muestra un gráfico con la curva de tensión deformación típica de un termoplástico dúctil, de la cual puede obtener información de diferentes propiedades como rigidez, resistencia, tenacidad y ductilidad. En resumen, una muestra es tensionada en una dirección hasta que se rompe, manteniendo la velocidad de ensayo constante, y como resultado se obtiene un completo perfil de comportamiento mecánico a una temperatura dada. De la curva se pueden calcular módulo de tensión, elongación a la rotura y resistencia a la rotura entre otras propiedades. Cabe destacar que hay aspectos de la

estructura molecular que influyen en el comportamiento mecánico. Los más importantes son el peso molecular, la polaridad de la molécula y finalmente la movilidad. Por lo tanto, estos aspectos serán tenidos en cuenta por el experto a la hora de elegir los mejores DMs para un modelo QSPR.

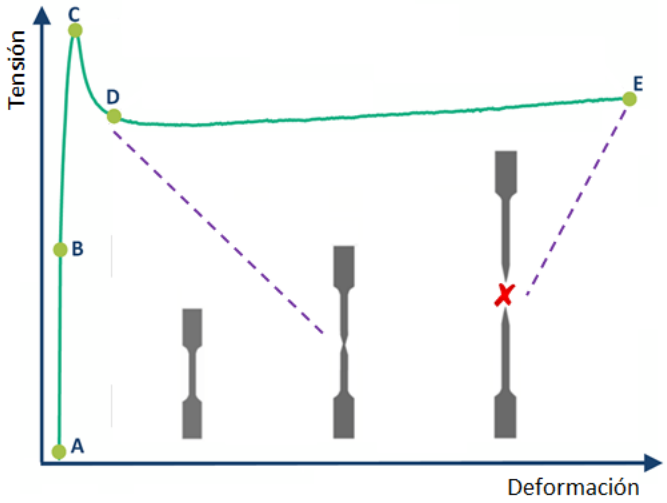

Fig. 2. Curva típica tensión-deformación proveniente del ensayo de tracción para un polímero dúctil.

Otro aspecto a tener en cuenta en el desarrollo de un modelo predictivo es la cantidad y calidad de los valores incorporados en la base de datos. Si bien los datos de materiales en general abundan, para el campo de los polímeros en particular resulta difícil armar bases con datos fidedignos y homogéneos, sobre todo para las propiedades mecánicas. Nosotros hemos creado una dataset de 77 termoplásticos, para las propiedades derivadas del ensayo de tracción. Está compuesta por 9 familias químicas, y el test se realizó a una temperatura de 20-25 °C, siguiendo normas [16].

B. Indice de Refraccion

Las propiedades ópticas están relacionadas tanto con el grado de cristalinidad como con la estructura real del polímero. En particular, una de las más importantes es el índice de refracción (IR), que es una medida de la capacidad del polímero para refractar o doblar la luz a medida que pasa a través del polímero. El IR es igual a la relación del seno de los ángulos de incidencia y la refracción de la luz que pasa a través del mismo. También se define como la relación entre la velocidad de la luz en el vacío y la velocidad de la luz en el material. La magnitud de IR está relacionada con la densidad de la sustancia y varía de 1.000 y 1.3333 para vacío y agua, a aproximadamente 1.5 para muchos polímeros y 2.5 para pigmento blanco, óxido de titanio (IV) (dióxido de titanio). El valor de IR es a menudo alto para los cristales y depende de la longitud de onda de la luz incidente y de la temperatura. Por lo general, se informa para la longitud de onda de la línea D de sodio transparente a 298 K. Los IRs para polímeros varían de 1.35 para politetrafluoroetileno a 1.67 para poliarilsulfona.

Una de las aplicaciones más actuales donde cobra especial interés el IR son las fibras ópticas. Hoy en día, casi todas las telecomunicaciones se producen a través de fibras ópticas en lugar de cables metálicos. La transmisión de señal con cables metálicos se realiza a través de electrones, mientras que la transmisión a través de fibras ópticas se realiza a través de fotones. Estas pueden ser fibras de vidrio que están revestidas con un polímero altamente refractivo de modo que la luz que ingresa a un extremo del filtro se transmite a través de la fibra, y emerge del otro extremo con poca pérdida de energía. También, puede estar formada por una combinación de materiales poliméricos, las llamadas POFs por su nombre del inglés *plastic optical fibers*, que, por lo general, consiste en un núcleo con un revestimiento y un recubrimiento exterior. El núcleo realiza la transmisión real de los fotones; el revestimiento restringe la luz para que viaje dentro del núcleo con poca pérdida de potencia de señal y poca distorsión de pulso; y el recubrimiento ayuda a proteger el material interno contra daños y presiones externas (Figura 3). Las POFs, presentan varias ventajas frente a las clásicas de SiO_2: son flexibles (la fibra óptica tradicional es rígida), livianas y de bajo costo de manejo y mantenimiento. En este sentido, la demanda de materiales poliméricos que cubran las necesidades de estos productos se ha incrementado exponencialmente, debido a su versatilidad y al creciente uso para instalaciones domésticas. Sin embargo, la velocidad de transmisión de datos de las POFs todavía presenta desafíos, ya que no alcanza la de las fibras de SiO_2 [17-19]. Una consideración importante en la composición de las POFs es que el núcleo, debe tener un IR más alto que la capa inmediata que lo reviste. Esta diferencia hace que la luz siga el camino a lo largo del núcleo, sin desviarse. En este sentido, la polaridad de la molécula del polímero está directamente relacionada con el IR, y por este motivo será tenido en cuenta por el experto en el modelado. Además, ahora hay un interés creciente en los polímeros de alto IR, ya que ciertas aplicaciones, como lentes, recubrimientos antirreflectantes y sensores de imagen, requieren menos material cuanto mayor sea el IR [20,21]. Por lo tanto, debido a la creciente demanda de materiales con características específicas y bajo costo, se hace necesario crearlos desde la etapa de diseño molecular.

Fig. 3. Corte de una fibra óptica donde se aprecian los dos componentes más importantes: el núcleo y el revestimiento. Además, esta fibra lleva capas de recubrimiento y protección.

Trabajamos con un dataset ensamblado y curado por nosotros a partir de otros dos, conteniendo 227 polímeros pertenecientes a

17 familias químicas, con un rango de IR de 1.3010 a 1.7100, medidos a 298 K [22].

C. Machine Learning y Visual Analytics

El aprendizaje automático, mayoritariamente conocido por su nombre en inglés, *machine learning*, es el área dedicada a la creación de algoritmos con la capacidad de mejorar su desempeño a través de la experiencia, es decir, algoritmos que puedan inducir modelos automáticamente. 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. La particularidad de estos algoritmos es que son guiados por los datos, es decir, si los 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 *machine learning* permiten generar automáticamente modelos, a partir de datos, para resolver estos tipos de problemas. Resumiendo, el *machine learning* es una estrategia de aprendizaje de una solución a partir de patrones presentes en los datos. Entonces, la calidad y la cantidad de estos datos son 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 en los conjuntos de Entrenamiento y Prueba, que tendrán diferentes tareas asignadas durante el proceso de aprendizaje y evaluación de este.

Los modelos de *machine learning*, 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: modelos de regresión, modelos de clasificación y modelos de ranking.

De acuerdo con 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.

Todos los datos no pueden ser utilizados para todas las tareas que se quieran analizar. ¿Es necesario explorar todas y cada una de las características o atributos?, ¿son todas igualmente importantes para determinada tarea? Los algoritmos de *machine learning* 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: *feature selection* y *feature learning*.

Machine learning y visual analytics son dos áreas interrelacionadas que se complementan entre sí en el análisis de datos. El *machine learning* utiliza algoritmos para aprender patrones a partir de datos y realizar predicciones o tomar decisiones automatizadas. El visual analytics se enfoca en el uso de herramientas visuales para explorar y comunicar información compleja de manera más efectiva. La combinación de ambos puede ayudar a los usuarios a comprender mejor los patrones y tendencias en los datos, a través de la identificación visual de los patrones encontrados por los algoritmos de aprendizaje automático. Además, el uso de herramientas de visualización puede ayudar a los usuarios a explorar y evaluar modelos de *machine learning*, así como tomar decisiones sobre los datos para que los filtre algunos y que las predicciones de los modelos sean más precisas.

III. MATERIALES Y MÉTODO

En esta sección se describen los conjuntos de datos y las herramientas informáticas empleados en el desarrollo de los modelos predictivos QSPR para propiedades mecánicas y ópticas. La metodología aplicada es en sí misma el resultado que se presenta en este trabajo, por este motivo se detalla en la siguiente sección.

A. Conjuntos de datos

Se utilizaron dos conjuntos de datos, uno para cada tipo de propiedad. En el caso de las propiedades mecánicas consta de 77 polímeros lineales, termoplásticos y amorfos que pertenecen a diferentes familias químicas: poliestirenos, polióxidos/éteres/acetales, poliésteres/tioésteres, polivinilos, poliamidas/tioamidas, polifenilenos, polisulfuros, poliimidas/tioimidas, policetonas/tiocetonas, y polisulfonas/sulfóxidos/sulfonatos/sulfonamidas. Estos polímeros fueron ensayados siguiendo normas (ASTM D638, ASTM D882-83 y DIN 53504.53A), dentro de una ventana de temperatura de 20-25 ºC, por debajo de las Tg [16]. En cuanto al conjunto de datos de IR, este consta de 227 polímeros que también pertenecen a diferentes familias químicas: poliolefinas, polióxidos, poliéteres, poliésteres, poliacetatos, polivinilos, poliamidas, poliimidas, policetonas, polifenilenos, policarbonatos, poliftalatos, polimetacrilatos, poliacrilatos, polisililenos, polisulfuros y polisulfonas. Los valores de IR se encuentran en un completo rango que va de 1.3010 a 1.7100 [22]. Los conjuntos de datos fueron creados y/o curados por nuestro equipo.

En cuanto a la división de los datos en entrenamiento y testeo, esta se realizó de modo que todas las familias químicas estuvieran representadas en cada conjunto.

B. Métodos y Herramientas Computacionales

En este trabajo se utilizaron principalmente dos herramientas, una de analítica visual llamada Videan desarrollada en nuestro grupo de investigación, y otra llamada Weka dedicada al aprendizaje automático y la minería de datos, desarrollada en la Universidad de Waikato en Nueva Zelanda.

VIDEAN es el acrónimo de *Visual and Interactive DEscriptor ANalysis*, que 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. De esta manera, los subconjuntos de descriptores seleccionados contienen información no redundante y más completa y específica desde el punto de vista fisicoquímico. VIDEAN muestra diferentes tipos de representaciones visuales coordinadas que capturan las relaciones e interacciones entre descriptores y también entre descriptor-propiedad:

- Grafos no dirigidos: permiten evitar conjuntos de descriptores redundantes y comparar entre subconjuntos de descriptores alternativos.
- Grafo bipartito: 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.

Weka es una sigla que significa *Waikato Environment for Knowledge Analysis*. Es una plataforma de software de código abierto y gratuita que proporciona una amplia gama de algoritmos de aprendizaje automático, técnicas de preprocesamiento de datos, herramientas de visualización y evaluación de modelos. Las características principales de Weka son:

- Una amplia variedad de algoritmos de aprendizaje automático que incluyen clasificación, regresión, clustering, etc.
- Una interfaz gráfica de usuario (GUI) fácil de usar, lo que la hace accesible para usuarios no técnicos.
- También es posible usar Weka a través de línea de comandos, lo que lo hace muy útil para flujos de trabajo automatizados.

En los trabajos previos que dieron origen a este se utilizaron diferentes métodos de machine learning provistos por Weka como: Redes Neuronales, Árboles de Decisión, Bosques Aleatorios, Bayes Simple (Naïve Bayes), K Vecinos más cercanos, Regresión Lineal, etc. Cada algoritmo tiene sus propias fortalezas y debilidades que fueron exploradas en cada trabajo de desarrollo de modelo predictivo QSPR para cada una de las propiedades.

IV. Resultados y discusiones

En esta sección se describe la metodología híbrida que hemos desarrollado para crear modelos predictivos para propiedades de interés de materiales poliméricos. Las primeras propiedades con que se trabajó fueron térmicas (Tg) y mecánicas derivadas del ensayo de tracción [16,23-25]. Estos primeros trabajos partían de una base de datos de las estructuras de polímeros y sus propiedades en estudio a los que se les calculaba los correspondientes DMs con softwares específicos como Dragon. El número de DMs que se pueden calcular es muy alto (orden de los miles), y para hacer una primera reducción se empleaban criterios matemáticos como por ejemplo eliminar las columnas que estaban altamente correlacionadas con otras. Este proceso se generaba automáticamente, por lo que no se tenía control sobre cuales columnas permanecían en el dataset y cuales eran eliminadas. Luego se aplicaba *feature selection* para obtener diferentes subconjuntos (subsets) de DMs y finalmente se los entrenaba y se obtenían los modelos predictivos correspondientes. En esta instancia participaba fuertemente el experto para poder elegir entre los modelos con mejores métricas de evaluación aquel que contuviera mejor complementariedad de información y mayor explicabilidad desde un punto de vista físico químico. Esta metodología se suele denominar Black-Box o caja negra, y está resumida en la Figura 4 en color verde. Si bien los resultados fueron promisorios, existía la necesidad de incrementar la intervención del experto para guiar con más precisión el proceso ajustando la información entrenada, limitando los DMs a solo aquellos que estuvieran fuertemente relacionados con la propiedad de estudio. Un ejemplo de esto fue la predicción del índice de refracción [22] donde desde un inicio el experto eligió DMs que tenían información del aspecto polar de la molécula, ya que la polaridad está muy relacionada con el valor del índice de refracción. Esta metodología con mayor intervención del experto se describe en la Figura 4 con color violeta. Como puede observarse, al final se comparan los mejores modelos de ambos caminos seguidos.

Además, se trabajó con el aspecto dúctil utilizando propiedades del ensayo de tracción (aspecto dúctil=elongación a la rotura/módulo de tensión), donde el experto hizo una preselección de DMs que contuviera información de los aspectos que están fuertemente relacionados con el desempeño mecánico [26]. En la tabla 1, se muestran los mejores modelos obtenidos por ambos caminos para ambas propiedades, donde finalmente se eligieron como finales, los provenientes del enfoque Expert-in-the-Loop, siguiendo los criterios que se explican a continuación. Una vez que se llega a esta etapa final, se puede elegir cuál modelo será el recomendado, y los criterios usados para esto se basan en la cardinalidad (número de descriptores), desempeño e interpretabilidad. Se evita la redundancia de información, y por el contrario se busca la complementariedad, es decir el aporte de diferentes aspectos importantes relacionados con la estructura y la propiedad. Esto brinda robustez y confiabilidad al modelo ya que los usuarios, en este caso diseñadores de materiales, tienen la mirada de un experto que interpreta y clarifica la caja negra que suelen ser los modelos de *machine learning*.

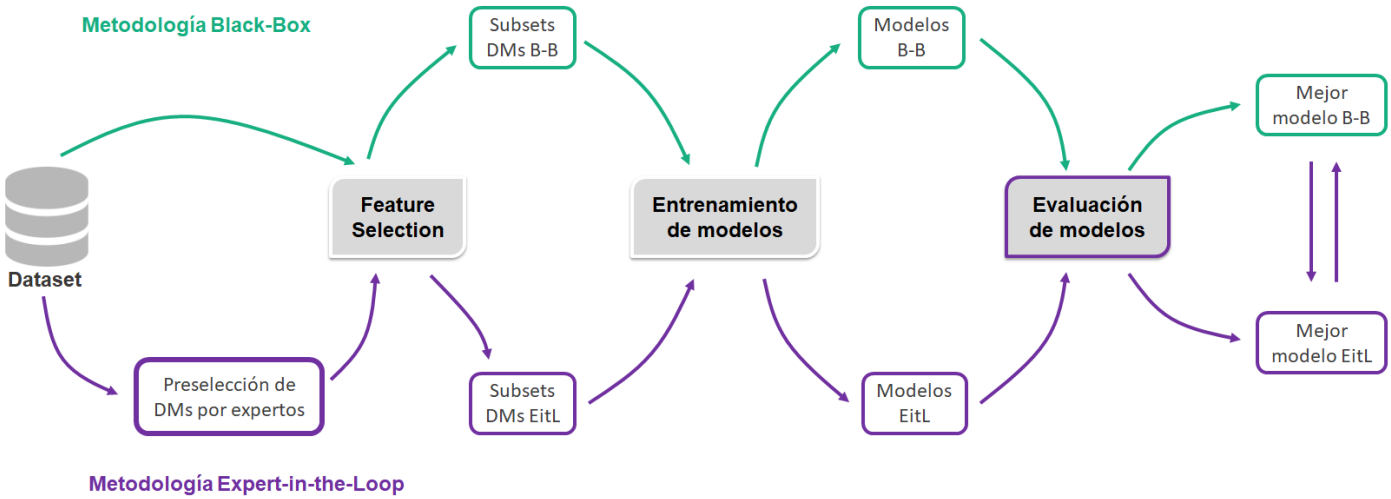

Fig. 4. Síntesis de los dos caminos metodológicos empleados para la inferencia de modelos predictivos. Ambos pueden retroalimentarse y/o compararse en el final. Arriba en verde la metodología denominada Black-Box (B-B) y abajo en violeta la de Expert-in-the-Loop (EitL).

TABLA I
COMPARACIÓN DE RESULTADOS DE LAS METODOLOGÍAS.

	Black-Box		Expert-in-the-Loop	
Propiedad	Índice de refracción	Aspecto dúctil	Índice de refracción	Aspecto dúctil
Cardinalidad	21	11	5	4
Métricas	0.9686 (R^2)	88.46 (%CC)	0.8987 (R^2)	88.46 (%CC)

Como se observa en la tabla I, utilizando una aproximación que involucre expertos como es Expert-in-the-Loop se reduce notablemente la cardinalidad de los modelos manteniendo una buena performance de los mismos. En los resultados, tanto para IR como para aspecto dúctil con la aproximación black-box se alcanzan valores de performance de 0.97 (R^2) para la propiedad óptica y 88.5 (%CC) para la propiedad mecánica. Sin embargo, recordemos que los dataset son de 227 y 77 polímeros respectivamente, por lo que la cardinalidad de ambos modelos (21 y 11 respectivamente) si bien es aceptable no es la deseada. Utilizando la metodología Expert-in-the-Loop, se logra reducir esta cardinalidad en un 76% y un 64% respectivamente. Esto significa que el modelo de IR tiene una cardinalidad de 5 y el modelo de aspecto dúctil de 4, conservando en el último caso el mismo rendimiento y en el caso de la propiedad óptica disminuyendo solo un 7%, conservando ambos una alta capacidad predictiva sin información redundante. Finalmente, cabe destacar que los modelos resultantes son más interpretables desde un punto de vista físicoquímico, ya que los DMs fueron seleccionados apropiadamente por un experto en el dominio, lo que hace que sean más confiables para los usuarios que no son expertos en aprendizaje automático, como los diseñadores de materiales, a quienes buscamos asistir con nuestras soluciones y herramientas informáticas.

V. CONCLUSIONES

En el presente trabajo se describe la metodología desarrollada con un enfoque híbrido para el desarrollo y mejora de modelos predictivos aplicado a propiedades de interés de materiales poliméricos. Este campo es muy poco explorado debido a la complejidad que representan las moléculas poliméricas y a la poca disponibilidad de datos fidedignos. No obstante, las metodologías ampliamente usadas para el diseño racional de drogas han sido el puntapié para trabajar en el campo de los materiales poliméricos. La metodología híbrida se desarrolló en forma evolutiva desde una intervención del experto solo al final del proceso de *machine learning*, hasta una intervención desde el inicio. En este sentido, el experto en propiedades de materiales realiza una preselección de los descriptores moleculares que ingresan al proceso de aprendizaje automático. Las etapas involucradas son dinámicas, ya que en cada una se toman decisiones para pasar a la siguiente, y finalmente se eligen los mejores modelos con criterios combinados donde participan tanto los informáticos como los expertos en materiales. De esta manera se logra robustez y confiabilidad en los modelos predictivos, esperando que las herramientas y modelos desarrollados por nuestro grupo sean útiles para un diseñador de nuevos materiales y que pueda utilizarlos como un test virtual para obtener un perfil de propiedades para sus prototipos, previo a la etapa de síntesis, ahorrando tiempo y recursos. Con la implementación de estas nuevas herramientas informáticas se podrán descartar aquellos prototipos que no reúnan los requerimientos buscados, antes de sintetizarlos.

AGRADECIMIENTOS

Este trabajo fue subsidiado parcialmente por el Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Argentina (Proyectos PIP 112-2017-0100829 y PIP 11220210100683CO), por la Universidad Nacional del Sur (UNS), Bahía Blanca, Argentina (Proyectos PGI 24/N052 y

PGI 24/M158), y por la Agencia Nacional de Promoción Científica y Tecnológica (Proyectos PICT 2018-4533 y PICT 2019-3350).

REFERENCIAS

[1] Ashby, Michael; Hugh Shercliff; David Cebon (2019). Materials: engineering, science, processing and design (4th ed.). Butterworth-Heinemann. ISBN 978-0-7506-8391-3

[2] 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.

[3] Utraki, 2002; Adams et al., 2004; Holdren, 2011, Ashby, Michael; Hugh Shercliff; David Cebon (2019). Materials: engineering, science, processing and design (4th ed.). Butterworth-Heinemann. ISBN 978-0-7506-8391-3]

[4] Audus, D. J., & de Pablo, J. J. (2017). Polymer informatics: opportunities and challenges.

[5] 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.

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

[7] Hansch, C., & Fujita, T. (1964). p-σ-π Analysis. A method for the correlation of biological activity and chemical structure. *Journal of the American Chemical Society*, 86(8), 1616-1626.

[8] Cherkasov, A., Muratov, E. N., Fourches, D., Varnek, A., Baskin, I. I., ⋯, Consonni, V. (2014). QSAR modeling: where have you been? Where are you going to? *J. of medicinal chemistry*, *57*(12), 4977-5010.

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

[10] Nosengo, N. (2016). Can artificial intelligence create the next wonder material? *Nature News*, *533*(7601), 22.

[11] George, J., & Hautier, G. (2020). Chemist versus Machine: Traditional Knowledge versus Machine Learning Techniques. *Trends in Chemistry*.

[12] Schwab, K. (2017). *The fourth industrial revolution*. Currency.

[13] Butler, K. T., Davies, D. W., Cartwright, H., Isayev, O., & Walsh, A. (2018). Machine learning for molecular and materials science. *Nature*, *559*(7715), 547-555.

[14] Ristoski, P., Gentile, A. L., Alba, A., Gruhl, D., & Welch, S. (2020). Large-scale relation extraction from web documents and knowledge graphs with human-in-the-loop. *Journal of Web Semantics*, *60*, 100546.

[15] Ramprasad, R., Batra, R., Pilania, G., Mannodi-Kanakkithodi, A., & Kim, C. (2017). Machine learning in materials informatics: recent applications and prospects. *npj Computational Materials*, *3*(1), 54.

[16] Palomba, D., Vazquez, G. E., & Díaz, M. F. (2014). Prediction of elongation at break for linear polymers. *Chemometrics and Intelligent Laboratory Systems*, *139*, 121-131.

[17] D. Kalymnios. Plastic Optical Fibres (POF). O.D.D. Soares (Ed.), Trends in Optical Fibre Metrology and Standards. NATO ASI Series (Series E: Applied Sciences), Springer, Dordrecht (1995), 10.1007/978-94-011-0035-9_37.

[18] M.S. Moslan, M.H.D. Othman, A. Samavati, M.A.M. Salim, M.A. Rahman, A.F. Ismail, H. Bakhtiar. Fabrication of polycarbonate polymer optical fibre core via extrusion method: the role of temperature gradient and collector speed on its characteristics. Opt. Fiber Technol., 55 (2020), p. 102162

[19] A. Theodosiou, K. Kalli. Recent trends and advances of fibre Bragg grating sensors in CYTOP polymer optical fibres. Opt. Fiber Technol., 54 (2020), p. 102079

[20] C.E. Carraher Jr. Carraher's Polymer Chemistry CRC Press (2016).

[21] E.K. Macdonald, M.P. Shaver. Intrinsic high refractive index polymers: intrinsic high refractive index polymers. Polym. Int., 64 (1) (2015), pp. 6-14

[22] Schustik, S. A., Cravero, F., Ponzoni, I., & Díaz, M. F. (2021). Polymer informatics: Expert-in-the-loop in QSPR modeling of refractive index. *Computational Materials Science*, *194*, 110460.

[23] Palomba, D., Vazquez, G. E., & Díaz, M. F. (2012). 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.

[24] Cravero, F., Martínez, M. J., Ponzoni, I., & Diaz, M. F. (2019). Computational modelling of mechanical properties for new polymeric materials with high molecular weight. *Chemometrics and Intelligent Laboratory Systems*, *193*, 103851.

[25] Cravero, F., Díaz, M. F., & Ponzoni, I. (2022). Polymer informatics for QSPR prediction of tensile mechanical properties. Case study: Strength at break. *The Journal of Chemical Physics*, *156*(20), 204903.

[26] 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.* Aveiros, Portugal.