Modelos bioinformáticos para la predicción de compuestos biológicamente activos en cáncer de colon

Paula Carracedo Reboredo

Tesis doctoral UDC / 2023

Directores:
Dr. Carlos Fernández Lozano
Dra. Sonia Arrasate Gil

Programa oficial de doctorado en Tecnologías de la Información y las Comunicaciones

UNIVERSIDADE DA CORUÑA

Dr. Carlos Fernández Lozano, Profesor Titular de Universidad del Departamento de Ciencias de la Computación y Tecnologías de la Información de la Universidade da Coruña.

Dra. Sonia Arrasate Gil, Profesora agregada del Departamento de Química Orgánica e Inorgánica de la Universidad del País Vasco/Euskal Herriko Unibertsitatea UPV/EHU.

AUTORIZAN:

La presentación para su deposito de la tesis que dirigen y que fue realizada por Doña Paula Carracedo Reboredo con el número de identidad 33282674H con título "Modelos bioinformáticos para la predicción de compuestos biológicamente activos en cáncer de colon".

Y para que así conste, firma esta autorización en A Coruña, a 12 Junio 2023

Los directores de la tesis

Fdo. Carlos Fernández Lozano

Fdo. Sonia Arrasate Gil

A mis padres

Agradecimientos

A mis directores, la Dra. Sonia Arrasate Gil y el Dr. Carlos Fernández Lozano, quiero agradecer profundamente su tiempo, su dedicación, su paciencia e inestimable ayuda.

También quiero mostrar mi agradecimiento al Dr. Humberto González Díaz y al grupo OMS y sus responsables, por haberme permitido iniciar esta andadura.

A todas las personas que, estando fuera, me hicieron sentir como en casa.

A mis amigos Antón y Rocío, porque allá donde estén, estoy segura de que se alegrarán por mí.

Y mis padres, mis hermanas y a Ana, por darme siempre todo sin pedir nunca nada.

Y muy especialmente a mi hijo Mateo, por su alegría infinita.

Resumo

Para o descubrimento de fármacos é preciso atopar novos compostos con propiedades químicas específicas que permitan o tratamento de enfermidades. Nos últimos anos, o enfoque empregado nesta procura ten un compoñente importante nas ciencias da computación, co rápido auxe das técnicas de aprendizaxe automática. Cos obxectivos marcados pola Medicina de Precisión e cos novos retos xerados, cómpre establecer metodoloxías computacionais robustas, estandarizadas e reproducibles para acadar os obxectivos propostos.

Na actualidade, os modelos preditivos baseados en Machine Learning cobraron gran importancia no paso previo ós estudos preclínicos, conseguindo reducir drasticamente os custos e os tempos de investigación no descubrimento de novos fármacos. Os modelos quimioinformáticos poden predecir diferentes resultados (actividade, propiedade química, reactividade) en moléculas individuais ou sistemas moleculares complexos (síntese orgánica catalizada, reaccións, nanopartículas, etc.). Concretamente, a predición quimioinformática da enantioselectividade en sistemas complexos catalíticos é un obxetivo importante da síntese orgánica e na industria química.

A reacción de α-amidoalquilación enantioselectiva catalizada por ácidos de Brønsted é un proceso útil para a produción de novos catalizadores quirais (ferramentas) ou fármacos (produtos). A enantioselectividade é sensible a moitos factores, desde a natureza do sustrato e do catalizador ata as condicións experimentais (disolvente, temperatura, etc.). Polo tanto, as ferramentas computacionais capaces de predicir a enantioselectividade destas reaccións constitúen un instrumento valioso para o deseño racional de novos catalizadores e produtos quirais por parte dos químicos especializados en síntese orgánica de todo o mundo.

Elaborouse esta Tese de Doutoramento, unha investigación moi marcadamente multidisciplinar, na que se conxugan os esforzos de tres ramas da ciencia como son a medicina, a química e a informática. No que respecta á informática, base da Tese Doutoral, empregaranse técnicas de intelixencia artificial para avanzar creando novos modelos. As ferramentas bioinformáticas desenvoltas serven de apoio fundamental para os avances médicos na comprensión dunha enfermidade tan complexa e multifactorial como o cancro e tamén no avance químico na identificación de estruturas de

compostos que os fan comportarse activamente contra o cancro de colon. Todo este esforzo conxunto culmina no desenvolvemento de novos modelos de aprendizaxe automática baseados na intelixencia artificial para a loita contra o cancro de colon e na implantación da primeira biblioteca R para o cálculo de descritores moleculares, así como o primeiro servidor web público para o seu cálculo en liña, validación dos modelos e unha versión de escritorio do software para uso local.

Resumen

Para el descubrimiento de fármacos es necesario encontrar nuevos compuestos con propiedades químicas específicas que permitan el tratamiento de enfermedades. En los últimos años, el enfoque utilizado en esta búsqueda presenta un componente importante en las ciencias de la computación, con el aumento vertiginoso de las técnicas de aprendizaje automático. Con los objetivos marcados por la Medicina de Precisión y los nuevos retos generados, es necesario establecer metodologías computacionales robustas, estandarizadas y reproducibles para alcanzar los objetivos planteados.

Actualmente, los modelos predictivos basados en Machine Learning han cobrado gran importancia en el paso previo a los estudios preclínicos, logrando reducir drásticamente los costos y tiempos de investigación en el descubrimiento de nuevos fármacos. Los modelos de quimioinformática pueden predecir diferentes resultados (actividad, propiedad química, reactividad) en moléculas individuales o sistemas moleculares complejos (síntesis orgánica catalizada, reacciones, nanopartículas, etc.). Específicamente, la predicción quimioinformática de la enantioselectividad en sistemas complejos catalíticos es un objetivo importante de la síntesis orgánica y la industria química.

La reacción de α-amidoalquilación enantioselectiva catalizada por ácidos de Brønsted es un procedimiento útil para la producción de nuevos catalizadores quirales (herramientas) o fármacos (productos). La enantioselectividad es sensible a muchos factores, desde la naturaleza del sustrato, nucleófilo y del catalizador hasta las condiciones experimentales (disolvente, temperatura, etc.). Por lo tanto, las herramientas computacionales capaces de predecir la enantioselectividad de estas reacciones constituye un valioso instrumento para el diseño racional de nuevos catalizadores y productos quirales por parte de los químicos especializados en síntesis orgánica de todo el mundo.

Se ha elaborado esta Tesis Doctoral, de investigación muy marcadamente multidisciplinar, en la que se aúnan esfuerzos de tres ramas de la ciencia como son la médica, química e informática. En lo referente a la informática, base de la Tesis Doctoral, se partirá de las técnicas de inteligencia artificial para avanzar creando nuevos modelos.

Las herramientas bioinformáticas desarrolladas sirven como apoyo fundamental para el avance médico en la comprensión de una enfermedad tan compleja y multifactorial como el cáncer y también en el avance químico de la identificación de estructuras de compuestos que hagan a los mismos comportarse activamente contra el cáncer de colon. Todo este esfuerzo conjunto culmina con el desarrollo de nuevos modelos de aprendizaje automático basados en inteligencia artificial para la lucha contra el cáncer de colon y la implementación de la primera librería en R para el cálculo de descriptores moleculares, así como del primer servidor web público para su cálculo en línea, validación de los modelos y una versión de escritorio del software para su uso local.

Abstract

For drug discovery it is necessary to find new compounds with specific chemical properties that allow the treatment of diseases. In recent years, the approach used in this search has an important component in computer science, with the rapid rise of machine learning techniques. With the objectives set by Precision Medicine and the new challenges generated, it is necessary to establish robust, standardized and reproducible computational methodologies to achieve the proposed objectives.

Currently, predictive models based on Machine Learning have become very important in the step prior to preclinical studies, drastically reducing costs and research times in the discovery of new drugs. Chemoinformatics models can predict different outcomes (activity, chemical property, reactivity) in single molecules or complex molecular systems (catalysed organic synthesis, reactions, nanoparticles, etc.). Specifically, chemoinformatics prediction of enantioselectivity in complex catalytic systems is an important goal in organic synthesis and chemical industry.

The Brønsted acid-catalysed enantioselective α-amidoalkylation reaction is a useful procedure for the production of new chiral catalysts (tools) or drugs (products). Enantioselectivity is sensitive to many factors, from the nature of the substrate and catalyst to the experimental conditions (solvent, temperature, etc.). Therefore, computational tools capable of predicting the enantioselectivity of these reactions are a valuable tool for the rational design of new catalysts and chiral products by organic synthesis chemists worldwide.

This Doctoral Thesis has been elaborated, a very markedly multidisciplinary investigation, in which the efforts of three branches of science such as medicine, chemistry and computer science are combined. With regard to computing, the basis of the Doctoral Thesis, artificial intelligence techniques will be used to advance by creating new models. The bioinformatics tools developed serve as fundamental support for medical advances in the understanding of a disease as complex and multifactorial as cancer and also in the chemical advancement of the identification of structures of compounds that make them behave actively against colon cancer. . All this joint effort culminates in the development of new machine learning models based on artificial intelligence for the fight against colon cancer and the implementation of the first R

library for the calculation of molecular descriptors, as well as the first public web server for its calculation online, validation of the models and a desktop version of the software for local use.

Tabla de contenidos

1 Introducción

El propósito de esta introducción es realizar una breve descripción del problema tratado en esta tesis doctoral y del contexto en el que se enmarca. Se definen los objetivos perseguidos por esta investigación y la hipótesis de partida de la misma. Por último, se resume el contenido de cada uno de los capítulos en los que está estructurado el presente documento.

Contexto

El cáncer es una enfermedad causada por el crecimiento descontrolado de las células que, entre otras cosas, modifican su forma y tamaño. Este crecimiento desorganizado puede originarse porque nacen más células, porque las células existentes no se mueren o por ambos fenómenos a la vez. Esta situación presenta dos características fundamentales que la hace potencialmente peligrosa para el organismo, en primer lugar porque las células se reproducen sin responder a los mecanismos de regulación y control y, en segundo lugar, porque invaden regiones o zonas que corresponden a otras células.

Se habla de cáncer de colon o Cáncer Colorrectal (CCR) cuando el proceso descrito anteriormente se produce en las células del colon y recto, que conforman el segmento final del sistema gastrointestinal. En el 80-90 % de los casos, el proceso de carcinogénesis se inicia a partir de mutaciones sobre determinados genes de las células de la mucosa colónica, que dan lugar inicialmente a una lesión precursora llamada pólipo, que en el transcurso de 10 a 15 años puede evolucionar a CCR.

A pesar de los avances en la comprensión de la etiología, la biología y la genética el CCR sigue siendo el tercer cáncer más frecuentemente diagnosticado en todo el mundo y es la segunda causa principal de muerte por cáncer. Según las proyecciones demográficas, se espera que la carga mundial de cáncer colorrectal aumente en un 72 %, pasando de 1,8 millones de nuevos casos en 2018 a más de 3 millones en 2040 [1], con aumentos sustanciales previstos en los países con ingresos bajos y medios.

Aunque son muchos los recursos dedicados a lo largo de los últimos años y el conocimiento que se tiene sobre la patogenia y los posibles tratamientos de esta enfermedad es cada vez mayor, el elevado coste que supone cada avance hace necesaria la exploración de otras formas alternativas a los ensayos preclínicos y clínicos tradicionales. Los elevados costes económicos asociados a los mismos, la dificultad en la búsqueda de muestras representativas para subtipos determinados, los problemas éticos que conllevan y la lentitud en la obtención de los resultados, son algunos de los principales inconvenientes que hacen que muchos de los modelos farmacológicos teóricos se queden en el papel al no tener un cierto éxito garantizado.

Es por ello que, en las fases de investigación preclínica, el desarrollo de nuevas herramientas para el descubrimiento de fármacos es un objetivo de gran importancia para la química médica. La implementación de herramientas basadas en química computacional pueden ayudar a predecir la reactividad química y la actividad biológica de compuestos que ayudan a refinar su uso, reducir costos en términos de recursos humanos, materiales, tiempo, y sacrificio de animales de laboratorio.

El aprendizaje automático (ML del ingés *Machine Learning*) es una rama de la Inteligencia Artificial (IA) que actúa como un proceso de inducción de conocimiento. Su objetivo es crear programas capaces de generalizar comportamientos a partir de una información suministrada en forma de ejemplos. Además, el aprendizaje automático también se centra en el estudio de la complejidad computacional de los problemas, por lo que gran parte de la investigación realizada en ML está enfocada al diseño de soluciones factibles a esos problemas. El diseño de experimentos y la validación de los resultados obtenidos con ML son vitales en cualquier estudio de investigación. Además, se pueden utilizar diferentes enfoques para comparar diferentes resultados de diferentes métodos y poder escoger así el que más se aproxima a los objetivos buscados.

La aplicación de técnicas de ML está ganando importancia en el procesamiento de la información química y el modelado de los problemas de reactividad química. En este trabajo se ha desarrollado, por un lado, un modelo quimioinformático mediante el cálculo de descriptores moleculares de cadenas de Markov y algoritmos de ML para predecir la respuesta activa de compuestos biológicos en CCR. Por otro, se ha desarrollado un modelo que combina la teoría de la perturbación y algoritmos ML para predecir el rendimiento de una reacción dada de manera que, a partir de estructuras moleculares conocidas, se puede establecer el comportamiento

que tendrán otras sin tener que reproducirlas en el laboratorio, reduciendo muy considerablemente la necesidad de recursos.

La información recopilada sobre la patogenia del CCR y sus tratamientos es mucha y la buena utilización de la misma podría ser muy útil en el desarrolo de modelos de predicción de comportamientos celulares en donde éstos se pueden desarrollar utilizando aprendizaje automático para predecir la hidrofobicidad de la superficie celular después de las perturbaciones de todas las variables de entrada.

En la actualidad, es en las terapias dirigidas donde se centra la mayor investigación de creación de fármacos contra el cáncer. Son la piedra angular de la medicina de precisión, una forma de medicina que usa información de los genes y proteínas de una persona para prevenir, diagnosticar y tratar enfermedades. La predicción de las Interacciones Fármaco-Proteína (IFP) para las proteínas diana implicadas en determinadas vías, es un objetivo muy importante en la química médica de hoy en día. Se puede abordar este problema usando modelos de acoplamiento molecular o ML para una proteína específica, lo cual podría abrir nuevas vías en terapias para determinados tipos de cáncer, entre ellos el de colon.

Descripción del problema

A pesar de ser uno de los más estudiados, el CCR es uno de los tumores malignos con mayor morbimortalidad del mundo, tanto en hombres como en mujeres. Puede permanecer silente muchos años pero si se detecta de manera precoz tiene, de manera general, buen pronóstico. Los programas de cribado poblacional han contribuído de manera considerable a su detección temprana, pero cuando los síntomas aparecen, suele estar localmente avanzado y en este estado, el pronóstico empeora. La diseminación metastásica, las recaídas y resistencia terapéutica endombrecen el pronóstico. A día de hoy, el CCR detectado precozmente es totalmente curable mediante cirugía y el uso de quimioterapia posterior. Se sabe que estos tumores presentan una alta tasa de respuesta a estos tratamientos y tienen impacto en términos de supervivencia global y supervivencia libre de progresión. Sin embargo, las secuelas tanto físicas como psicológicas causadas por la toxicidad de las quimioteraias estándar y las colostomías a las que en muchas ocasiones obliga la cirugía, hacen necesaria la búsqueda de alternativas terapéuticas. Además, la tasa de recurrencia es alta, y la resistencia a los fármacos contra el cáncer aumenta la tasa de fracaso del tratamiento.

Las terapias dirigidas consisten en la aplicación de un tipo de fármaco que interfiere en la función de una molécula previamente identificada en la célula tumoral, pero no en las células integrantes del tejido sano. Estas dianas desempeñan un papel fundamental en el crecimiento anormal y/o la capacidad de disemización y metástasis del tumor y deberían de ser óptimas cuando se indiquen en un contexto molecular apropiado porque la potencial especificidad que aportan este tipo de fármacos conlleva una mayor eficiencia terapéutica, evitando la toxicidad innecesaria en aquellos pacientes que, presumiblemente, no tuvieran potencial beneficio del tratamieto.

La búsqueda de nuevos fármacos y terapias más efectivas es esencial para mejorar los resultados de tratamiento y reducir la mortalidad. Aunque en las últimas décadas se han logrado avances significativos, todavía hay casos en los que las terapias convencionales, como la cirugía, la quimioterapia y la radioterapia, no son suficientemente efectivas. Además, algunos tumores de colon pueden volverse resistentes a los tratamientos disponibles, lo que limita su eficacia a largo plazo.

Otro problema que presentan los tratamientos convencionales, como la quimioterapia, son los posibles efectos secundarios que provocan y que afectan de manera considerable a la calidad de vida de los pacientes.

Objetivos

Los principales objetivos de la presente tesis doctoral, se pueden resumir en los siguientes:

- Analizar las posibles interacciones de compuestos biológicamente activos frente al cáncer de colon a partir de la estructura del compuesto.
- Predecir así el resultado obtenido en ensayos clínicos.
- Identificar aquellos que mejor cumplen como potenciales dianas terapéuticas de la enfermedad.
- Implementar los mejores modelos de manera abierta en servidores web de altas prestaciones que permiten obtener nuevas predicciones sobre futuros compuestos no conocidos durante la generación del modelo.

Hipótesis

La investigación de fármacos contra el CCR busca desarrollar terapias más selectivas y específicas que puedan reducir los efectos secundarios y mejorar la tolerabilidad de los tratamientos.

Los modelos quimioinformáticos son un tipo específico de modelos utilizados en el descubrimiento y diseño de fármacos. Estos modelos establecen una relación cuantitativa entre la estructura química de un compuesto y su actividad biológica, permitiendo predecir la actividad de nuevos compuestos en función de su estructura molecular.

Estos modelos se basan en la premisa de que la actividad biológica de un compuesto está determinada por sus características estructurales y propiedades físico-químicas que se pueden representar mediante descriptores moleculares, que son variables numéricas que capturan información sobre la estructura y propiedades del compuesto, como su tamaño, forma, carga, solubilidad, etc.

Mediante la recopilación de datos experimentales de actividad biológica de un conjunto de compuestos, junto con los correspondientes descriptores moleculares más relevantes para la actividad biológica que se quiere predecir, se obtiene un conjunto de datos al que se pueden aplicar diversas técnicas de aprendizaje automático para desarrollar un modelo matemático que relacione los descriptores moleculares con la actividad biológica, realizando análisis estadísticos y validaciones cruzadas para evaluar la calidad y robustez del modelo. Finalmente, la validación del modelo se realiza utilizando conjuntos de datos diferentes a los utilizados en su construcción, para verificar su capacidad de generalización y predecir correctamente la actividad de nuevos compuestos.

Una vez validados se puede utilizar para predecir la actividad biológica de nuevos compuestos antes de su síntesis y evaluación experimental. Esto permite filtrar rápidamente compuestos y enfocar los recursos en aquellos con mayor probabilidad de ser activos.

La reacción de α-amidoalquilación enantioselectiva catalizada por ácidos de Brønsted es un procedimiento útil para la producción de nuevos fármacos. En este contexto, los Catalizadores de Ácido Fosfórico Quiral (CPA) son catalizadores versátiles para este

tipo de reacciones y la selección y diseño de nuevos catalizadores CPA para diferentes reacciones enantioselectivas tiene un doble interés porque se pueden obtener nuevos catalizadores CPA (herramientas) y fármacos o materiales quirales (productos). Sin embargo, este proceso es difícil y requiere mucho tiempo si se aborda desde una perspectiva experimental de ensayo y error. Las herramientas quimioinformáticas pueden ayudar a comprender el mecanismo de las reacciones y a diseñar nuevos catalizadores

Es por todo ello que se plantea la siguiente hipótesis de trabajo sobre la que se fundamenta la presente Tesis Doctoral:

El desarrollo de nuevos modelos quimioinformáticos para la predicción de compuestos biológicamente activos frente al cáncer de colon, utilizando diferentes aproximaciones para cuantificar la actividad mediante modelos de relación cuantitativa entre estructura y actividad y de síntesis enantioselectiva por reacción α-amidoalquidación catalizada por ácidos de Brønsted, mejorará los resultados obtenidos con aproximaciones convencionales.

Se espera que los nuevos modelos generados sean útiles en diversas áreas de la investigación farmacéutica, como el cribado virtual de bibliotecas de compuestos, el diseño de nuevos fármacos y la optimización de propiedades farmacocinéticas, teniendo en cuenta que son herramientas predictivas y sus resultados deben ser confirmados experimentalmente antes de su aplicación práctica.

Estructura de la tesis

La presente tesis se estructura en 7 capítulos. Se resume a continuación el contenido de cada uno de los ellos.

En el primero, la introducción, se realiza una breve descripción del problema tratado y del contexto en el que se enmarca. Se definen también los objetivos perseguidos por esta investigación y la hipótesis de partida de la misma.

El capítulo 2 indica cuáles son los fundamentos necesarios para comprender el contenido de la presente Tesis Doctoral. Dado que se trata un trabajo altamente multidisciplinar, con generación de reacciones en laboratorio, análisis y generación de datos a partir de las mismas, y desarrollo de modelos de machine learning de

predicción, se exponen los fundamentos biológicos de la patología a estudio y los fundamentos computacionales necesarios.

En el capítulo 3 se expone el estado de la cuestión. Se ha revisado para ello la literatura existente y se han analizado investigaciones previas, teorías y enfoques relacionados con la temática multidisciplinar de la tesis. Se presentan por ello los estudios más recientes y relevantes sobre la temática, proporcionando un contexto sólido que permite comprender la relevancia y originalidad de la investigación propuesta. Se ha dividido la sección en dos grandes puntos, primero se mostrarán diferentes trabajos de predicción de compuestos biológicamente activos y, seguidamente, se mostrarán las técnicas de machine learning utilizadas más habitualmente para la generación de las predicciones.

En el capítulo 4 se presentan las dos aproximaciones seguidas en el desarrollo de la tesis doctoral: modelos quimioinformáticos para el descubrimiento y diseño de fármacos y modelos de síntesis enantioselectiva por reacción de α-amidoalquilación catalizada por ácidos de Brønsted para la producción de nuevos fármacos.

En el capítulo 5 se exponen y se analizan los resultados obtenidos para los dos modelos propuestos, indicando primero los datos utilizados y su procedencia, para después comentar la experimentación realizada y finalmente presentar los resultados obtenidos.

En el capítulo 6 se presentan las conclusiones extraídas tras el desarrollo de la presente Tesis Doctoral.

Finalmente, se presentan los futuros desarrollos que podrían ser abordados a partir del trabajo de investigación realizado a lo largo de estos años.

2 Fundamentos

En este capítulo se indican cuáles son los fundamentos necesarios para comprender el contenido de la presente tesis doctoral. Dado que se trata un trabajo altamente multidisciplinar, con generación de reacciones en laboratorio, análisis y generación de datos a partir de las mismas, y desarrollo de modelos de machine learning, se exponen los fundamentos biológicos de la patología a estudio y los fundamentos computacionales necesarios.

Fundamentos biológicos

Para poder analizar en profundidad una patología multifactorial y compleja, como el cáncer, es necesario comprender los fundamentos biológicos de la enfermedad. En los siguientes apartados del presente capítulo se resumirá: epidemiología, factores de riesgo, prevención, etiopatogenia, diagnóstico, tratamiento y dianas terapeúticas del CCR. De acuerdo a estos pasos, será posible comprender el proceso de trabajo seguido en la presente tesis para generar modelos bioinformáticos para la predicción de compuestos biológicamente activos en CCR.

Epidemiología

Según datos del Observatorio Global del Cáncer (GLOBOCAN) en 2020 se produjeron 19,3 millones de nuevos casos y 10 millones de muertes por cáncer en todo el mundo, de los cuales el CCR contribuyó con aproximadamente 1,93 millones (10 %) de nuevas incidencias y 0,94 millones (9,4 %) de muertes. La incidencia y la mortalidad del CCR varían considerablemente entre regiones del mundo y se asocian a la situación socioeconómica de cada país. Según el Banco Mundial, los nuevos casos y las muertes son más notables en las zonas con niveles de renta más altos y menores en las zonas con niveles de renta más bajos[2].

En 2020, el CCR fue el cáncer más diagnosticado, de un total de 36 cánceres, entre los hombres en 18 de los 186 países de todo el mundo y entre las mujeres en 6 de los 185 países.

Tab. 2.1.: **Incidencia y defunciones de los cánceres más frecuentes en 2020** [3].

Tipo	NºCasos	% Total	Nº Defunciones	% Total
Mama	2.261,519	11,7	684,996	6,9
Pulmón	2,206,771	11,4	1,796,144	18
Próstata	1,414,259	7,3	375,304	3,8
Piel no melanoma	1,198,073	63,731	0,6	
Colon	1,148,515	6,0	576,858	5,8
Estómago	1,089,103	5,6	768,793	7,7
Hígado	905,677	4,7	830,180	8,3
Recto	732,210	3,8	339,022	3,4

Sin embargo, en 2018, el CCR era el más diagnosticado entre los hombres en 10 de los 185 países, y ningún país tenía el CCR como el cáncer más diagnosticado entre las mujeres [4]. Por lo tanto, la tasa de incidencia de CCR ha aumentado alrededor del 10 % en los últimos dos años en los hombres. Las mujeres predominaban en el 3,24 % de los países. El CCR es más frecuente entre los hombres que entre las mujeres y más de cuatro veces más frecuente en los países de ingresos altos que en los de ingresos bajos. Las muertes también fueron aproximadamente 2,5 veces mayores en los países de ingresos altos que en los de ingresos bajos. En la Figura 2.1 se muestra la incidencia de CCR en ambos sexos.

Entre los CCR, el cáncer de colon fue el predominante y representó el 59,5 % de los casos nuevos, el 61,9 % de las muertes. Por su parte, el cáncer de recto tuvo una incidencia del 37,9 % y una mortalidad del 36,3 % para ambos sexos. El cáncer de colon ocupa por sí solo el quinto lugar en cuanto a nuevos casos de cáncer y muertes en comparación con todos los cánceres. En cambio, el cáncer de recto es el octavo y el décimo cáncer por incidencia y mortalidad, respectivamente.

El cáncer se ha convertido en la enfermedad no transmisible con mayor mortalidad en el mundo, siendo la segunda causa de muerte tras las enfermedades cardiovasculares.

Según el informe de *Las Cifras del Cáncer en España*, editado por la Sociedad Española de Oncología Médica (SEOM) en colaboración con la Red Española de Registros de Cáncer (REDECAN), los tumores más frecuentemente diagnosticados en el mundo en el año 2020 fueron los de mama, que ocupa la primera posición, pulmón, colon y recto, próstata y estómago, todos ellos con más de un millón de casos. En concreto

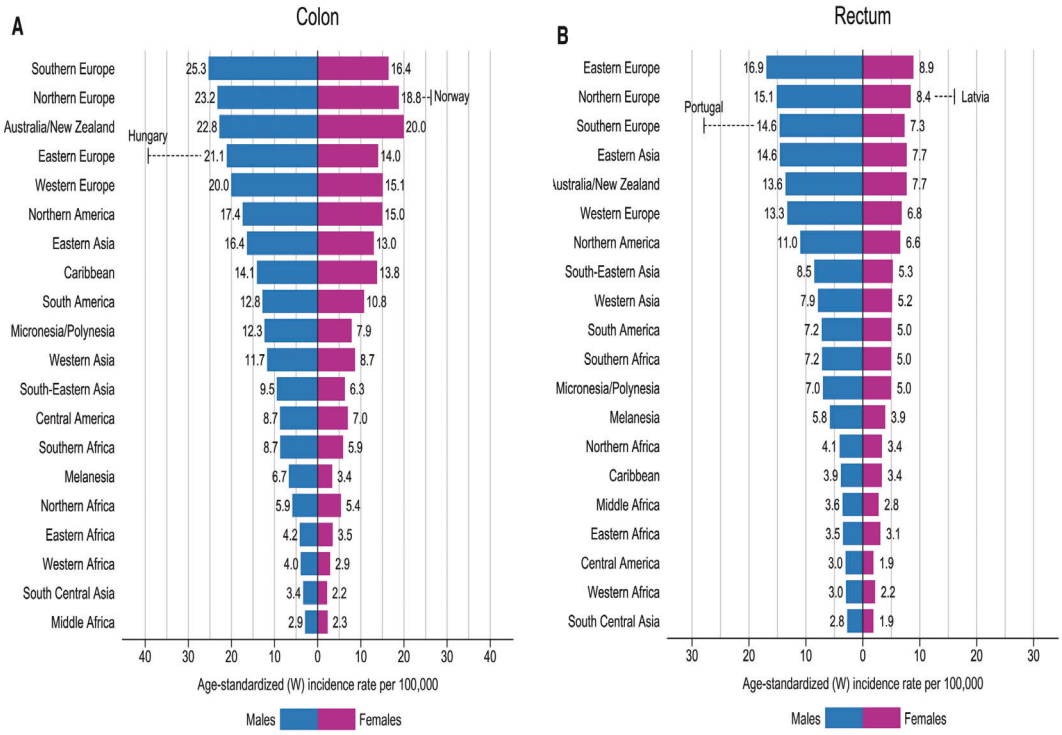

Fig. 2.1.: Incidencia mundial del CCR por sexos 2020. Imagen obtenida de [3].

el CCR fue diagnosticado a 1.931.590 personas, lo que supone más de un 10 % del total de cánceres diagnosticados.

Los casos esporádicos de CCR, sin antecedentes familiares ni alteraciones genéticas heredadas, representan el 60-65 % de todos los casos de CCR [5]. El inicio del CCR esporádico se atribuye en gran medida a mutaciones genéticas somáticas adquiridas o alteraciones epigenéticas inducidas por factores de riesgo modificables. El 35-40 % restante de los pacientes con CCR muestra susceptibilidad a componentes hereditarios [6]. Estos componentes incluyen antecedentes familiares sin predisposición genética evidente, síndromes de cáncer hereditario como el síndrome de Lynch, variaciones genéticas de baja penetrancia y otras aberraciones genéticas hereditarias desconocidas [7].

Cabe señalar que los factores ambientales contribuyen sustancialmente a la carcinogénesis en todos los casos de CCR, incluso en pacientes con antecedentes familiares, las alteraciones genéticas adquiridas pueden ser una causa importante del desarrollo del CCR.

Los programas de detección precoz de CCR sufrieron una disminución de su cobertura, sobre todo en los tres primeros meses de pandemia, aunque de forma general la tasa

de participación a estos programas se mantuvo igual. Hay que destacar que en el programa de cribado de cáncer colorrectal la caída de la cobertura fue mayor que la de otros, por ejemplo el de mama, debido fundamentalmente a las características propias del programa, que hicieron que fuera más difícil de retomar o que se hiciese más tarde. Todo esto sugiere que en los años posteriores a 2020 se pueda producir un aumento de los diagnósticos en estadios más avanzados a causa del retraso diagnóstico [8].

Por todo ello, dado que las estimaciones de la incidencia se realizan a partir de proyecciones realizadas con datos de años anteriores, el número de cánceres finalmente diagnosticado en 2020 fue menor al esperado y el número estimado de cánceres incidentes presentado en *las cifras del cáncer en España 2020* fue superior a lo que finalmente fue la realidad. Del mismo modo, no está claro cómo todo esto afectó al número de diagnósticos de cáncer de los años 2021 y 2022, y cómo afectará a 2023, aunque muy posiblemente el efecto ya será mucho menor.

Las cifras anteriores ponen de manifiesto que el CCR se ha convertido en un desafío de salud global con alta carga financiera por lo que la prevención resulta crucial para reducir su riesgo de CCR y así también su incidencia.

La mayoría de las neoplasias colorrectales se producen a partir de los 50 años, sin embargo, los adultos jóvenes y los adolescentes también pueden desarrollar CCR. Los hombres tienen una media de 68 años cuando se les diagnostica cáncer de colon, mientras que las mujeres tienen de media 72 años. En ambos sexos el diagnóstico de cáncer de recto se realiza de media a los 63 años de media [9].

Factores de Riesgo del CCR

El desarrollo del CCR está relacionado con factores de riesgo tanto no modificables como modificables. Entre los primeros, que no pueden ser controlados por los individuos se engloban, entre los más importantes, el sexo, edad, raza, antecedentes personales de pólipos adenomatosos y Enfermedad Inflamatoria Intestinal (EII), y los antecedentes familiares. En cuanto a los segundos, que pueden ser modificados por los individuos, se incluyen los relacionados con hábitos o estilos de vida individuales.

Antecedentes familiares

Son importantes factores de riesgo para padecer CCR la Poliposis Adenomatosa Familiar (PAF) y el síndrome de Lynch, también conocidos como Cáncer Colorrectal Hereditario No Polipósico (CCNPH), son las dos formas más comunes de cáncer de colon hereditario [10]. Ambas afecciones son el resultado de ciertas mutaciones en la línea germinal que contribuyen a aumentar el riesgo de CCR entre los miembros de la familia.

- La PAF es una enfermedad genética rara causada por la mutación del gen APC que inhibe la formación de tumores en el intestino. Cientos de pequeños pólipos adenomatosos recubren el intestino grueso y el recto, generalmente durante la adolescencia. Los pólipos persisten proliferando por todo el colon, con eventual transformación en malignos. El riesgo de desarrollar cáncer colorrectal en individuos con PAF no tratada es de casi el 100 % antes de los 40 años [11].

- El Síndrome de Lynch es la propensión hereditaria más frecuente al CCR y es una enfermedad hereditaria autosómica dominante. El síndrome de Lynch está causado por mutaciones de la línea germinal en uno de los genes de reparación de errores del ADN (MMR del inglés *Mismatch Repair*), incluidos MLH1, MSH2, MSH6 o PMS2, o la molécula de adhesión de células epiteliales (EpCAM del inglés *Epithelial Cellular Adhesion Molecule*). Para los portadores de mutaciones MMR de línea germinal, el riesgo de por vida de adquirir cáncer colorrectal se predice en más del 60 %. El cáncer colorrectal afecta a varias generaciones de una familia y se desarrolla pronto en la vida, con una edad media de diagnóstico de aproximadamente 45 años [11].

Enfermedad inflamatoria intestinal

Los pacientes con EII, que incluye las patologías colitis ulcerosa y enfermedad de Crohn, tienen un mayor riesgo de desarrollar cáncer colorrectal. La inflamación crónica es el motor de la progresión neoplásica, dando lugar a lesiones precursoras displásicas que pueden surgir en múltiples áreas del colon. El CCR asociado a colitis comparte muchas similitudes moleculares con el CCR esporádico, y las investigaciones preclínicas han demostrado un papel potencial del microbioma en concierto con el sistema inmunitario del huésped en el desarrollo del cáncer colorrectal asociado a colitis [12].

Diabetes mellitus tipo 2

La Diabetes Mellitus tipo 2 (DM2) también es un factor riesgo de desarrollar CCR. El efecto de la hiperinsulinemia en el colon es un mecanismo subyacente que podría explicar un mayor riesgo de cáncer de colon proximal. Los Factores de Crecimiento similares a la Insulina (IGF del inglés *Insulin-like Growth Factor*), que funcionan en el crecimiento y el desarrollo y se sobreexpresan en las células cancerosas junto con receptores IGF específicos, desempeñan un papel en el desarrollo de muchas neoplasias malignas. Los IGF lo hacen aumentando la progresión del ciclo celular y suprimiendo al mismo tiempo la apoptosis. La insulina promueve el cáncer estimulando los receptores de insulina y reduciendo las cantidades de proteínas de unión al IGF en el organismo [13].

Obesidad e inactividad física

La obesidad se asocia a un riesgo elevado de CCR [14]. El mecanismo subyacente es que la obesidad promueve la resistencia a la insulina o la hiperinsulinemia, la inflamación crónica, el estrés oxidativo, el daño del ADN y niveles elevados del factor de crecimiento similar a la insulina-1 (IGF-1), estimulando la proliferación celular [15]. La obesidad y la inactividad física aparecen fuertemente relacionados y existe evidencia que demuestra una relación entre la actividad física y los resultados positivos del cáncer, que incluyen menos fatiga, mejor calidad de vida y una supervivencia más prolongada. Entre los mecanismos investigados y propuestos para explicar este vínculo se encuentran la reducción de la grasa corporal total y visceral, la disregulación metabólica, la inflamación crónica, el estrés oxidativo y la función inmunológica mejorada asociada con la participación en la actividad física [16].

Alcohol

El consumo regular de alcohol, tanto semanal como diario, se asocia significativamente con un mayor riesgo de CCR [16]. El metabolismo del alcohol implica la conversión del etanol en sus metabolitos, que pueden causar efectos cancerígenos en el colon. La producción de metabolitos del etanol puede verse influida por la microbiota del colon, otro factor mediador recientemente reconocido de la carcinogénesis del colon. El desarrollo de aductos de ADN, el estrés oxidativo y la peroxidación lipídica, los cambios epigenéticos, la disfunción de la barrera epitelial y los efectos moduladores

inmunológicos están relacionados con la producción de acetaldehído y otros metabolitos del alcohol. Además, los alcohólicos son susceptibles a una dieta deficiente en folato y fibra, así como a alteraciones circadianas, que pueden exacerbar el cáncer de colon inducido por el alcohol [17].

Tabaco

Es un factor de riesgo modificable para desarrollar CCR, y el riesgo de CCR aumenta con el incremento del número de cigarrillos consumidos [17]. Uno de los carcinógenos presentes en los cigarrillos es la nicotina. La nicotina ha aumentado la proliferación alterando la expresión de los receptores y los patrones de fosforilación en varias vías mitogénicas. La exposición a la nicotina provoca una mayor fosforilación del receptor del factor de crecimiento epidérmico (EGFR del inglés *Epidermal Growth Factor Receptor*) y un aumento de la expresión de 5-LOX (Lipoxigenasa) en el cáncer de colon. Del mismo modo, la nicotina aumenta el crecimiento de las células de cáncer colorrectal mediante la regulación al alza de los receptores de acetilcolina y noradrenalina. La nicotina también estimula la angiogénesis y la neovascularización en el cáncer de colon a través de aumentos del factor de crecimiento del endotelio vascular (VEGF del inglés *Vascular Endothelial Growth Factor*), 5-LOX, COX-2 [18].

Dieta

El consumo de carne roja y procesada, aumenta el riesgo de CCR en un 20-30 %, pero las carnes blancas no se asocian con el riesgo de CCR[19]. Por otro lado, el consumo de patrones dietéticos ricos en fibra, como verduras, frutas, cereales integrales y cereales, se asocia a una baja incidencia de CCR y además son recomendables para su prevención [20]. Se desconoce el mecanismo que subyace a la asociación entre el consumo excesivo de carne roja y procesada y el riesgo de CCR. Sin embargo, los posibles mecanismos subyacentes de esta asociación incluyen sustancias carcinógenas formadas durante el proceso de cocinado que pueden aumentar la exposición a compuestos *N*-nitroso, aminas heterocíclicas e hidrocarburos aromáticos policíclicos. Las moléculas de hierro hemo de la carne roja pueden producir carcinógenos y actuar como mutágenos del ADN. Los PUFAs, los ácidos biliares, el ácido siálico no humano y los patógenos infecciosos son también potenciales impulsores mecánicos[21].

Fármacos

La Terapia Hormonal Sustitutiva (THS), que se indica en la menopausia, se asocia con un menor riesgo de CCR en mujeres, mortalidad relacionada con el y mortalidad por todas las causas [22]. El consumo regular de aspirina y otros Antiinflamatorios No Esteroideos (AINE), se asocian como factores protectores del CCR. El uso de aspirina y AINE a largo plazo conlleva el riesgo de sufrir una hemorragia gastrointestinal importante o un infarto de miocardio debido a los inhibidores selectivos de la COX-2. Debido a estos efectos secundarios, no se recomiendan en la población general y solo se recomienda el uso diario de dosis bajas de aspirina para prevenir las enfermedades cardiovasculares y el CCR en determinadas personas, a partir de 50 años con alto riesgo de enfermedad cardiovascular [23].

Prevención

La prevención primaria del cáncer de colon implica adoptar medidas para reducir el riesgo de desarrollar la enfermedad. A continuación, se mencionan algunas estrategias que pueden ayudar en la prevención del cáncer de colon[24, 5, 25, 26]:

- **Dieta saludable**: Seguir una dieta rica en frutas, verduras, granos enteros y fibra, y baja en carnes rojas procesadas y alimentos altos en grasas saturadas puede ayudar a reducir el riesgo de cáncer de colon. Además, limitar el consumo de alcohol y evitar el tabaquismo también puede ser beneficioso.

- **Actividad física regular**: Mantener un estilo de vida activo y realizar actividad física regular puede disminuir el riesgo de cáncer de colon. Se recomienda hacer al menos 150 minutos de actividad física moderada o 75 minutos de actividad física vigorosa por semana.

- **Mantener un peso saludable**: El sobrepeso y la obesidad están asociados con un mayor riesgo de cáncer de colon. Mantener un peso saludable a través de una alimentación equilibrada y actividad física regular puede contribuir a la prevención.

- **Screening o cribado**: Participar en programas de detección temprana de cáncer de colon, como la colonoscopia, sigmoidoscopia flexible, pruebas de sangre oculta en heces o pruebas de ADN en heces, puede ayudar a identificar pólipos o cáncer en etapas tempranas, cuando son más tratables.

- **Reducción de factores de riesgo**: Evitar la exposición innecesaria a factores de riesgo conocidos, como la radiación y ciertos productos químicos industriales, puede contribuir a la prevención del cáncer de colon.

- **Medicamentos preventivos**: En algunos casos, el médico puede recomendar medicamentos como la aspirina o los inhibidores de la Cox-2 para reducir el riesgo de cáncer de colon en personas con alto riesgo.

Es importante tener en cuenta que la prevención primaria del cáncer de colon no ofrece una protección absoluta, pero puede reducir significativamente el riesgo de desarrollar la enfermedad.

En cuanto a la prevención secundaria, el cribado se considera el método más eficaz para prevenir el desarrollo de CCR, ya que aumenta la detección temprana y permite la extirpación de lesioines precancerosas[25, 26]. Se incluyen aquí diversas pruebas, como los análisis de heces, que es una opción de detección no invasiva y que es ampliamente utilizado en los programas de cribado poblacional, pues no se requiere preparación especial [27]. Si la prueba es positiva, los métodos endoscópicos invasivos, como la colonoscopia, la colonografía por TC o la sigmoidoscopia, se realizarán para confirmar los resultados anormales. La Asociación Española contra el Cáncer (AECC) recomienda el cribado regular para personas mayores de 45 años, y debe ser iniciado antes en personas con alto riesgo de CCR debido a úlceras, colitis, adenomas previos o antecedentes familiares[28]].

Etiopatogenia

La patogenia del cáncer de colon es un proceso complejo y multifactorial que se desarrolla gradualmente en las células epiteliales de la mucosa intestinal. La duración de esta transición se estima entre los 10 y 15 años , a lo largo de los cuales se produce la acumulación secuencial de alteraciones genéticas[29]. Los principales factores que predisponen a padecer esta patología son:

- **Factores genéticos y hereditarios**: algunos casos de cáncer de colon tienen una predisposición hereditaria debido a mutaciones genéticas en las líneas germinales, que aumentan el riesgo de desarrollar cáncer de colon.

- **Lesiones precursoras**: la mayoría de los casos de cáncer de colon se desarrollan a partir de lesiones precursoras conocidas como pólipos, que son crecimien-

tos anormales en el revestimiento del colon que pueden sufrir trasformaciones malignas con el tiempo.

- **Alteraciones genéticas**: se trata de mutaciones en genes específicos, como los genes APC, KRAS y p53, que son comunes en el cáncer de colon y desempeñan un papel importante en la progresión del tumor.

- **Inflamación crónica**: la presencia de enfermedades inflamatorias y crónicas del intestino, como la enfermedad de Crohn o la colitis ulcerosa, aumenta el riesgo de desarrollar cáncer de colon, ya que la inflamación perpetuada en el tiempo, puede causar daño en el ADN y promover la proliferación celular anormal.

- **Factores ambientales y estilo de vida**: existen varios factores externos que pueden influir en el desarrollo del cáncer de colon, como una dieta rica en grasas saturadas y carnes rojas, falta de actividad física, obesidad, tabaquismo y consumo excesivo de alcohol. Estos factores pueden aumentar la inflamación, el estrés oxidativo y el daño en el ADN, lo que contribuye al desarrollo del CCR.

Es importante tener en cuenta que estos factores no son mutuamente excluyentes y que el cáncer de colon puede resultar de la interacción de múltiples factores genéticos, ambientales y de estilo de vida, siendo la patogenia exacta variable de un individuo a otro. Así pues, el CCR es una enfermedad compleja y no se puede definir simplemente por la acumulación de alteraciones genéticas ni tampoco por la mutación de un gen específico que genera el desarrollo de la enfermedad. Además de estas alteraciones moleculares es necesario un ambiente favorable para su desarrollo, que esté vinculado a un estado adecuado para la proliferación de mediadores inflamatorios inmunológicos, desarrollo de nueva vascularización y otros muchos factores que tienen que ser propicios para que se desarrolle una célula tumoral en este ambiente [30].

La mayoría de los casos de cáncer colorrectal son esporádicos, no relacionados con predisposición genética o antecedentes familiares. Sin embargo, del 20 al 30 % de los pacientes con cáncer colorrectal tienen antecedentes familiares de cáncer colorrectal y 5 % de estos tumores surgen en el contexto de un síndrome de herencia mendeliana [31].

En muchos pacientes, el desarrollo de un cáncer colorrectal está precedido por una lesión neoplásica benigna que se genera por el crecimiento anormal sobre la mucosa

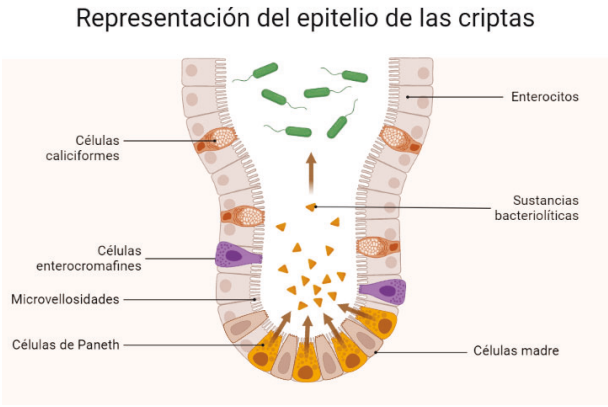

Fig. 2.2.: Representación del epitelio de las criptas. (Imagen generada en herramienta BioRender.com).

del colon que crece hacia su luz. Para entender mejor el proceso de transición de pólipo a CCR se debe conocer la histología del colon, esquematizada en la imagen 2.2

El colon está revestido por una capa única de células epiteliales columnares que se organizan en invaginaciones llamadas criptas. Cada una de estas criptas se divide en dos zonas funcionalmente distintas. Una proliferativa, que se compone de células madre que producen otras células indiferenciadas, que a su vez se dividen varias veces y migran hasta la zona de las criptas donde se diferencian dando lugar a las cuatro líneas celulares que coexisten en estas criptas:

- Células de absorción, fundamentalmente enterocitos.
- Células de Paneth, cuya función principal es la secreción de sustancias bacteriolíticas, tales como lisozimas, fosfolipasa A y defensinas.
- Células caliciformes, secretoras de moco.
- Células enteroendocrinas, producen y segregan varios polipéptidos.

Todas estas células son clones, ya que derivan todas de una misma célula madre, y se someten a la muerte celular programada o apoptosis. La vida de estas células, salvo las de Paneth, es de 4 ó 5 días. Debido al crecimiento reactivo de estas células aparecen pólipos intestinales, que son crecimientos de mucosa intestinal hacia la luz y por definición visible a simple vista. Los polipos pueden ser de diferentes tipos (Figura 2.3).

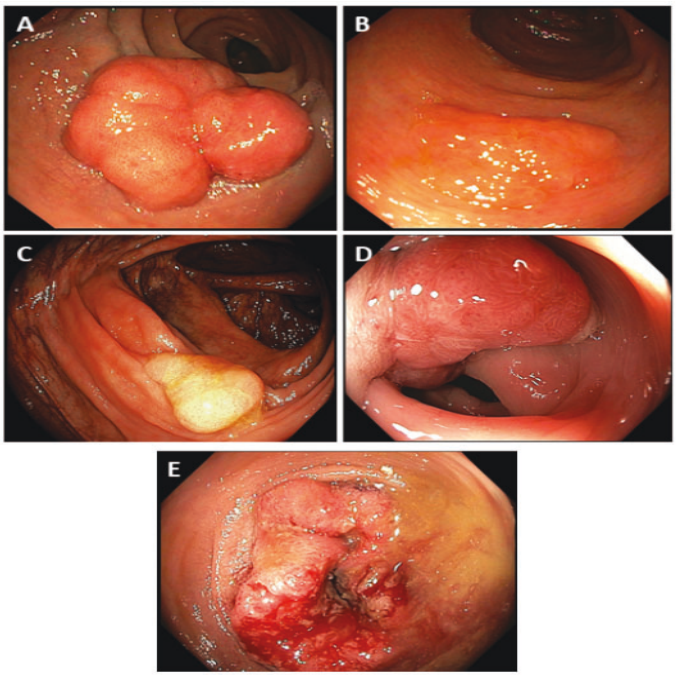

Fig. 2.3.: Imagen endoscópica de diversos tipos de lesiones precursoras de CCR: adenomas tubulares, túbulo-vellos, serrados, etc. Este tipo de imágenes se encuentran disponibles en fuentes como [32].

Polipos adenomatosos. Los pólipos adenomatosos son muy frecuentes y presentan gran potencial de malignización. De todos ellos, el adenoma velloso es el de mayor potencial maligno, pero todos los tipos histológicos presentan componente velloso en mayor o menor proporción, por lo que todos se pueden considerar premalignos. El examen histológico los clasifica según su arquitectura glandular en:

- **Tubulares**: son los que mayormente se detectan en las piezas de biopsia, constituyendo el 85 % de los pólipos adenomatosos. Contienen menos de un 20 % de componente velloso.

- **Túbulo - vellosos**: representan el 10 %.

- **Vellosos**: presentan más de un 80 % de componente velloso y son los menos fercuentes, representando el 5 % de los pólipos adenomatosos.

Polipos hiperplásicos y/o serrados. Son aquellos con arquitectura en dientes de sierra. Un tamaño mayor de 1 cm y la localización proximal son los determinantes de su malignificación. Se dividen en:

- **Pópilo hiperplásico**: suponen el 80-90 % de las lesiones serradas. Son difíciles de diferenciar por su color parecido a la mucosa sana.

- **Adenoma o pólipo serrado sésil**: supone el 15-20 % de los pólipos serrados, y se considera la lesión preneoplásica clave en la vía serrada de la carcinogénesis. Su localización en colon proximal y sus características, hacen que este tipo de pólipos pasen frecuentemente desapercibidos.

- **Adenoma serrado tradicional**: es poco frecuente (1-6 %).

Pólipos inflamatorios. Secundarios a un proceso regenerativo de un foco inflamatorio sin potencial de degeneración neoplásica.

Pólipos hamartomatosos. Son consecuencia de la proliferación de células maduras de la mucosa.

Con base en estudios histológicos y epidemiológicos realizados en la década de 1970, se reconoció que la mayoría de los carcinomas colorrectales probablemente se originaban a partir de pólipos adenomatosos premalignos. Fearon y Vogelstein propusieron el modelo original de varios pasos para la carcinogénesis colorrectal en 1988 [33]. Así pues, durante décadas, el proceso degenerativo conocido como secuencia adenoma - carcinoma explicó la gran mayoría de los cánceres colorrectales. A día de hoy, y gracias a los avances tecnológicos, se conoce que no todos los CCR presentan la mencionada secuencia y en torno al 10 - 15 % están incluidos dentro de la denominada vía serrada, cuya principal característica es la inestabilidad de microsatélites que se define por la inactivación de los genes reparadores del ADN debida a cambios epigenéticos.

Factores genéticos. A nivel molecular destacan dos grupos de genes específicos que están vinculados básicamente al control del ciclo celular, que son los que clásicamente se han estudiado, y están principalmente alterados en esta patología[34]. Dentro de ellos se encuentran:

- Genes supresores de tumores, cuya función se encuentra vinculada al control del cliclo celular como respuesta a alteraciones genéticas. En el CCR este grupo de genes se caracterizan por presentar una pérdida de función.

- Oncogenes, que también participan en la regulación del ciclo celular fomentando el crecimiento y la proliferación celular. Cuando este grupo de genes

se encuentra alterado, se pierde la regulación de la proliferación celular y se caracteriza por presentar una ganancia de su función.

Así pues, la ganancia de mutaciones acumulativas en ciertos oncogenes y la pérdida de función de genes supresores de tumores conducen a la alteración del epitelio colónico. Esta transición puede producirse a través de varias vías [35, 36, 37], cuyo conocimiento es fundamental para mejorar la detección temprana, el diagnóstico y el desarrollo de enfoques terapéuticos más efectivos. Algunas de las principales vías son las siguientes:

- **Vía de la inestabilidad cromosómica**: también conocida como la vía CIN (del inglés Cromosomal Instability Pathway) es una de las principales vías moleculares implicadas en el desarrollo del CCR[36]. Se caracteriza por una alteración cromosómica a nivel estructural, de manera que a una mutación de uno de los alelos se suma una segunda mutación que genera la inestabilidad.

- **Vía de los microsatélites**: clásicamente estudiada por el CCNPH o Síndrome de Lynch[38, 39], se caracteriza por mutaciones genéticas heredadas en genes de reparación del ADN, específicamente en los genes MLH1, MSH2, MSH6, PMS2. Estos genes son responsables de corregir los errores en el ADN durante la división celular, y mutaciones en ellos pueden llevar a una acumulación de errores en el ADN, lo que aumenta el riesgo de desarrollo de CCR.

- **Vía del adenoma serrado**: recientemente se ha descubierto que está asociada con mutaciones en genes como BRAF y KRAS, que están implicados en la regulación del crecimiento celular y la supervivencia [40]. Las mutaciones en el gen BRAF, en particular, se consideran un marcador molecular importante de la vía del adenoma serrado [41].

Genes implicados en la patogenia del CCR.

Se han identificado varios genes que desempeñan un papel importante en el desarrollo y progresión del cáncer de colon. Algunos de los genes más destacados implicados en el cáncer de colon son:

- Gen de la poliposis adenomatosa colónica (APC del inglés Adenomatous Polyposis Coli): se encuentra en el locus 5q 21q 22, forma parte de los genes supresores de tumores y en los estadíos precoces de la enfermedad se encuentra

alterado en la mayoría de los casos, regulando de manera negativa las señales que promueven el crecimiento y dando lugar a la proliferación del epitelio normal hacia un adenoma. Se encuentra mutado hasta en el 80 % de los pólipos esporádicos y los adenocarcinomas [42]. La línea germinal de mutación del gen APC es también responsable de la formación de múltiples adenomas en la poliposis adenomatosa familiar (PAF) [43].

- Protooncogén KRAS: su alteración provoca una ganancia de función que promueve la alteración en la proliferación celular más allá de las señales exógenas de crecimiento [44]. Se establece que se necesita la mutación previa del gen APC para que se de esta alteración [45], la cual promueve la tumorogénesis al causar la hiperproliferación de células colorrectales, tanto en el estadío temprano del adenoma como posteriormente en la transformación maligna a carcinoma.

- Gen DCC (del inglés Delete in Colon Cancer), situado en el cromosoma 18q, participa en la progresión, invasión y metástasis del tumor.

- Gen p53: localizado en el cromosoma 17p, pertecece a los genes supresores de tumores. Parece estar involucrado en la transformación maligna que supone la conversión de adenoma a adenocarcinoma que se produce en fases tardías. En condiciones normales este gen, ante alteraciones celulares, promueve su reparación conduciendo a la célula a la apoptosis. Por lo tanto, si pierde su función, las células dañadas no van a ser reparadas ni morirán. La pérdida de este gen es infrecuente en adenomas pero aparece en más del 75 % de adenocarcinomas [46].

En la imagen 2.4 se muestra la secuencia de mutaciones implicadas en la vía CIN. La inactivación inicial del gen APC, que conduce al desarrollo de adenomas a partir de la mucosa normal, las mutaciones posteriores en KRAS y las alteraciones genéticas de los genes en 18q dan como resultado el crecimiento y la progresión del adenoma y, finalmente, la transición del carcinoma del adenoma estuvo mediada por la pérdida bialélica o la inactivación del gen p53.

Se ha descubierto que la vía del adenoma serrado se asocia con mutaciones genéticas específicas, como las mutaciones en el gen BRAF, que es un oncogén involucrado en la regulación de las vías de señalización celular. También se ha observado una mayor frecuencia de metilación del promotor del gen MLH1, que afecta a los mecanismos de reparación del ADN.

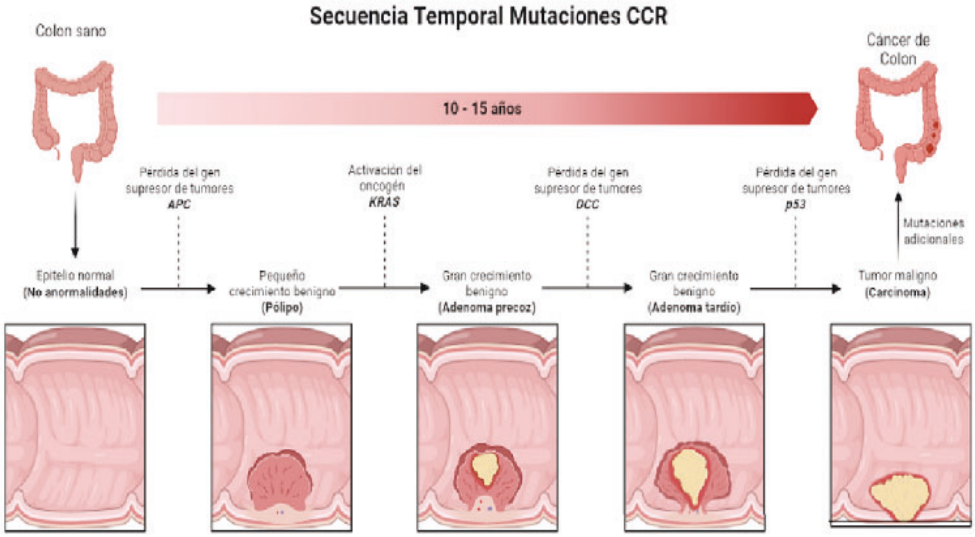

Fig. 2.4.: Secuencia temporal de mutaciones de la vía CIN. (Imagen generada en BioRender.com).

- Mutaciones en el gen BRAF: el gen BRAF codifica una proteína que forma parte de la vía de señalización del EGFR [47]. En la vía del adenoma serrado, se han identificado mutaciones en el gen BRAF en una proporción significativa de los adenomas serrados y cánceres colorrectales asociados. Estas mutaciones están asociadas con un mal pronóstico y una mayor progresión hacia el CCR [48].

- Metilación del promotor del gen MLH1: en algunos casos de adenomas serrados y cánceres colorrectales asociados, se ha observado la metilación del promotor del gen MLH1. Esta metilación epigenética puede llevar a la inactivación del gen MLH1 [49], que está involucrado en la reparación del ADN, lo que contribuye a la progresión hacia el CCR.

Sin embargo, los avances recientes en las tecnologías de secuenciación molecular, que permiten la caracterización de fenotipos de pacientes con cáncer a gran escala, han destacado los cambios epigenéticos como un sello distintivo del cáncer. Las modificaciones epigenéticas incluyen la hipermetilación y hipometilación del ADN, el silenciamiento transcripcional de genes supresores de tumores, las alteraciones en genes involucrados en el control del ciclo celular y en la reparación del ADN (genes Mismatch) [50] y la modificación de las histonas, segundo mecanismo genético más común relacionado con el CCR esporádico. Se ha encontrado que estos factores juegan un papel clave en la patogénesis de una amplia variedad de cánceres a través de la regulación de estado de la cromatina, expresión génica y otros eventos nucleares [50], que ocurren en estadios tempranos de la carcinogénesis colorrectal.

Dada la naturaleza reversible de los cambios epigenéticos, la posible modificación de los mismos se ha instaurado como una potencial terapia contra el cáncer. La comprensión de la vía del adenoma serrado es importante para la detección temprana, el diagnóstico y el manejo adecuado de los pacientes con pólipos colorrectales. Sin embargo, es necesario realizar estudios adicionales para profundizar en los mecanismos moleculares y las características clínicas asociadas con esta vía.

Diagnóstico

Existen diversas pruebas utilizadas para el diagnóstico de CCR, pero no todas se realizan en todos los pacientes, siendo factores para su elección la sospecha clínica, la edad y el estado general, entre otros. Además, la evaluación de los antecedentes de cáncer personales y familiares de primer segundo y tercer grado es fundamental para obtener información detallada en el proceso de diagnóstico [7].

El diagnóstico definitivo necesita confirmación mediante el estudio histológico de muestra de biopsia, que si el tumor se ha diseminado, puede realizarse de alguna lesión metastásica. Una vez que se confirma el CCR, son necesarios estudios de imágenes que evalúen el estadio local, la presencia de ganglios linfáticos agrandados y metástasis a distancia y el riesgo de obstrucción.

De manera general, el diagnóstico y la extensión del CCR se realiza en la práctica clínica mediante la combinación de las siguientes pruebas:

- Colonoscopia: permite la observación directa del interior de todo el recto y el colon, así como la toma de biopsias, pero no describe con precisión la localización y la diseminación del tumor.

- Biopsia: única prueba que permite dar un diagnóstico definitivo del cáncer colorrectal.

- Tomografía computarizada y Resonancia magnética: utilizadas para medir el tamaño del tumor y detectar la diseminación del mismo a otros órganos.

- Ecografía endorrectal: se utiliza frecuentemente para determinar a qué profundidad se ha extendido el cáncer en recto y es útil como ayuda para planificar el tratamiento.

- Tomografía por emisión de positrones: es una forma de crear imágenes de los órganos y los tejidos internos del cuerpo. No se usan regularmente para todas las personas con cáncer colorrectal, pero existen situaciones específicas en las que se recomienda su estudio.

Además, se debe estudiar la presencia o ausencia de biomarcadores genéticos específicos que permitan introducir tratamientos individualizados, cuya eficacia puede ser superior a la de los estándar. En los últimos años, se ha trabajado intensamente en la identificación de nuevos biomarcadores para el diagnóstico no invasivo del CCR. Actualmente existen algunos que permiten predecir el riesgo de invasión, aparición de metástasis o resistencia a regímenes terapéuticos específicos [51] y pueden traducirse con éxito a la práctica clínica [52].

Hay una gran cantidad de candidatos a biomarcadores de diagnóstico, dependiendo de sus tipos. Así, los genes con alteraciones en la metilación pueden afectar a la reparación del ADN, apoptosis y migración celular y respaldan la predicción de resultados clínicos, como el tratamiento y el pronóstico de supervivencia, la aparición de metástasis y la resistencia a la terapia [53]. Los biomarcadores de proteínas permiten detectar CCR[52, 54], estadificar el cáncer existente y predecir la respuesta a un tratamiento específico [52]. La detección de ADN tumoral circulante de células cancerosas en fluidos corporales se puede aplicar al diagnóstico y permiten determinar el tipo y grado de tumor, la reaparición del pronóstico y la respuesta al tratamiento [54]. Los microARN pueden regular negativamente la expresión génica y pueden constituir un potencial biomarcador, posibilitando diagnósticos más tempranos y un abordaje más personalizado. Su papel en la práctica clínica podría estar asociado también en el estadio temprano, el pronóstico, crecimiento tumoral o riesgo predictivo de adenomas de alto riesgo, recurrencia, etc. Los miARN circulantes como nuevos biomarcadores siguen teniendo varios desafíos que superar antes de su aplicación clínica. Se necesita una mayor investigación sobre el origen y la función biológica de los miARN. Además, se requiere una explicación de los mecanismos a través de los cuales los miARN podrían estar involucrados en la resistencia a la quimioterapia y otras terapias dirigidas [55].

La evaluación de diferentes biomarcadores relacionados con la microbiota puede ser una herramienta no invasiva útil para prevenir, diagnosticar e incluso tratar el CCR [56], ya que los productos metabólicos de los microbios intestinales pueden desencadenar una respuesta inflamatoria, producir especies reactivas de oxígeno, toxinas o mediadores, que pueden causar daño en el ADN y disfunción inducida o daño en las células epiteliales [57].

Tratamiento

Gracias a los avances, tanto en los tratamientos primarios como en los adyuvantes, el tiempo de supervivencia en el CCR ha mejorado en los últimos años. De manera general, el objetivo ideal es lograr la extirpación completa del tumor y las metástasis, lo que en su mayoría requiere intervención quirúrgica [58]. Sin embargo, a pesar de la aparición de numerosos programas de detección para reducir la incidencia del CCR, casi una cuarta parte de los mismos se diagnostican en una etapa avanzada con metástasis, y el 20 % de los casos restantes pueden desarrollar metástasis metacrónicas, por lo tanto, el control quirúrgico curativo por sí solo no es suficiente [5].

Durante muchos años, en los pacientes con CCR metastásico (CCRm) la cirugía y la quimioterapia fueron las primeras líneas de tratamiento contra la enfermedad, pero los individuos con enfermedad diseminada han tenido, de manera general, mal pronóstico. Para aquellos pacientes con lesiones irresecables o que no toleran la cirugía, el objetivo es la reducción máxima del tumor y la supresión de su crecimiento y diseminación, siendo la radioterapia y la quimioterapia las principales estrategias para controlar la enfermedad, pudiéndose administrar antes o después de la cirugía como tratamiento neoadyuvante.

Por lo tanto, el principal tratamiento para el CCR metastásico no resecable es la terapia sistémica, la cual incluye quimioterapia citotóxica, terapia biológica, inmunoterapia y sus combinaciones. Los ensayos clínicos completados en los últimos años han demostrado que adaptar el tratamiento a las características moleculares y patológicas del tumor mejora la supervivencia global [59]. Es por ello muy importante determinar el perfil genómico para así poder detectar variantes somáticas susceptibles de tratamientos efectivos.

Los detalles genómicos, transcripcionales, proteómicos y epigenéticos nunca han sido tan accesibles como en los últimos años, debido a la evolución de las tecnologías de secuenciación. Las alteraciones en la diferenciación, proliferación y supervivencia celular resultantes de cambios en el perfil genético contribuyen al inicio y desarrollo del cáncer. Sobre la base de la identificación de estas heterogeneidades, los tratamientos dirigidos a enzimas específicas, receptores de factores de crecimiento y transductores de señales hacen posible una terapia personalizada contra el cáncer, de modo que muchos procesos celulares oncogénicos pueden interferirse de manera eficiente, lo que promete una erradicación precisa del cáncer [60]. En cualquier caso,

hasta la fecha, no existe un régimen universal que pueda tratar fácilmente a todos los pacientes con la misma eficacia, y el conocimiento actual sobre el CCR también ha ido avanzando, lo que ha dado lugar a la identificación de nuevos objetivos.

Dianas terapéuticas del CCR

Existen diversas vías que median en la iniciación, progresión y migración del CCR y todas ellas son potenciales dianas en las que se pueden interferir con éxito para detener la progresión del tumor. Algunas de las dianas utilizadas en el cáncer de colon incluyen:

- Inhibidores EGFR: esta proteína juega un papel muy importante en el crecimiento y la proliferación celular y por ello es un objetivo terapéutico importante en el CCR. Los fármacos utilizados aquí son anticuerpos monoclonales que se dirigen al EGFR y pueden bloquear su actividad, lo que dificulta el crecimiento y la supervivencia de las células cancerosas.

- Inhibidores de la vía de señalización de RAS: esta vía juega un papel crítico en el crecimiento y la proliferación celular. Mutaciones en los genes RAS (como KRAS y NRAS) son frecuentes en el CCR y pueden conducir a la activación constante de la vía de señalización de RAS.

- Inhibidores de la vía de señalización de PI3K/AKT/mTOR: es una vía intracelular importante implicada en la supervivencia y el crecimiento celular. Fármacos que inhiben componentes de esta vía, pueden interferir con la angiogénesis (formación de nuevos vasos sanguíneos) y reducir el suministro de nutrientes al tumor, inhibiendo así su crecimiento.

- Inhibidores de angiogénesis: el proceso de formación de nuevos vasos sanguíneos desempeña un papel crucial en el crecimiento y la propagación de los tumores. Fármacos dirigidos contra el receptor VEGF pueden inhibir la formación de nuevos vasos sanguíneos en el tumor, lo que reduce su suministro de nutrientes y oxígeno provocando la muerte celular.

- Inmunoterapia: estimula el sistema inmunológico para atacar las células cancerosas. Los inhibidores de punto de control inmunitario, como los inhibidores de PD-1 (del inglés *Programmed cell Death Protein 1*) y PD-L1 (del inglés *Program-*

med Death-Ligand 1), han demostrado eficacia en subgrupos de pacientes con CCR.

- Vía de señalización de Wnt beta-catenina: esta vía desempeña un papel clave en la regulación de la proliferación celular y la diferenciación. Las mutaciones en el gen APC y otros genes asociados con esta vía son comunes en el cáncer de colon. Algunos medicamentos en investigación están diseñados para inhibir la actividad de la vía de Wnt/beta-catenina y reducir el crecimiento del tumor.

- Vía de reparación de ADN: en algunos casos de cáncer de colon, las células tumorales tienen defectos en la reparación del ADN, lo que las hace más susceptibles a ciertos medicamentos. Por ejemplo, los inhibidores de la polimerasa PARP (Poli ADP Ribosa Polimerasa) pueden ser efectivos en tumores con mutaciones en los genes de reparación del ADN como BRCA1 y BRCA2.

Es importante destacar que la elección de la terapia dirigida dependerá de varios factores, incluyendo el estado genético del tumor, el estadio de la enfermedad y las características individuales del paciente. El médico especialista en oncología determinará la mejor opción terapéutica basada en la evaluación individual del paciente.

Terapias dirigidas

También llamadas de precisión, son estrategias terapéuticas que se basan en el ataque y destrucción de células cancerígenas sin afectar a las células sanas. Están diseñadas para interferir en los procesos de crecimiento y supervivencia de las células cancerosas, proporcionando un enfoque más preciso y dirigido en comparación con la quimioterapia convencional. Pueden inhibir directamente la proliferación, diferenciación y migración celular. El microambiente del tumor, incluidos los vasos sanguíneos locales y las células inmunitarias, también podría verse alterados por fármacos dirigidos para impedir el crecimiento del tumor y promulgar una vigilancia y un ataque inmunitarios más fuertes [61].

En el tratamiento del cáncer de colon, se utilizan diversas terapias dirigidas que apuntan a dianas específicas en las células cancerosas. Algunas de las más utilizadas actualmente son:

- **Terapia anti-EGFR**: fármacos que bloquean la señalización del EGFR en las células cancerosas, inhibiendo su crecimiento y propagación[62].

- **Terapia anti-VEGF**: la inhibición VEGF trata de interferir en la formación de nuevos vasos sanguíneos en el tumor, reduciendo así el suministro de nutrientes y oxígeno al tumor y frenando su crecimiento.

- **Terapia anti-HER2**: en algunos casos de CCR, las células cancerosas pueden expresar una proteína llamada HER2, responsable del crecimiento y propagación del tumor. La inhibición de esta proteína es otro de los objetivos de las terapias dirigidas.

- **Terapia anti-BRAF**: los pacientes con mutaciones del gen BRAF pueden beneficiarse de terapias dirigidas específicas. Los fármacos utilizados se dirigen a las mutaciones del gen BRAF y bloquean su actividad, inhibiendo el crecimiento tumoral.

- **Terapia anti-PARP**: Los inhibidores de la PARP se utilizan en casos CCR con deficiencias en los sistemas de reparación del ADN, como las mutaciones en los genes BRCA1 y BRCA2. Estos medicamentos interfieren con la capacidad de las células cancerosas para reparar el ADN dañado, lo que lleva a su muerte.

Es importante destacar que la elección y la combinación de terapias dirigidas dependen de varios factores, como las características moleculares y genéticas del tumor, el estadio de la enfermedad y las consideraciones individuales del paciente. Suelen ser utilizadas en combinación con la cirugía, la quimioterapia y la radioterapia según las necesidades específicas de cada paciente, en concreto:

Debido a la creciente eficacia de tratamientos como los agentes anti-EGFR y los agentes anti-angiogénesis, que son medicamentos innovadores que inhiben numerosas vías inmunológicas, se están desarrollando otros similares a un ritmo sin precedentes [60].

La nanotecnología es un campo de rápido crecimiento en la administración de fármacos que ofrece varias ventajas sobre los métodos tradicionales. Los mecanismos innovadores de administración específicos para el colon permitirían la distribución local de una alta concentración de medicamentos en el colon, lo que mejoraría la farmacoterapia al tiempo que reduciría la toxicidad sistémica y otros resultados

adversos. Se han creado nanotransportadores teranósticos para monitorizar y tratar enfermedades utilizando un único sistema de administración [63, 64, 65].

Terapias génicas

Este tipo de terapias utilizan componentes genéticos para ayudar a sustituir o reparar genes que funcionan mal. La terapia génica también puede emplearse para desencadenar una respuesta inmunológica o como tratamiento en sí mismo.

Las mutaciones y aberraciones genéticas tienen una gran relevancia en la progresión del cáncer colorrectal. El potencial para inhibir el CCR puede residir en la modificación y corrección de estos genes defectuosos y en la prevención de los genes sobreexpresados [66].

La principal ventaja de la terapia génica es la transferencia de genes concretos a células tumorales específicas, lo que suprime la función aberrante del gen mutante y frena la progresión del tumor.

La inmunoterapia tumoral ha despertado gran interés científico porque resulta muy prometedora desde el punto de vista terapéutico. La terapia con anticuerpos monoclonales, la terapia con inhibidores del punto de control inmunitario, las vacunas contra el cáncer, la terapia celular adoptiva, la inhibición del complemento y el tratamiento con citocinas son algunas de las técnicas de inmunoterapia utilizadas frente al CCR [67]. La mayoría de ellas se encuentran en fases tempranas de estudio, pero los resultados han resultado alentadores en varios estudios de investigación [68, 69, 70].

- Anticuerpos monoclonales: son moléculas pequeñas que juegan un importante papel dentro de las terapias dirigidas, ya que penetran en las células e inactivan dentro de ellas enzimas seleccionadas, interfiriendo así en el crecimiento de las células tumorales e incluso desencadenando la apoptosis [71]. Además, ciertos anticuerpos monoclonales funcionan en células que no son cancerosas, como las células inmunitarias, lo que ayuda a manipular el sistema inmunitario para atacar el cáncer humano. Actualmente se están realizando estudios clínicos sobre compuestos activos contra la molécula de adhesión de células epiteliales, contra el antígeno carcinoembrionario (CEA) y contra las mucinas [72, 16].

- Terapia con inhibidores de puntos de control inmunitarios CTLA-4: molécula de punto de control inmunológico que se une a las estructuras de las células presentadoras de antígenos, inhibe la activación de las células T, cuya función está regulada negativamente por el ligando del receptor de muerte programada, que se une al receptor PD-1 de las células T y suele ser estimulado por sus diversos ligandos [73, 74]

- Otros: las vacunas, la inhibición del complemento y las terapias con citoquinas [75] son otras de las estrategias investigadas. Las primeras tratan de estimular el sistema inmunitario empleanto células cancerosas de pacientes combinadas con un adyuvante inmunoestimulante para provocar actividad inmunitaria antitumoral y prevenir la recaída del cáncer de colon tras la cirugía [76]. La depleción del complemento es un tipo eficaz de inmunoterapia para el CCR, ya que puede ralentizar la progresión tumoral al aumentar las respuestas inmunitarias del huésped frente al cáncer y disminuir el efecto inmunosupresor del microentorno tumoral. Las citocinas son componentes importantes de la inmunología tumoral, especialmente en el CCR, donde el proceso inflamatorio y las respuestas inmunogénicas impulsan el desarrollo tumoral [77]

Fundamentos Computacionales

Una vez se han comentado los fundamentos biológicos de la enfermedad, a continuación se presentan los fundamentos computacionales en los que se basa el proceso de trabajo seguido en la presente tesis para generar modelos bioinformáticos para la predicción de compuestos biológicamente activos en CCR.

Dichos modelos, necesitan medir/cuantificar en términos numéricos, la actividad biológica de diferentes fármacos a partir de sus reacciones y de las condiciones experimentales a las que son sometidos. Por ello, se introducirán en las siguientes secciones los modelos quimioinformáticos, descriptores moleculares y la teoría de la perturbación. A partir de la descripción numérica de los diferentes fenómenos biológicos, se desarrollarán, siguiendo diferentes perspectivas, modelos de machine learning de predicción de actividad biológica.

Modelos quimioinformáticos

Bajo las premisas de que la estructura de una molécula define su actividad biológica y moléculas estructuralmente similares pueden tener una actividad biológica similar, los modelos quimioinformáticos, que relacionan numéricamente las estructuras químicas de las moléculas con su actividad biológica, permiten, a través de sistemas matemáticos, predecir las propiedades fisicoquímicas y biológicas del destino que tendrá un nuevo compuesto a partir del conocimiento de su estructura química y de los estudios experimentales existentes.

Los modelos quimioinformáticos integran técnicas informáticas y estadísticas para realizar una predicción teórica de la actividad biológica que permita el diseño teórico de posibles nuevos fármacos, evitando el proceso de prueba y error de la síntesis orgánica. Al ser una ciencia que existe únicamente en un entorno virtual, permite prescindir de ciertos recursos como equipos, instrumentos, materiales y personal de laboratorio. Con un enfoque en las relaciones entre la estructura química y la actividad biológica, el diseño de candidatos para nuevos fármacos es mucho más rápido y barato.

Para realizar un estudio quimioinformático se necesitan tres tipos de información:

1. **Estructura molecular** de diferentes compuestos que tienen el mismo mecanismo de acción (farmacodinámica), y por lo tanto son considerados como ligandos a un objetivo biomolecular común.

2. **Datos de actividad biológica** de cada uno de los ligandos en estudio.

3. **Propiedades Físicoquímicas**, que se definen a través de los descriptores moleculares, de los ligandos calculados por medios computacionales a partir de la estructura molecular generada virtualmente por técnicas computacionales.

En el tipo prospectivo, los resultados en forma de ecuación o modelo quimioinformático permiten predecir la actividad biológica de compuestos aún no sintetizados que se generan prácticamente en poco tiempo, pero que deben compartir características estructurales de los ligandos incluidos en el estudio para no salirse del patrón químico o rango de valores de los descriptores. El otro tipo, la retrospectiva, analiza las moléculas ya existentes (las de síntesis y bioensayos) para comprender sus interrelaciones no obvias entre estructuras y actividades biológicas. La preparación de los datos de

entrada es el paso más crucial ya que el resultado se obtiene de forma automatizada y solo depende de la entrada.

La metodología quimioinformátiva es interdisciplinar, por lo que recibe información de Química Orgánica y Farmacología. La forma en que los modelos quimioinformáticos premian esta situación y que constituye el objetivo de esta metodología, es a través del diseño dirigido de ligandos que aún no existen, pero que a través de las ecuaciones generadas han mostrado una alta probabilidad de éxito farmacológico. Cuando se cuenta con información recolectada de la literatura o de un laboratorio, se utiliza una herramienta estadística llamada regresión lineal múltiple, tomando como variable dependiente los valores de actividad biológica de los ligandos y como variables independientes, los descriptores calculados.

El tiempo de una simulación molecular realizada mediante herramientas computacionales es mucho menor que el tiempo que llevaría la síntesis y bioensayos de nuevos compuestos, que pueden ser meses o incluso años. Esta ventaja permite tomar una serie de moléculas y gracias a la rapidez en la obtención de los resultados, alimentar directamente al laboratorio de síntesis en el proceso continuo del proyecto. Así,los modelos quimioinformáticos predicen nuevas estructuras nunca antes evaluadas y las propone a los químicos orgánicos para que sean sintetizadas y posteriormente llevadas a los bioensayos cuyos resultados confirmen o contradigan los valores predichos por el modelo quimioinformático. En un caso óptimo, a través de este ciclo operativo se obtienen mejores candidatos que a través de puro ensayo y error. Esto ahorra tiempo, dinero, recursos y evita el fracaso de quienes desarrollan nuevos fárcamos.

Las ventajas de la quimioinformática son el bajo costo, ya que no utiliza instrumentos de laboratorio, ni reactivos químicos, y además, existe software libre para la generación de los modelos que brindan interfaces que facilitan el manejo y diseño. Además, la construcción de las moléculas y el cálculo de los descriptores puede ser extremadamente rápido. Entre sus desventajas podemos mencionar la necesidad de capacitación en metodologías computacionales (diferentes sistemas operativos e interfaces gráficas, manejo de bases de datos, desarrollo de software) y en este sentido, la resolución de diferentes problemas computacionales (compatibilidad, actualizaciones, registros, formatos de datos) así como el hecho de tener que disponer de datos sobre la actividad biológica de las moléculas procedentes de una misma fuente, el cambio de perspectiva en la forma de trabajar, etc.

Descriptores moleculares

Los Descriptores Moleculares (DM) juegan un papel clave en muchas áreas de investigación. Se pueden definir como representaciones numéricas de la molécula que describen cuantitativamente su información estrucutural. Pero no toda la información contenida en una molécula, sino sólo una parte, puede extraerse a través de medidas experimentales. En las últimas décadas ha habido un interés creciente en cómo capturar y convertir, de forma teórica, la información codificada en la estructura molecular en uno o más números que se utilizan para establecer relaciones cuantitativas entre estructuras y propiedades, actividades biológicas y otros aspectos experimentales. De esta forma, los DM se han convertido en una herramienta muy útil para realizar la búsqueda de similitudes en repositorios moleculares, ya que pueden encontrar moléculas con propiedades fisicoquímicas similares según su similitud con los valores de los descriptores calculados.

Desde el inicio de su aplicación se han definido miles de descriptores moleculares, que codifican moléculas de diferentes formas, pudiendo dar una descripción genérica de la molécula entera (descriptores 1D), cuyo cálculo es más sencillo que aquellos descriptores que definen propiedades calculadas a partir de estructuras bidimensionales y tridimensionales (2D y 3D), que definen características más específicas, pero cuyo cálculo es más complejo.

Se ha argumentado que el número de descriptores atómicos y moleculares desarrollados hasta la fecha constituyen un arsenal suficiente para la búsqueda de nuevos fármacos a desarrollar. Sin embargo, una de las causas de la falta de ajuste en los modelos puede ser la propia naturaleza de la muestra o la selección inadecuada de los descriptores estructurales. Esto último puede deberse al procedimiento de selección utilizado, o incluso a la insuficiente capacidad de los modelos para describir el fenómeno. Todo esto es motivo suficiente para continuar la búsqueda de nuevos descriptores estructurales o atómicos que puedan utilizarse en estudios de modelos basados en modelos quimioinformáticos.

Los descriptores moleculares se pueden dividir, principalmente, en dos categorías bien diferenciadas:

- Experimentales: como log P, refractividad molar, momento dipolar, polarizabilidad y, en general, propiedades físico-químicas aditivas.

- Teóricos: se derivan de una representación simbólica de la molécula y pueden clasificarse según los diferentes tipos de representación molecular. Éstos a su vez, se clasifican en:

    1. **Constitucionales:** reflejan propiedades generales de naturaleza molecular.

    2. **Topológicos:** su cálculo se realiza a través de la teoría de grafos.

    3. **Geométricos:** se derivan de esquemas empíricos y codifican la capacidad de la molécula para participar en diferentes tipos de interacciones.

    4. **Electrónicos:** referidos a las propiedades electrónicas.

    5. **Fisicoquímicos:** definen el comportamiento de la molécula frente a reacciones. externas.

Atendiendo al número de dimensiones, los descriptores moleculares pueden clasificarse como se puede ver en la Figura 2.5.

Descriptores 0D. Son los más fáciles de calcular e interpretar. Se incluyen en esta categoría todos aquellos descriptores moleculares para cuyo cálculo no se necesita información estructural de la molécula ni conectividad entre átomos y por tanto son independientes de cualquier problema de conformación y no necesitan optimización de la estructura molecular.

Suelen mostrar una degeneración muy alta, es decir, tienen valores iguales para varias moléculas, como los isómeros. Su contenido de información es bajo, pero pueden jugar un papel importante en el modelado de varias propiedades fisicoquímicas o participar en modelos más complejos.

Ejemplos de estos descriptores son el número de átomos, el número de enlaces de un determinado tipo, el peso molecular, el peso atómico medio o la suma de propiedades atómicas como los volúmenes de Van der Waals.

Descriptores 1D. En esta categoría se pueden incluir todos los descriptores moleculares que permiten calcular información a partir de fracciones de una molécula. Suelen representarse como huellas dactilares, que no son más que vectores binarios

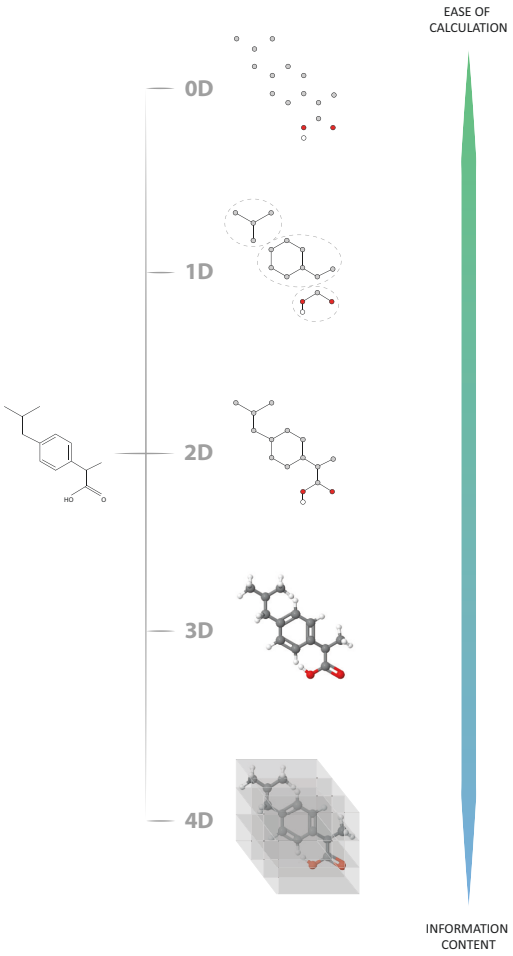

Fig. 2.5.: Representación de la información codificada por los diferentes descriptores moleculares según sus dimensiones.

en los que el 1 indica la existencia de una subestructura y el 0 indica su ausencia. Esta forma de representación tiene una gran ventaja y es que permite realizar cálculos muy rápidamente para encontrar similitudes entre moléculas. Al igual que los 0D, estos descriptores se pueden calcular fácilmente, se interpretan de forma natural, no requieren optimización de la estructura molecular y son independientes de cualquier problema de conformación. Suelen mostrar una degeneración media-alta y suelen ser muy útiles para modelar propiedades tanto fisicoquímicas como biológicas.

Dentro de los descriptores 1D se hace referencia a los basados en el recuento de grupos funcionales químicos, como el número total de átomos de carbono primarios, número de cianatos, número de nitrilos, etc, y los llamados fragmentos centrados en átomos, que se basan en el recuento de diferentes fragmentos de la molécula. Ejemplos de los últimos mencionados son hidrógeno unido a un heteroátomo, hidrógeno unido a un carbono alfa y flúor unido a un carbono primario.

Descriptores 2D. Describen propiedades que se pueden calcular a partir de representaciones bidimensionales de moléculas. Se obtienen a través de la teoría de grafos, independientemente de la conformación de la molécula. Su cálculo se basa en una representación gráfica de la molécula y presentan propiedades teóricas de estructura que se conservan por isomorfismo, es decir, propiedades con valores idénticos para gráficos isomorfos. La parte invariante puede ser un polinomio característico, una secuencia de números o un solo índice numérico obtenido aplicando operadores algebraicos a matrices que representan estructuras moleculares y cuyos valores son independientes de la numeración o etiquetado de los vértices.

Generalmente se derivan de una estructura molecular degradada en hidrógeno. Pueden ser sensibles a una o más estructuras características de la molécula, como tamaño, forma, simetría, ramificación y ciclicidad, y también pueden codificar información química sobre el tipo de átomo y la multiplicidad de enlaces. De hecho, generalmente se dividen en dos categorías:

1. **Índice de topología estructural:** codifican solo información sobre la adyacencia y la distancia de los átomos en la estructura molecular.

2. **Índice topoquímico:** cuantifican información sobre la topología pero también sobre las propiedades específicas de los átomos, como su identidad química o el estado de hibridación.

Descriptores 3D. Los descriptores tridimensionales están relacionados con la representación 3D de la molécula e incluyen la conformación de la estructura molecular, donde se consideran las distancias entre enlaces, ángulos de enlace, ángulos diédricos, etc., pudiendo luego describir las propiedades estereoquímicas de los moléculas. Su cálculo es más complejo que en los anteriores y puede requerir el análisis de muchas conformaciones moleculares.

Los descriptores 3D más comunes incluyen representaciones de moléculas de tipo farmacóforo, definidas como un conjunto de características estéricas y electrónicas necesarias para garantizar interacciones supramoleculares óptimas con un objetivo biológico específico y desencadenar o bloquear su respuesta biológica, donde características tales como centros hidrofóbicos o donantes de enlaces de hidrógeno, que se sabe o se cree que son responsables de la actividad biológica, se mapean en posiciones en una molécula. A continuación, se calculan y registran las distancias dependientes de la conformación entre estos puntos. Los farmacóforos de tres puntos se utilizan ampliamente, pero se han introducido farmacóforos de cuatro puntos más potentes, que pueden requerir el análisis de millones de posibles farmacóforos para un compuesto de prueba. Los descriptores 3D complejos se calculan, por ejemplo, para identificar conformaciones activas de un compuesto o para identificar características críticas que diferencien actividad en series de análogos. Al mismo tiempo, este tipo de cálculo es necesario para generar la ”forma de farmacóforo"de una molécula de consulta y buscar en bases de datos compuestos con características 3D similares. Además, el uso de descriptores de tipo de farmacóforo es fundamental para la derivación de modelos quimioinformáticos 3D o 4D.

¿Cuántos descriptores moleculares son necesarios?. En términos generales, el número de descriptores a utilizar dependerá de las herramientas computacionales disponibles y del número de moléculas incluidas en el estudio. El error más frecuente consiste en permitir que la operación matemática del modelo de regresión lineal agregue un espacio numérico sobredimensional para describir cada molécula de forma independiente sin poder establecer una regla con poder predictivo y confiable. Esto sucede cuando el número de descriptores excede el número de moléculas. Por otro lado, la búsqueda exhaustiva puede aplicarse a todos los casos excepto a los más simples, ya que el espacio de búsqueda no es práctico cuando hay un bajo número de descriptores moleculares. La confiabilidad del modelo puede verse afectada, no solo por la presencia de ruido, sino también por la correlación de descriptores redundantes y también por la presencia de descriptores irrelevantes. Por lo tanto, las técnicas de selección de variables se utilizan en gran medida para remediar esta situación y

mejorar la precisión y el poder predictivo de los modelos de clasificación o regresión [78].

En los últimos años, la comunidad científica ha prestado mucha atención a las técnicas dedicadas a la selección de variables, es decir, la selección de descriptores moleculares en modelos quimioinformáticos. Dado que hay miles de descriptores disponibles para describir una molécula y, a menudo, no hay un conocimiento *a priori* sobre qué características son las más responsables de una propiedad específica, se exploran subconjuntos de los descriptores más apropiados a través de diferentes estrategias. Hoy en día existen muchas herramientas de software para el cálculo de descriptores moleculares, cada una con sus ventajas y desventajas (facilidad de uso, licencias, número de descriptores, etc.).

Teoría de la perturbación

La teoría de perturbación en quimioinformática es un enfoque fundamental utilizado para predecir y describir propiedades moleculares y las interacciones químicas en sistemas moleculares complejos. Basada en los principios de la mecánica cuántica y la teoría del campo molecular, la teoría de perturbación permite analizar los efectos de pequeñas perturbaciones en la estructura y las propiedades de las moléculas. Por norma general, los modelos se generan con un dominio de aplicación reducido, centrados únicamente solo en un conjunto de condiciones, por ejemplo una propiedad específica, una proteína objetivo o una línea celular. Es muy interesante poder considerar múltiples condiciones de ensayo al mismo tiempo, lo que se facilita mediante la teoría de la perturbación.

En el ámbito de la quimioinformática, las perturbaciones pueden manifestarse en cambios en la geometría molecular, sustitución de átomos o grupos funcionales, interacciones intermoleculares, efectos de solvatación, entre otros. Mediante la aplicación de métodos de perturbación, se logra cuantificar los cambios en las energías, las propiedades espectroscópicas y otras propiedades moleculares como resultado de estas perturbaciones. Comienzan con una función de referencia que mide la probabilidad de que un fármaco sea activo bajo ciertas condiciones y utiliza operadores PT (PTO) para dar cuenta de las desviaciones (perturbaciones) de las variables de entrada de este fármaco con respecto a una población de fármacos ensayados en las mismas condiciones.

La teoría de perturbación en quimioinformática encuentra aplicación en diversos campos, entre ellos:

- Descriptores moleculares: Los métodos de perturbación se emplean para calcular y evaluar descriptores moleculares, los cuales proveen información sobre las características físico-químicas de las moléculas. Estos descriptores son esenciales en el diseño de fármacos, la predicción de la actividad biológica y la evaluación de propiedades químicas relevantes.

- Interacciones moleculares: La teoría de perturbación es utilizada para analizar y comprender las interacciones entre moléculas, como en el estudio de enlaces de hidrógeno, interacciones de apilamiento, fuerzas de van der Waals, interacciones electrostáticas, entre otras. Estos análisis permiten comprender las estructuras y propiedades de los sistemas moleculares y son útiles en el diseño de materiales y la predicción de propiedades fisicoquímicas.

- Efectos de solvatación: La teoría de perturbación se aplica para estudiar los efectos de la solvatación, es decir, las interacciones de una molécula con el solvente circundante. Esto es relevante para la predicción de constantes de disociación, solubilidad, estabilidad y otras propiedades de las moléculas en solución.

En síntesis, la teoría de perturbación en quimioinformática es una herramienta esencial para comprender y predecir propiedades moleculares y las interacciones químicas en sistemas complejos. Al emplear los principios de la mecánica cuántica y la teoría del campo molecular, se logra obtener descripciones cuantitativas de los cambios en las energías y propiedades moleculares debido a perturbaciones específicas. Esto resulta de gran utilidad en áreas como el diseño de fármacos, la química computacional y otros campos afines.

3 Estado de la cuestión

En este capítulo se expone el estado de la cuestión, se ha revisado para ello la literatura existente y se han analizado investigaciones previas, teorías y enfoques relacionados con la temática multidisciplinar de la tesis. Se presentan por ello los estudios más recientes y relevantes sobre la temática, proporcionando un contexto sólido que permite comprender la relevancia y originalidad de la investigación propuesta. Se ha dividido la sección en dos grandes puntos, primero se mostrarán diferentes trabajos de predicción de compuestos biológicamente activos y, seguidamente, se mostrarán las técnicas de machine learning utilizadas más habitualmente para la generación de las predicciones.

Un fármaco se puede definir como una molécula que interacciona con una entidad funcional del organismo, denominada diana terapéutica o diana molecular, modificando de alguna manera su comportamiento. Los fármacos conocidos actúan sobre dianas conocidas, pero el descubrimiento de nuevos fármacos que puedan modificar el curso de una enfermedad o mejorar la eficacia de los tratamientos existentes es uno de los principales objetivos de la investigación en el campo de la química y la biología.

El desarrollo de un nuevo fármaco puede tardar hasta 12 años y se estima que su coste medio, hasta que llega al mercado, es de aproximadamente mil millones de euros. El tiempo y los costos involucrados están en gran parte asociados con la gran cantidad de moléculas que fallan en una o más etapas de su desarrollo, ya que se estima que solo 1 de cada 5,000 medicamentos finalmente llega al mercado. Las estadísticas anteriores muestran que el descubrimiento y desarrollo de nuevos fármacos es un proceso muy complejo y costoso. Este proceso se ha llevado a cabo durante mucho tiempo utilizando métodos exclusivamente experimentales. Los avances tecnológicos de las últimas décadas han promovido el nacimiento del término *in silico*, término que ya es común en los laboratorios de biología, y que designa un tipo de experimento que no se realiza directamente sobre un organismo vivo (estos se denominan experimentos *in vivo*) o en un tubo de ensayo u otro entorno artificial fuera del organismo (experimentos denominados *in vitro*), pero se llevan a cabo virtualmente a través de simulaciones informáticas de procesos biológicos.

La complejidad de la biología moderna ha hecho que estas herramientas computacionales sean imprescindibles para la experimentación biológica, ya que permiten codificar modelos teóricos con gran precisión y son capaces de procesar grandes cantidades de información, facilitando y acelerando el proceso de desarrollo de nuevos fármacos.

Predicción de compuestos biológicamente activos y herramientas quimioinformáticas

El desarrollo de un nuevo fármaco comienza con la fase de búsqueda, mediante cribado de alto rendimiento, de aquellas moléculas o compuestos que muestran actividad biológica frente a una diana terapéutica o diana molecular, que es el lugar del cuerpo donde se pretende que actúe el fármaco. A esta fase le sigue la generación de *leads*, donde las moléculas previamente seleccionadas son validadas y refinadas estructuralmente para aumentar su potencia con respecto al objetivo.

Además, se esperan propiedades farmacocinéticas apropiadas, es decir, tasas adecuadas de absorción, distribución, metabolismo y eliminación, así como bajas tasas de toxicidad y efectos adversos. Se hablará de todo ello a continuación y se finalizará la sección separando los trabajos identificados en el estado de la cuestión agrupados por el tipo de estudio, área de conocimiento o patología en la que se ha centrado la búsqueda de compuestos biológicamente activos.

Propiedades farmacocinéticas

El concepto de similitud de fármacos, establecido a partir del análisis de las propiedades fisicoquímicas y características estructurales de compuestos existentes o candidatos, ha sido ampliamente utilizado para filtrar compuestos con propiedades indeseables en términos de Administración, Distribución, Metabolismo, Eliminación y Toxicidad (ADMET) [79]. El estudio de las fases ADMET por las que pasa un fármaco tras ser administrado a un individuo es otra de las tareas fundamentales en el desarrollo de nuevos compuestos [80]. La alteración en un paciente de alguna de estas fases ADMET (por ejemplo, problemas de excreción por algún tipo de insuficiencia renal, aumento del volumen de distribución en personas obesas, problemas de absorción por patología gastrointestinal o problemas en el metabolismo del fármaco por deterioro hepático), puede influir en la concentración final del fármaco, modificando

la respuesta esperada del organismo y haciendo necesario disminuir o aumentar la dosis del fármaco, según cada caso. Por ello, es fundamental en las primeras etapas de la investigación estimar el comportamiento de las propiedades farmacocinéticas de los compuestos, para lo que se han desarrollado herramientas para mejorar y acelerar esta fase de desarrollo, como por ejemplo Chemi-Net [81].

Otro ejemplo sería el del coeficiente de reparto octanol-agua, que es una medida de la hidrofobicidad o afinidad lipídica de una sustancia disuelta en agua. Los compuestos químicos con valores elevados de este coeficiente suelen acumularse en las porciones lipídicas de los organismos, produciendo así toxicidad. Por el contrario, los compuestos de bajo coeficiente tienden a distribuirse en el agua o el aire, por lo que podrían eliminarse del organismo sin acumularse [82]. Además, la estimación adecuada de la vida media de eliminación de un fármaco tendría aplicaciones potenciales para las primeras evaluaciones farmacocinéticas y, por lo tanto, proporcionaría una guía para diseñar candidatos a fármacos con un perfil de exposición favorable *in vivo* [83].

El metabolismo es la principal vía de eliminación del organismo para la mayoría de los 200 medicamentos más comercializados. El estudio de la estabilidad de los microsomas hepáticos fortificados con NADPH es habitual en la investigación de nuevos fármacos para predecir el aclaramiento y así poder estimar la exposición máxima a un fármaco para una dosis dada [84]. Además, el hígado es el principal órgano implicado en el metabolismo de los fármacos y, por tanto, el daño hepático causado por los fármacos a menudo ha dificultado el desarrollo de nuevos fármacos. La evaluación del riesgo de daño hepático para los candidatos a fármacos es una estrategia eficaz para reducir el riesgo de que un estudio no avance con el descubrimiento de nuevos fármacos.

La toxicidad es una de las principales causas del fracaso de la investigación y el desarrollo de fármacos. Datos internacionales mostraron que en el período de 2006 a 2010, la toxicidad representó el 22 y el 54 % de los fracasos en la investigación y desarrollo de fármacos, en las etapas clínica y preclínica, respectivamente. Las Reacciones Adversas a Medicamentos (RAM), que pueden aumentar la morbilidad, se producen de forma más significativa en pacientes hospitalizados y, además de la carga clínica, también suponen un importante coste económico. Por todas estas razones, la detección virtual en etapa temprana de moléculas candidatas a fármacos juega un papel clave en la industria farmacéutica para prevenir RAM. Por lo tanto, es fundamental estudiar las propiedades toxicológicas lo antes posible y dar prioridad a los principales compuestos que representan la menor amenaza en la etapa de

descubrimiento de éxitos, aumentando así las posibilidades de éxito durante el desarrollo clínico.

Los modelos predictivos de toxicidad generados por ordenador, pueden orientar las pruebas de seguridad de fármacos en la dirección correcta y, en consecuencia, acortar el tiempo requerido y ahorrar costos durante el desarrollo de medicamentos en el laboratorio. Algunas reacciones adversas pueden ser parte de la acción farmacológica natural de un fármaco que no se puede evitar, pero más a menudo pueden ser impredecibles en la etapa de desarrollo. Pueden ocurrir debido a la administración de un fármaco o al uso combinado de dos o más.

Para avanzar en la investigación de compuestos contra agentes infecciosos, estos compuestos deben mostrar una falta relativa de toxicidad en células de mamíferos. Antes de la aplicación clínica de un fármaco, debe pasar por una fase preclínica, para estudiar sus efectos y realizar ensayos clínicos de seguridad de los fármacos [85]. La gran cantidad de efectos adversos potenciales que pueden ocurrir con el consumo de fármacos, solos o en combinación, dificulta la detección de muchos de estos efectos adversos durante el desarrollo temprano, por lo que se han propuesto herramientas como SDHINE [86]. En [87], están desarrollando una nueva metodología para predecir los efectos secundarios de los medicamentos, que también puede ayudar a revelar las posibles causas de los efectos adversos. En general, también se han desarrollado herramientas computacionales para distinguir entre compuestos cancerígenos y no cancerígenos [88].

Interacciones farmacológicas

Determinar el objetivo a alcanzar es lo más importante y más propenso a errores en el desarrollo de un tratamiento terapéutico para una enfermedad, donde las fallas son potencialmente costosas debido a los largos plazos y gastos del desarrollo de fármacos. El análisis de ILP se ha convertido en un requisito previo crucial para el descubrimiento de nuevos fármacos [89, 90, 91, 92, 93]. Los experimentos in vitro se usan comúnmente para identificar los CPI, pero no es factible realizar esta tarea solo a través de enfoques experimentales, y los avances en el aprendizaje automático para predecir los CPI han hecho grandes contribuciones al descubrimiento de fármacos. Para mejorar la tarea de predecir las interacciones ligando-proteína, se utilizan herramientas como [94]. Además, el estudio de estas interacciones es fundamental

para obtener trazas de nuevos fármacos y predecir sus efectos secundarios a partir de fármacos aprobados y candidatos [95].

Por otro lado, la identificación de la viabilidad de las proteínas diana es otro de los pasos preliminares del descubrimiento de fármacos [96, 97, 98]. Determinar la capacidad de una proteína para unirse a fármacos con el fin de modular su función, llamada capacidad de fármacos, requiere una cantidad no trivial de tiempo y recursos. Esta tarea cuenta con la ayuda de nuevas funciones desarrolladas en eFindSite [99], que es un software independiente disponible de forma gratuita, que utiliza el aprendizaje automático supervisado para predecir la capacidad farmacológica de una proteína determinada.

Los ligandos cristalizados en el Banco de Datos de Proteinas (BDP) pueden tratarse como formas inversas de los sitios activos de las proteínas correspondientes. La similitud de forma entre una molécula y los ligandos de los BDP indica la posibilidad de que la molécula se una a ciertos objetivos [100]. Las proteínas de membrana están involucradas en muchos mecanismos biomoleculares esenciales como un factor clave para permitir el transporte de pequeñas moléculas y señales en ambos lados de la membrana celular. Por lo tanto, la identificación precisa de los sitios de unión de proteínas y ligandos de membrana mejorará significativamente el descubrimiento de fármacos. Con este fin, MPLs-Pred [101] se ha desarrollado como una herramienta disponible gratuitamente para usuarios generales. La predicción de interacciones a partir de métodos Multi Kernel permite identificar posibles pares de interacción fármaco-objetivo [102].

Enfermedades infecciosas

Las enfermedades infecciosas son causadas por microorganismos patógenos como las bacterias, los virus, los parásitos o los hongos. Estas enfermedades pueden transmitirse, directa o indirectamente, de una persona a otra.

Los antibióticos son el tratamiento de elección de las enfermedades bacterianas, pero el aumento y mal uso de los mismos ha provocado la aparición de resistencias bacterianas a muchos de los antibióticos actuales, de ahí la necesidad de generar nuevos compuestos que puedan combatir organismos multirresistentes. Así, se está investigando mucho la Halicina [103], que es una molécula con capacidad bactericida que se ha mostrado muy prometedora atacando bacterias difíciles de tratar con los

antibióticos actuales. Su estructura es diferente a la de los antibióticos convencionales y muestra capacidad bactericida frente a un amplio espectro filogenético que incluye *mycobacterium tuberculosis* (bacteria para la que se siguen buscando tratamientos más eficientes [104]), y enterobacterias, *clostridium difficilae* y *acinetobacter baumanii*. Pero las resistencias no son exclusivas de las bacterias, y la resistencia de los tratamientos actuales para la malaria, producida por protozoo parásito *plasmodium falciparum*, es preocupante pues afecta a más de 200 millones de personas en el mundo. Si bien el tratamiento de elección incluye combinaciones de medicamentos [105], ya hay estudios en marcha para identificar nuevas combinaciones que puedan eludir los mecanismos de resistencia actuales[106].

Los virus pueden causar enfermedades infecciosas banales como el resfriado común y la gripe, para las que actualmente no hay tratamiento específico (es el sistema inmunitario el encargado de eliminarlo del organismo) o el que hay es preventivo (vacunas). Pero los virus también causan enfermedades graves como SIDA, ébola (causante de gran número de estudios [107] por la epidemia ocurrida en el año 2016), o la COVID-19 (en cuyo tratamiento curativo o preventico se centran la mayor parte los esfuerzos actuales a nivel mundial). Además, la coinfección de determinados virus es frecuente en determinadas poblaciones, como por ejemplo la coinfección por el virus de la inmunodeficiencia humana tipo 1 y el virus de la hepatitis C (VIH-VHC). En este caso el tratamiento de la coinfección es un desafío debido a las consideraciones especiales a tener en cuenta para garantizar la seguridad hepática y evitar interacciones farmacológicas. Por ello se buscan fármacos que sean efectivos frente a múltiples patógenos y con menos toxicidad que puedan proporcionar una estrategia terapéutica en determinadas coinfecciones [108].

Cáncer

El cáncer es un importante problema de salud pública en todo el mundo y, por tanto, es imperativo que se desarrollen nuevos fármacos para su tratamiento. El objetivo principal de la investigación del cáncer es descubrir el método de tratamiento más efectivo para cada paciente con cáncer, ya que no todos responden igual a un tratamiento específico debido a factores externos, como el uso de productos de tabaco y dietas poco saludables, e internos, como la heterogeneidad de las células cancerosas y las condiciones inmunitarias. Dado que el número de pacientes con cáncer en todo el mundo aumenta cada año, sería invaluable poder predecir correctamente la respuesta o la falta de respuesta del cáncer a un fármaco específico.

La palabra cáncer se refiere a la proliferación descontrolada de células anormales, que cuando superan a las células normales, dificultan que el cuerpo funcione como debe. Aunque la palabra cáncer se utiliza de forma general, el término engloba un abanico de enfermedades con patogenia, evolución y tratamiento muy diferentes.

En general, se han estudiado e identificado muchas dianas terapéuticas que pueden ser susceptibles de tratamiento. Las técnicas computacionales han sido ampliamente utilizadas para predecir la actividad de muchos compuestos sobre estas dianas, por lo que por ejemplo se sabe que los Receptores Acoplados a Proteína G (RAPG) juegan un papel clave en muchos mecanismos de señalización celular cuya alteración puede estar involucrada en la patogénesis del cáncer [109]. La proteína 4, que contiene bromodominio (BRD4), se ha convertido en una diana terapéutica prometedora para muchas enfermedades, como el cáncer, la insuficiencia cardíaca y los procesos inflamatorios. La nitroxolina es un antibiótico que mostró potencial sobre BRD4 con actividad inhibidora contra las líneas celulares de leucemia, y se ha demostrado que es eficaz contra esas líneas celulares de leucemia [110]. La indolamina 2,3-dioxigenasa (IDO), un punto de control inmunológico, es un objetivo prometedor para la inmunoterapia contra el cáncer. Se han identificado tres inhibidores de IDO con actividad potente mediante métodos de aprendizaje automático [111], pero aún no se han aprobado para uso clínico. También se han realizado intentos para predecir la respuesta al mismo fármaco de diferentes tipos de tumores [112], incluidos el cáncer de mama, el cáncer de mama triple negativo y el mieloma múltiple. También se han utilizado métodos computacionales para estudiar las histonas desacetilasas (HDAC), que son una clase importante de objetivos enzimáticos para la terapia del cáncer, y se han buscado compuestos inhibidores de estos mediante técnicas computacionales [113].

Por otro lado, la proteína fosfoinositida 3-quinasa (PI3K) juega un papel clave en una vía de señalización intracelular responsable de muchos procesos en respuesta a señales extracelulares, como la regulación del ciclo celular, la supervivencia celular, el crecimiento celular, la angiogénesis, etc. Tumores vasculares en niños a menudo muestran mutaciones en esta molécula, por lo que se ha convertido en un objetivo farmacológico prometedor para la quimioterapia contra el cáncer. Los inhibidores de PI3K han cobrado importancia como estrategia viable de tratamiento del cáncer, ya que controlan la mayoría de las características del cáncer, incluido el ciclo celular, la supervivencia, el metabolismo, la motilidad y la inestabilidad genómica. Se realizó una evaluación virtual basada en la estructura para identificar los inhibidores de PI3K [114]. Otra glicoproteína de membrana estudiada como diana, y para la que

se han realizado simulaciones de inhibición, es la la glucoproteína de membrana P-gp [115]. También se han utilizado diferentes líneas de células tumorales para estudiar los anticuerpos como un posible tratamiento [116] cuantificando los niveles de proliferación y apoptosis para predecir su funcionamiento.

Las moléculas se pueden identificar como agentes anticancerígenos a través de dos métodos de descubrimiento de fármacos ampliamente utilizados [117]: descubrimiento de fármacos basado en objetivos (TDD, target-first, biología química directa) y descubrimiento de fármacos basado en fenotipos (PDD, función primero , biología química inversa).

Las quinasas son una de las familias más grandes que se consideran objetivos farmacológicos atractivos para las enfermedades neoplásicas debido a su papel fundamental en la transducción de señales y la regulación de la mayoría de las actividades celulares [118]. Como resultado, los inhibidores de quinasas han cobrado gran importancia en el descubrimiento de fármacos contra el cáncer en las últimas dos décadas, ya que, a pesar del considerable esfuerzo académico y de la industria, el conocimiento químico actual de los inhibidores de quinasas es limitado y, por lo tanto, se han desarrollado herramientas como Kinformation [119] , que se basa en métodos de aprendizaje automático para automatizar la clasificación de estructuras de quinasas y se espera que este enfoque mejore el modelado de proteínas quinasas tanto en conformaciones activas como inactivas.

El crecimiento neoplásico y la diferenciación celular son características fundamentales del desarrollo tumoral. Está bien establecido que la comunicación entre las células tumorales y las células normales, a través de canales que contienen tejido conectivo, incluidos enlaces gap, nanotubos tunelizados y hemocanales, regula la diferenciación y proliferación tumoral, la agresividad y la resistencia al tratamiento. Se ha propuesto que los nuevos enfoques computacionales [120] para la identificación y caracterización de estos sistemas de comunicación y su señalización asociada podrían proporcionar nuevos objetivos para prevenir o reducir las consecuencias del cáncer.

Se ha postulado que las nuevas combinaciones de fármacos pueden mejorar la terapia personalizada contra el cáncer. Usando varios tipos de información genómica sobre líneas celulares de cáncer, objetivos de fármacos e información farmacológica, es posible predecir la sinergia de combinación de fármacos mediante la regresión del

nivel de sinergia entre dos fármacos y una línea celular, así como clasificar si existe sinergia o antagonismo entre ellos. [121]

Para el diagnóstico actual de muchos cánceres, las medidas morfométricas nucleares se utilizan para hacer un pronóstico preciso en la última etapa, pero el diagnóstico temprano sigue siendo un gran desafío. La evidencia reciente destaca la importancia de las alteraciones en las propiedades mecánicas de las células individuales y sus núcleos como impulsores críticos para la aparición del cáncer. La detección de cambios sutiles en la morfometría nuclear con resolución de una sola célula mediante la combinación de imágenes de fluorescencia y aprendizaje profundo [122] permite discriminar entre células normales y líneas celulares de cáncer de mama, lo que abre nuevas vías para el diagnóstico temprano de enfermedades y el descubrimiento de fármacos.

Trastornos del Sistema nervioso Central

La neuropatía periférica inducida por quimioterapia (CIPN, por sus siglas en inglés) es un efecto secundario adverso común de la quimioterapia contra el cáncer, que puede causar dolor extremo e incluso incapacitar al paciente. La falta de conocimiento sobre los mecanismos de toxicidad multifactorial de ciertos compuestos ha impedido la identificación de nuevas estrategias de tratamiento, pero los modelos computacionales de neurotoxicidad de fármacos [123] se utilizan al principio del desarrollo de fármacos para detectar compuestos de alto riesgo y seleccionar fármacos candidatos más seguros.

Muchos trastornos del Sistema nervioso Central (SNC), tanto procesos neurodegenerativos como traumatismos, requieren múltiples estrategias para abordar la neuroprotección, reparación y regeneración de las células. El conocimiento acumulado en procesos neurodegenerativos y tratamientos neuroprotectores puede utilizarse, a través de técnicas computacionales como el ML, para identificar combinaciones de fármacos que puedan reutilizarse como posibles agentes neuroprotectores [124]. Otra rama de la neurología que genera gran interés científico es el estudio de las enfermedades neurodegenerativas, como la enfermedad de Alzheimer, principal causa de demencia y patología que actualmente no tiene cura. Varios estudios han reportado que la expresión de ROCK2, pero no de ROCK1, ha aumentado significativamente en el tejido nervioso humano de pacientes con trastornos neurodegenerativos, por lo que la supresión de la expresión de ROCK2 se considera una diana farmacológica

para el tratamiento de esta enfermedad. [125]. En el mismo sentido, el 5-HT1A es un receptor cerebral utilizado como biomarcador de trastornos degenerativos. Se ha trabajado para predecir los compuestos que se unirán a este receptor [126]. En general, en base a igual número de fármacos aprobados o retirados para el tratamiento de patologías del SNC, se han estudiado posibles fragmentos discriminativos que permitan buscar otros compuestos similares para el tratamiento de patologías del SNC [127].

Las infecciones del sistema nervioso central (SNC) son una causa muy importante de morbimortalidad. El paso de fluidos y de solutos al SNC está estrechamente regulado a través de la barrera hematoencefálica (BHE). La penetración de cualquier fármaco en el líquido cefalorraquídeo (LCR) depende del tamaño molecular, lipofilicidad, unión a proteínas plasmáticas y su afinidad por transportadores. En la búsqueda de fármacos indicados para las infecciones del SNC, es fundamental predecir la capacidad de los mismos para atravesar esta barrera [128], pudiendo desechar desde un principio aquellos que no lo hagan en unas concentraciones mínimas.

Diabetes Mellitus tipo 2 y obesidad

La diabetes mellitus tipo 2 es la patología endocrina más común en el mundo, causando muchas complicaciones en muchos sistemas de órganos que pueden conducir a un acortamiento de la vida y una reducción considerable en la calidad de vida de los pacientes que la padecen. Por ello, la industria farmacéutica ha realizado muchos esfuerzos en la búsqueda de tratamientos eficaces que puedan curar esta enfermedad o, en su defecto, minimizar las lesiones producidas en órganos diana por el exceso de glucosa en sangre. Una de las ramas de investigación se centra en la inhibición de los cotransportadores de glucosa dependientes de sodio (SGLT1 y SGLT2), a través de los cuales se absorbe la glucosa. Se han desarrollado inhibidores duales, pero continúa la búsqueda de compuestos destinados a reducir la absorción de glucosa por SGLT1 [129].

Otra rama de la investigación dentro de la endocrinología se centra en la fisiopatología y tratamiento de la obesidad. Se sabe que los receptores nucleares PPAR (Peroxisome Proliferator Activated Receptors) son factores de transcripción que se activan mediante la unión de ligandos específicos y regulan la expresión de genes implicados en el metabolismo de lípidos y glucosa. Estos receptores han sido propuestos como dianas terapéuticas para enfermedades metabólicas, y se ha desarrollado

ISE (Iterative Stochastic Elimination) [130], una herramienta que permite distinguir compuestos agonistas de PPARs.

Cuando el componente del complemento C1 está sobreactivado, su regulación puede verse alterada produciendo un daño tisular que vuelve a activar el sistema del complemento, produciendo así un círculo de activaciones que se perpetúa y produce un daño cada vez mayor. Los tratamientos para inhibir C1 son caros, por lo que continúa la búsqueda de inhibidores más baratos [131].

Como se mencionó, el desarrollo de nuevos medicamentos es un proceso complejo que requiere muchos recursos. Por ello, la búsqueda de nuevas indicaciones clínicas para fármacos existentes se ha convertido en una alternativa para acelerar y reducir los costes del proceso. Así, el término reposicionamiento de fármacos se refiere al proceso de desarrollo de un compuesto para su uso en una patología distinta a su indicación actual, aprovechando los beneficios de la abundancia, variedad y fácil acceso a productos farmacéuticos y datos biomédicos [132]. Un enfoque prometedor para el reposicionamiento de medicamentos es aprovechar los algoritmos de aprendizaje automático para aprender patrones en los datos biológicos disponibles relacionados con los medicamentos y vincularlos con enfermedades específicas para tratar [133, 134]. Por ejemplo, las indicaciones de los compuestos contra la malaria, la tuberculosis y el carcinoma de células grandes ya se están reutilizando para predecir las interacciones de las proteínas mediante el cálculo de la precisión mediante la comparación de la similitud de las interacciones de los medicamentos aprobados para otras indicaciones [135].

La OMS propuso una clasificación que asigna códigos a los compuestos según sus características terapéuticas, farmacológicas y químicas, así como los sitios de actividad in vivo. La capacidad de predecir los códigos ATC de los compuestos puede ayudar en la creación de bibliotecas químicas de alta calidad para la detección de compuestos y el reposicionamiento de fármacos [136].

Algoritmos de Machine Learning

Fue en 1964 cuando Hansch *et al.* [137] propusieron la ecuación de Hansch. Se trataba de un modelo de regresión lineal que utilizaba descriptores fisicoquímicos (como el parámetro de hidrofobicidad, el parámetro electrónico y el parámetro estérico), utilizado descriptores 2D. Así pues, utilizando un algoritmo predictivo

como la regresión lineal y descriptores moleculares de las secuencias, se inició el campo de estudio de la quimioinformática.

No fue hasta 1998, cuando Ajay *et al.* [138] introdujeron el concepto de *Drug-likeness*. Para ello, construyeron un modelo capaz de predecir con un alto rendimiento si una molécula era un fármaco o no. Lo hicieron a partir de descriptores 1D y 2D. Este trabajo fue pionero en el campo del descubrimiento de fármacos basado en algoritmos de ML.

Debido a la baja disponibilidad de datos, han sido pocos los trabajos basados en ML en el campo del descubrimiento de fármacos antes del año 2000. Con el avance de las técnicas biotecnológicas y computacionales, se han generado cada vez más datos de moléculas que se han puesto a disposición del público en general. Además, grandes iniciativas han desarrollado repositorios públicos donde la información sobre un gran número de moléculas está disponible de forma estandarizada. Fue en 2004 cuando se publicó la primera versión de dos bases de datos que serán de gran importancia para este campo: PubChem y ZINC.

Posteriormente, en 2006 y 2008 se publicaron DrugBank y ChEMBL, respectivamente. Fue este hecho, la disponibilidad de este gran número de bases de datos públicas, lo que permitió el desarrollo y entrenamiento de nuevos modelos de Machine Learning para ayudar en el cribado de nuevos fármacos. Para cuantificar este incremento, la Figura 3.1 muestra un histórico del número de artículos publicados en PubMed en el campo. Para realizar la figura, se estratificó la búsqueda según los diferentes algoritmos revisados en esta tesis doctoral. En concreto, se realizó una búsqueda booleana en PubMed con los siguientes términos: nombre del algoritmo (*'Artificial Neural Networks', 'Support Vector Machines', 'Naive Bayes' o 'Random Forest'*) & término '*Drug Discovery*' & periodo 1964-2021.

Como puede verse en la Figura 3.1, entre 2004 y 2008 se produjo un enorme crecimiento en el uso de algoritmos de ML. En concreto, las máuinas de vector soporte (SVM del inglés Support Vector Machines) ha sido con diferencia el algoritmo más utilizado en los últimos años. A partir de 2008 también se observó un punto de inflexión en el uso de las redes neuronales. Fue en este año cuando se liberó la librería Tensorflow. A partir de este momento, la aplicación de las Redes de Neuronas Artificiales (RNA) y especialmente el uso de modelos de *Deep Learning* impulsó el número de publicaciones en este campo, hasta convertirse en uno de los algoritmos más utilizados. La gran variedad de topologías de RNA existentes provocó este cre-

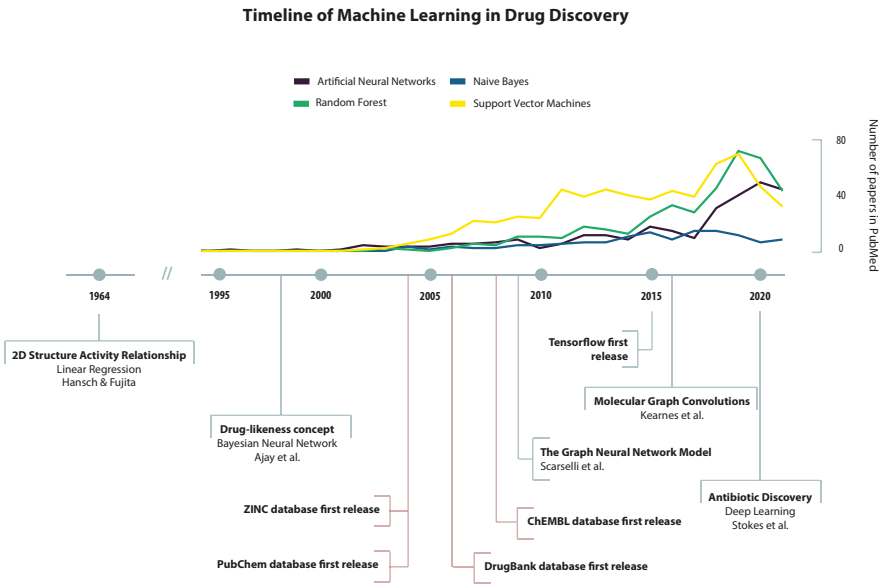

Fig. 3.1.: Cronología de los principales acontecimientos del aprendizaje automático en el campo del descubrimiento de fármacos. La figura representa los principales acontecimientos del aprendizaje automático en el campo del descubrimiento de fármacos. Además, se ha trazado una línea para mostrar el número de artículos a lo largo del tiempo. Cada algoritmo está representado por una línea de color. El eje Y representa el número de artículos publicados en PubMed.

cimiento. Por ejemplo, *The Graph Neural Network Model* [139], publicado en 2009, abrió un nuevo campo de aplicación en el descubrimiento de fármacos. Hasta entonces, la gran mayoría de los modelos utilizaban datos de modelos quimioinformáticos. Gracias a estos modelos, las entradas a las redes podían ser, por ejemplo, grafos moleculares, que también generaron un gran número de publicaciones científicas. Posteriormente, en 2016 se publicó *Molecular Graph Convolutions* [140], otro modelo que incluía las características de las redes convolucionales para el análisis de grafos moleculares. Fueron estos modelos los que se utilizaron en el trabajo de Stokes *et al.* [103]. Ese trabajo, publicado en 2020 en la revista Cell, demostró la capacidad de los modelos de aprendizaje profundo en este campo. Entrenaron un modelo de aprendizaje profundo capaz de predecir la actividad antibacteriana. Posteriormente, realizaron predicciones sobre múltiples bibliotecas químicas, descubriendo una molécula, la halicina, con actividad antibacteriana. Este descubrimiento se probó en laboratorio húmedo, reforzando la hipótesis establecida en los experimentos in silico.

Naïve bayes

En términos generales, los algoritmos de aprendizaje automático intentan encontrar la mejor hipótesis a partir de datos de interés dados. Por ejemplo, para un problema de clasificación, sería la clase de una muestra de datos desconocida para el modelo. Los clasificadores bayesianos asignan la clase más probable de cada muestra según la descripción dada por los valores vectoriales de sus variables. En su versión más simple, el algoritmo asume que las variables son independientes, es decir, facilita la aplicación del Teorema de Bayes [141]. Aunque esta suposición es poco realista (no todas las variables tienen la misma importancia), la familia de clasificadores que surge de la premisa anterior, conocida como NB (Naïve Bayes) obtiene resultados sobresalientes, aunque en algunos casos existen fuertes dependencias en su conjunto de atributos. Este algoritmo describe una forma sencilla de aplicar el teorema de Bayes a problemas de clasificación, y es un modelo simple y rápido que es capaz de trabajar con datos ruidosos. Es capaz de aprender de pequeños conjuntos de datos, lo que es una ventaja aunque no sufre si el volumen de datos es muy alto en términos de número de muestras. No es el algoritmo ideal para problemas de alta dimensionalidad con un elevado número de atributos ya que utiliza tablas de frecuencia para extraer conocimiento de los datos y trata cada variable como categórica y, en caso de trabajar con variables numéricas, debe realizar algún tipo de transformación. La generalización discriminativa de estos modelos, puede interpretarse como un modelo lineal aditivo [142].

Naïve Bayes en el descubrimiento de fármacos

Este modelo ha sido utilizado en drug discovery para la predicción de posibles targets de fármacos. En concreto, en [143] desarrollan un modelo bayesiano que integra diferentes fuentes de datos como efectos secundarios conocidos, o datos de expresión génica consiguiendo un modelo con un 90 % de accuracy sobre más de 2.000 moléculas y desarrollando también la validación experimental del proceso de screening. En [108] predicen moléculas que sean multi target con AUC 80 % para tratamiento de VIH/HCV a partir de datos obtenidos de ChEMBL y generando para cada una de ellas dos tipos de descriptores (MACCS del inglés Molecular ACCess System y ECFP6 del inglés Extended Connectivity Fingerprint) y validando los resultados mediante técnicas de acopplamiendo molecular. A partir de 5125 interacciones conocidas con cuatro subtipos diferentes de proteínas (enzimas, canales iónicos, RAPGs y receptores nucleares) que obtuvieron de KEGG y DrugBank e

interacciones al azar de STITCH en [95] generaron un modelo para la predicción de interacciones ligando-target con una precisión del 95 %. En [136] generan modelos de predicción de fármacos según el sistema ATM (del inglés Anatomical Therapeutic Chemical) de la WHO utilizando desde STITCH y ChEMBL el conjunto de datos utilizado y calculando tres tipos diferentes de descriptores moleculares (basados en estructura, interacción entre compuestos y en interacciones con dianas similares) con un accuracy del 65 %. En [104] siguieron un diseño experimental basado en ML y docking molecular para la predicción de posibles inhibidores de Mttopo I, proteína target para la tuberculosis con valores de AUC de 74 % y realizando la validación *in vitro* de sus resultados computacionales.

Además, a partir de la generación de diferentes descriptores moleculares de un conjunto de compuestos que dañaron el hígado en [144], predicen un posible daño hepático por la medicina tradicional china con una precisión del 72 %. En [111] a partir de 50 compuestos ChEMBL se generaron modelos de predicción con un AUC del 80 % para inhibidores de IDO, calculando descriptores quimioinformáticos y validando los resultados mediante técnicas de acoplamiento molecular.

Este tipo de modelo también se ha utilizado en [145] para predecir la toxicidad de elementos químicos obtenidos de publicaciones anteriores en Pubmed con un AUC superior al 80 % y validando las mejores predicciones en laboratorio. O en [114] para predecir inhibidores de PI3K a partir de descriptores 3D obtenidos de ChEMBL y BindigDB con valores de AUC del 97 % y además con fase de validación *in vitro* y mediante técnicas de acoplamiento computacional.

Máquinas de vector soporte

Vapnik introdujo las SVM a finales de los años setenta [146]. Son una de las técnicas más utilizadas por su buen funcionamiento y su capacidad de generalización en dominios de alta dimensionalidad, sobre todo en bioinformática [147, 148, 149]. En ML se utilizan conjuntos de puntos en un determinado espacio para aprender una manera de tratar nuevas observaciones. Los métodos basados en kernel utilizan esos puntos para saber como son de similares las nuevas observaciones y tomar una decisión. Los kernel codifican y miden la semejanza entre objetos [150, 151, 152].

La implementación base trabaja con problemas de dos clases en la que los datos están separados por un hiperplano aunque existen implementaciones para problemas

de regresión o supervivencia. Siendo n la dimensión de los datos, un hiperplano es un subespacio afín de dimensión n-1 que divide el espacio en dos mitades que se corresponde con las entradas de las dos clases [152]. En problemas de clasificación, el objetivo del SVM es encontrar el hiperplano que separe los ejemplos positivos de los negativos. Este hiperplano separa los ejemplos orientado de tal manera que la distancia entre la frontera y el dato más cercano de cada clase sea máxima; los puntos más cercanos son utilizados para definir los márgenes, conocidos como vectores soporte [153].

Las técnicas de ML, los métodos basados en kernel y los SVM más específicamente, han demostrado ser excepcionalmente eficientes en problemas de clasificación de alta dimensionalidad [154, 155] debido a su habilidad para generalizar en dichos espacios.

La mayor parte de los conjuntos de datos complejos no son linealmente separables, por lo que las SVM introducen el concepto de kernel. Una función kernel es una función que mapea el espacio de entrada a una dimensión superior, donde los datos puedan ser linealmente separables. Sin embargo, la inclusión de estos kernels requiere de un nuevo nivel de parametrización, donde la función del kernel y sus parámetros deben ser cuidadosamente seleccionados.

En el caso en el que los datos no sean linealmente separables, una de las técnicas empleadas es el truco del kernel. La idea es muy sencilla y parte de lo mencionado en las dos secciones anteriores. Las SVM buscan el hiperplano que mejor separe los datos maximizando la capacidad de generalización del modelo. Si los datos no son linealmente separables, para intentar que lo sean, el espacio de entrada inicial se puede mapear en un espacio de mayor dimensionalidad [156] (recibe el nombre de espacio de características). En este nuevo espacio debe estar definido el producto escalar, espacio de Hilbert.

Por otro lado, son muchos los kernels que pueden cumplir con el teorema de Mercer [157]. Una clasificación sencilla fue propuesta por Smola en su tesis doctoral, separándolos en kernels locales y globales [158]. Para los kernels locales solamente aquellos datos que están cerca de otros, tienen influencia en los valores del kernel, mientras que por el contrario, en los kernels globales, todos los puntos, por muy distantes que estén unos de otros, tienen influencia en los valores del kernel. Un ejemplo de kernel global podría ser el polinomial y de local, el kernel de función de base radial. Los

kernels más utilizados son, con diferencia, los lineales, los polinomiales de grado q, los función de base radial y los sigmoidales.

SVM en el descubrimiento de fármacos.

Se han utilizado en [159] con valores de accuracy del 83.9 % para clasificar fármacos a partir de su categorización en KEGG. En [98] proponen un nuevo framework para la predicción de redes complejas de interacción a partir de matrices de interacción con valores de F1 de 80 %. Es posible también predecir la estabilidad en los microsomas del hígado humano calculando diferentes descriptores e índices químicos de 25 conjuntos de datos de ChEMBL con valores cercanos al 70 % en validación [84].

A partir de datos de expresión génica, se han utilizado para predecir el efecto de un fármaco sobre alguna línea tumoral, obteniendo previamente información sobre los genes que intervienen en la respuesta del fármaco en diferentes tumores [112]. En concreto utilizaron tres tipos de tumores y mediante *transfer learning* obtuvieron valores de AUC cercanos al 70 %. Otra aplicación clínica en [123] para generar modelos de predicción de neuropatía periférica después de quimioterapia a partir de descriptores para predecir la toxicidad de 95 compuestos aprobados por la FDA con un MCC del 80 %. O en la búsqueda de fármacos contra la malaria con *accurary* de 90 % en [106] y realizando una fase final de validación experimental.

En concreto en [127] se usaron 760 descriptores (calculados con CORINA) y se seleccionaron aquellos que más información contenían mediante un enfoque de selección de características con una precisión del 89 %. Las SVM también fueron utilizados siguiendo un esquema de reducción de dimensionalidad sobre descriptores 3D en [160] para la predicción de inhibidores de HDAC1. También en combinación con un algoritmo genético para predecir, a partir de descriptores 2D, los compuestos (usaron 499) que inhiben la proteína de membrana P-gp (cáncer) en [115], obtuvieron buenos resultados validados posteriormente mediante técnicas de *docking* molecular.

Un ejemplo de uso avanzado de SVM es el aprendizaje de kernel múltiple (MKL) donde se generan diferentes combinaciones lineales de SVM con diferentes parámetros o kernels que tratan de resolver el mismo problema. Esto además permite integrar diferentes conjuntos de datos heterogéneos disponibles aunque a coste de incrementar el coste computacional de generación de resultados. En [102] demuestran como un modelo MKL es capaz de predecir interacciones entre fármaco-objetivo

con valores de AUC de 90 % integrando diferentes conjuntos de datos disponibles de 1332 interacciones conocidas (matriz de interacción, efecto secundario, patología o secuencia).

Sin embargo, en problemas complejos y usados conjuntamente con una metaheurística como un algoritmo genético puede producirse, con datasets de reducido tamaño (136 fármacos) el sobreentrenamiento del modelo con resultados no adecuados en validación (PubChem), como sucede en [131] para la predicción de candidatos inhibidores de C1.

Modelos basados en árboles

Dado el relativo éxito de los modelos basados en árboles en diferentes campos por los buenos resultados que habitualmente logran, existen múltiples implementaciones diferentes. Una aproximación que destaca sobre el resto es la que hace uso de árboles de decisión. En concreto, un árbol de decisión es una estructura de tipo jerárquico compuesta de nodos y ramas que los unen. Aunque existen implementaciones para diferentes tipos de problemas como regresión, supervivencia o detección de outliers, destaca la aproximación usada para problemas de clasificación. Dentro de la estructura jerárquica de un árbol de decisión se pueden identificar nodos raíz, nodos internos o nodos terminales. El nodo raíz suele representarse en la parte superior del árbol sin ramas que lleguen hasta él y con una o varias ramas que parten desde él. Con respecto a los nodos internos estos tienen una rama que llega hasta ellos y dos o más que parten desde ellos hasta el siguiente nivel de la jerarquía. Finalmente los nodos terminales no tienen ramas que partan de ellos pues están en el último nivel de la jerarquía.

De los diferentes algoritmos que usan árboles de decisión destaca por encima de todos Random Forest [161]. Se trata de un meta-algoritmo que utiliza un conjunto de árboles de decisión (*ensemble*) para construir una solución al problema que pretende resolver mediante aproximaciones de *bagging* y *boosting*. El modo de funcionamiento supone que RF haga crecer el bosque y los árboles de decisión resuelven el problema de forma individual, aportando cada uno de ellos un voto en la resolución del problema. Como cada árbol puede estar explorando una parte diferente del espacio de soluciones, al final RF debe determinar la solución general al problema y lo hace considerando la mayoría de votos.

Al estar diseñado como una aproximación de *bagging* lo que hace RF es separar el conjunto de datos que está analizando en 1/3 para validación y 2/3 para entrenamiento de tal forma que cada árbol de decisión individual es capaz de determinar de forma interna el error de generalización que está cometiendo, es lo que se conoce como *out of bag error* (OOB) y los autores han demostrado que es equivalente al error que cometería el algoritmo si se utiliza con validación cruzada. Finalmente, como cada árbol de decisión está siendo entrenado con ejemplos y atributos diferentes es posible, a partir del OOB, calcular un valor de importancia de cada uno de los atributos, pudiendo descartar el resto y reduciendo la dimensionalidad del problema. Es por ello que este algoritmo es particularmente útil para problemas de muy alta dimensionalidad y con ruido.

Modelos basados en árboles en descubrimiento de fármacos.

Se trata de uno de los algoritmos más utilizados, independientemente del tipo de problema a resolver y, a pesar de que no es posible identificar un modelo como el mejor para cualquier tipo de problema, RF es sin duda uno de los mejores en términos de rendimiento, velocidad y capacidad de generalización.

RF ha sido utilizado en [162] con datos de *The Cancer Genome Atlas* (TCGA) para la búsqueda de posibles reposicionamientos de fármacos antitumorales aprobados por la FDA en cáncer de colon, identificando cuatro interacciones prometedoras GLTP - Nilotinib, PTPRN - Venetoclax, VEGFA - Venetoclax and FABP6 - Abemaciclib. A partir de 211,888 interacciones fármaco-proteína de BindingDB, siguiendo un esquema de reducción de dimensionalidad mRMR (*max relevance and min redundancy*) en [163], fueron capaces de predecir la interacción entre compuesto y proteína con un accuracy superior al 90 % a partir de los descriptores generados con Open Babel y los *enrichment scores* de cada proteína a partir de GO and KEGG.

Es posible predecir además la interacción entre un compuesto y una ruta molecular a partir de datos de cMap (dispone de 7056 perfiles de microarrays de 5 líneas celulares, tratadas con 1309 compuestos diferentes), para ello en [164] calcularon descriptores y propusieron un nuevo modelo basado en árboles con resultados de AUC superiores al 90 %. Es más, es posible predecir la interacción de un determinado fármaco con moléculas de la membrana plasmática de las células RAPG mediante descriptores a partir de 1860 pares RAPG-fármaco con un accuracy del 87 % como en [165].

En [116] se generaron modelos de predicción para probar diferentes anticuerpos sobre líneas celulares tumorales cuantificando niveles de proliferación y apoptosis a partir de las variables seleccionadas por un RF para comprobar aquellas que mejor describen el fenotipo inducido por cada par anticuerpo-dosis.

También es posible calcular descriptores que no sean moleculares, sino *proteochemometrics* (512 descriptores) para predecir posibles inhibidores para SGLT1 en diabetes tipo II con un valor de MCC de 48 % como en [129].

Es más, la robustez del modelo y su gran rendimiento en tareas de predicción ha hecho posible que se utilice en la búsqueda de sinergías con varios fármacos en diferentes líneas celulares. En [121] predicen sinergias entre dos fármacos y una línea celular utilizando información genómica, dianas de fármacos e información farmacológica con un total de 583 combinaciones de fármacos para 31 tipos de líneas celulares tumorales. Partiendo de datos de expresión y mutación de genes en rutas moleculares relacionadas con cáncer, identificaron que los modelos basados en árboles eran los que mejor predecían el score de la sinergia. Incluso convirtiendo el problema en uno de clasificación mantuvieron valores de F1 de 95.4 %. Cabe mencionar un esfuerzo conjunto de múltiples investigadores que surge como *DREAM challenge* en el que a partir de 11,576 experimentos reportados por AstraZeneca de 910 combinaciones de fármacos sobre 85 líneas celulares de cáncer caracterizadas molecularmente (*expression, copy number variation, methylation, mutations*) [166] 160 equipos internacionales trataron de predecir las mejores sinergias entre pares de fármacos y biomarcadores para lo que se usaron diferentes aproximaciones. El equipo ganador utilizó un RF.

Redes de Neuronales Artificiales

Para poder describir las RNA, primero debemos mencionar la neurona artificial, lo que supondrá un elemento funcional de la red la cual recibe información de otros elementos y de alguna manera los procesa para acabar proporcionando una salida que puede ser procesada por otros elementos. Al igual que en la naturaleza, las neuronas artificiales se pueden comunicar unas con otras y las conexiones de las neuronas se representan mediante pesos que no es más que un valor que trata de expresar la fuerza sináptica de esa conexión entre dos neuronas. Cuando se evalúa una neurona artificial o elemento de procesado (EP) lo primero a tener en cuenta es el valor neto que representa el conjunto de todas las fuerzas que reciben. Una

vez calculado el valor neto se le aplica una función de activación para determinar la salida del EP. A partir del concepto de una neurona artificial, se pueden interconectar varias neuronas para formar una red donde las salidas de una neurona puede ser la entrada de otra neurona. Es necesario entender que la RNA necesita tener unos nodos de entrada que son los que obtienen la información desde el exterior, se dice que estas neuronas son la capa de la entrada de la red. La red también necesita unos nodos de salida, que están en la capa oculta, que transfieren el resultado de la RNA. El resto de nodos se conocen como nodos ocultos que transmiten información entre neuronas de la red y se agrupan en una o varias capas ocultas.

Muchos investigadores han definido parámetros específicos de la red como su topología, las funciones de activación que modifican la salida de la red, para obtener distintos tipos de RNA. Es importante mencionar que lo importante de una RNA no es solo la topología de interconexión de las neuronas y sus funciones de activación, sino que una parte fundamental es la relevancia de cada una de las conexiones de la red. Estos valores se obtienen en una fase de entrenamiento, que dependiendo del tipo de red puede ser supervisado, no supervisado y por refuerzo. Es necesario tener en cuenta que el entrenamiento de una red es un proceso costoso en tiempo, obtener la salida de una RNA supone evaluar todas las neuronas que conforman la red y en entrenamiento el proceso es iterativo. Por esta razón las RNA suelen tener un número no muy elevado de neuronas, lo que implica que todo el conocimiento está en la maraña de conexiones sin poder conocer que acción realiza cada parte, motivo por el que se considera que una RNA es una caja negra; se conoce las entradas de la red, la salida que produce, pero no que sucede en su interior.

Sin embargo, hay un caso particular de red neuronal, el *Deep Learning* (DL). Por lo general suele tener un número muy elevado de capas de neuronas conectadas entre si. El concepto detrás del DL es como se procesa la información. En este caso la información se procesa de forma jerárquica. Es decir, en DL cada capa de neuronas trata de obtener una representación más significativa de los datos. Las primeras capas extraen un nivel de características bajo, pero según se va profundizando en la red, las funciones simples se van combinando para poder representar relaciones más complejas.

El auge de las RNA y de los modelos de DL hablando de forma general surge a partir de la explosión computacional derivada del uso generalizado de las GPUs. Este salto cualitativo y cuantitativo supuso una reducción de meses/años en el entrenamiento de modelos complejos con miles de capas internas en la jerarquía a minutos o

segundos. Es más, supuso pasar de modelos de juguete a modelos que realmente se asemejan más a la estructura jerárquica biológica del cerebro humano real en cuanto a número de neuronas y capas. Desafortunadamente, aún es necesario esperar para ver modelos equivalentes en número de conexiones (donde realmente reside el conocimiento) al modelo biológico.

Sin embargo, a pesar del tremendo incremento en el uso de este tipo de técnicas algunas de las carencias asociadas al modelo siguen sin solucionarse: son cajas negras que a partir de una entrada emiten una salida pero no explican como han alcanzado esa conclusión, necesitan volúmenes de datos muy grandes (parcialmente solucionado con *transfer learning* y la posibilidad de reutilizar modelos entrenados para problemas parecidos) y la elección de alguna de las jerarquías de modelos existentes para un nuevo problema diferente no es tarea sencilla.

RNA en el descubrimiento de fármacos

Uno de los primeros artículos que utilizó RNA para la predicción de fármacos es [138], entrenaron una RNA y modelos basados en árboles a partir de datos de CMC (*Comprehensive Medicinal Chemistry*) y ACD (*Available Chemicals directory*), para fármacos y no fármacos, respectivamente. Para cada uno de los compuestos utilizaron descriptores 1D que contenían información sobre la molécula entera (peso molecular, número de enlaces de hidrógeno, etc.) y 2D que contenían información sobre la presencia o ausencia de grupos funcionales dentro de la molécula. Los mejores resultados fueron obtenidos con una RNA y ambos tipos de descriptores con un accuracy del 89 %.

A partir de 1003 sustancias químicas procedentes del *Carinogenic Potency Database* en [88] fueron capaces de predecir la carcinogénesis temprana de compuestos propuestos a ser fármacos para los que calculan seis tipos diferentes de descriptores con un modelo *deep learning* y un accuracy del 86 %. Del mismo modo, es posible iniciar una fase experimental en laboratorio para generar un conjunto (2130 compuestos) nuevo de posibles fármacos de interés en cardiotoxicidad y calcular con DRAGON 3456 descriptores de cada uno de ellos e incluir el análisis en un esquema de selección de características para finalizar con AUC de 76 %. A partir de interacciones compuesto-proteína positiva obtenidas de STITCH se utilizó un modelo *deep learning* con descriptores moleculares de los compuestos obtenidos de PubChem

y de proteinas con un AUC superior al 95 %, tomando para ello como interacciones negativas pares generados de manera aleatoria.

Además, como se comentó previamente es posible generar modelos de predicción DL utilizando *transfer learning* para predecir interacciones ligando proteína con tres tipos de descriptores diferentes y un AUC superior al 90 % en Tox21 [94].

Es posible trabajar con información en diferentes niveles biológicos, por ejemplo en [90] utilizan DL para predecir si es posible la unión fármaco-proteína utilizando datos transcriptómicos con un accuracy del 95 %. En general la mayoría de los trabajos revisados se centran en la predicción de la función de un determinado fármaco pero es posible aún avanzar en el campo en cuanto a la predicción de interación entre ligando y proteína con CNN (*Convolutional Neural Network*), mejorando los resultados obtenidos con Vina, software del estado del arte para experimentos de Docking a partir de los SMILES y los FASTA. El auge de las CNN en el análisis de imagen también ha permitido trabajos en los que, a partir de diferentes líneas celulares tumorales se generaron imágenes de inmunohistoquímica de cada una de ellas y se generaron modelos de clasificación para líneas normales y tumorales. Esta aproximación [122] sería, por lo tanto, de utilidad en el descubrimiento de fármacos. Existen también trabajos como [140] en el que este tipo de modelos se usan para la búsqueda de nuevas moléculas con posibilidades de funcionar como antiobióticos que aún no han alcanzado el estado del arte en generación de descriptores basados en grafos, pero que pueden suponer un avance.

En general la parte más compleja de este modelo es la obtención de datasets de suficiente tamaño y encontrar los mejores hiperparámetros para análisis de fármacos [167]. Incluso es necesario validar aproximaciones como las de *dropout* para establecer si mejoran el rendimiento predictivo en análisis quimioinformático que estudian la interacción fármaco-proteína como en [168]. Como se indicó previamente, el auge de estas técnicas hace que se apliquen sobre dominios nuevos, deba estudiarse cuidadosamente su funcionamiento y sea necesario adaptar el modelo adecuadamente. Es más, en el campo de descubrimiento de fármacos, se emplean para suplir las limitaciones de los métodos convencionales. Incluso ha permitido probar fármacos que se diseñaron con un propósito específico para otros propósitos [103], son las técnicas conocidas como de reposicionamiento y del que hemos mencionado algún ejemplo desarrollado con RF. El diseño de fármacos basado en estructuras se ha beneficiado por ser mucho más rápido y eficiente en coste que el diseño tradicional [169].

4 Solución propuesta

La investigación de fármacos busca desarrollar terapias más selectivas y específicas que puedan reducir los efectos secundarios y mejorar la tolerabilidad de los tratamientos. En este capítulo se presentan las dos aproximaciones seguidas en el desarrollo de la tesis doctoral: modelos quimioinformáticos para el descubrimiento y diseño de fármacos y modelos de síntesis enantioselectiva por reacción de α-amidoalquilación catalizada por ácidos de Brønsted para la producción de nuevos fármacos. Ambos modelos permiten filtrar rápidamente compuestos y enfocar los recursos en aquellos con mayor probabilidad de ser activos biológicamente en CCR.

Modelos quimioinformáticos

Los modelos quimioinformáticos son capaces de predecir diferentes salidas, por ejemplo la actividad, propiedades o reactividad química en sistemas moleculares complejos tales como reacciones metabólicas [170], nanopartículas [171], etc. Específicamente, la predicción de la reactividad química de reacciones complejas en la síntesis orgánica es un objetivo de gran importancia tanto para la investigación básica como para la industria química.

Los modelos quimioinformáticos pueden ser muy útiles en la predicción de resultados en reacciones estereoselectivas [172]. Sigman *et al.* escribieron algunos de los trabajos pioneros para la predicción de proporciones enantioméricas de productos [173, 174, 175]. Más recientemente, se han aplicado metodologías de quimioinformática para predecir la enantioselectividad de diferentes tipos de reacciones, algunas de estas reacciones son alquilación [176, 177], alilación [178], propargilación [179, 180], carbolitiación intramolecular [181], reacciones de acoplamiento deshidrogenativo tipo Heck C-C y C-N [182, 183, 184, 185], reacciones en cascada Heck-Heck [186], ciclopropanación asimétrica de alquenos catalizada por cobre [187] y reacción de Henry [188].

Aunque se han sintetizado y probado varios compuestos con actividad frente al CRC, la posibilidad de encontrar un fármaco eficaz es todavía demasiado baja. Además,

es un proceso costoso que requiere de grandes inversiones económicas y de tiempo [189]. Es, por lo tanto, un campo abierto que se aborda en la presente tesis doctoral.

Para ello, se propone un modelo que, como se explicó en el capítulo 2 de Fundamentos, parte de la caracterización a nivel numérico del problema. En este caso, diferentes descriptores moleculares que serán calculados y analizados mediante diferentes técnicas de ML de las comentadas en el capítulo de Estado de la Cuestión: modelos basados en árboles, máquinas de vector soporte, redes de neuronas artificiales y modelos lineales.

En concreto, se propone el uso de descriptores moleculares de cadenas de Markov (MCDs del inglés *Markov Chain Molecular Descriptors*). Se trata de un conjunto de descriptores ampliamente utilizados para resolver problemas quimioinformáticos. Hay diferentes tipos de MCDs, por citar alguno de los más relevantes, Markov Means o Promedios Markov, Entropía de Shannon o Markov Moments [190]. Sin embargo, se propone también el uso de otros tipos de índices que no se habían usado previamente para el cálculo de descriptores, como por ejemplo, Markov Singular Values o Autovalores de Markov [191]. Si bien es cierto que han sido utilizados en problemas quimioinformáticos previamente, hasta la fecha, no se han encontrado reportes del uso de Autovalores de Matrices de Markov para el cálculo de descriptores propiamente.

Además, el cálculo de los MCDs se realiza a menudo utilizando un software específico que no está disponible para usuarios generales y, al inicio del desarrollo de esta tesis doctoral, no existía ninguna biblioteca pública disponible para el cálculo de los mismos, siendo todas las opciones de pago por licencia. Este hecho limita la disponibilidad de procedimientos generales de quimioinformática para investigadores de síntesis orgánica que utilizan este tipo de modelos y, en cierta manera, no aporta luz sobre el cálculo de los descriptores. Por lo tanto, el desarrollo de nuevas herramientas de acceso libre para el cálculo de descriptores moleculares en general y específicamente de MCD es un área prometedora de investigación ya que el acceso libre a estas herramientas puede promover el desarrollo de modelos quimioinformáticos para áreas de investigación menos exploradas con este tipo de técnica.

Por todo ello, como parte del modelo y para la obtención de compuestos biológicamente activos frente al cáncer de colon se desarrolla y se pone a disposición pública el paquete RMarkovTI (disponible en GitHub en: `https://github.com/muntisa/RMarkovTI`) con el objetivo de implementar, en el lenguaje de programación R, un

paquete que permita calcular los siguientes descriptores: promedios de Markov y autovalores de Markov.

Ambos tipos utilizan por un lado la topología de grafos moleculares con propiedades físicas de 4 átomos para codificar la información molecular y, por otro lado, la teoría de cadenas de Markov para incluir las interacciones atómicas intramoleculares. El algoritmo se deriva de un software privado existente, MInD-Prot, pero difieren los pesos atómicos y se amplian los descriptores, además de hacerlo completamente abierto.

Las propiedades atómicas incluidas en este caso son las siguientes: número de electrones de valencia (Zv), radio atómico de Van der Waals (Rvdw), radio covalente (Rcov), masa atómica (m), volumen de Van der Waals (Vvdw), electronegatividad de Sanderson (SAe), polarizabilidad atómica (aPolar), potencial de ionización (IP) y afinidad electrónica (EA).

El cálculo de descriptores moleculares se realiza para cada propiedad y tipo de átomo. Todos los descriptores se calculan para seis tipos de átomos: C saturado (Csat), C insaturado (Cuns), halógeno (Hal), heteroátomos (Het), heteroátomos pero no halógenos (HetNoX) o todos los átomos (All). De esta forma, se calculan 54 descriptores (9 propiedades atómicas x 6 tipos de átomos).

El modelo quimioinformático final propuesto, además de disponer de un paquete de R abierto para el cálculo de los descriptores que es usable, explicable y ampliable en cuanto a la incorporación de nuevos descriptores, implementará el mejor modelo de ML generado en una aplicación web para su validación por los investigadores del campo, proporcionando así una rápida y potente herramienta que permite el diseño de modelos de regresión quimioinformáticos para la predicción del nivel de respuesta biológicamente activa en cáncer de colon. En el siguiente capítulo se mostrará la fase de experimentación, resultados y la herramienta web desplegada.

Modelos de síntesis enantioselectiva por reacción de α-amidoalquilación catalizada por ácidos de Brønsted

La búsqueda de compuestos biológicamente activos frente al CCR implica identificar y desarrollar moléculas que muestren actividad antitumoral específica contra las células cancerígenas del colon y recto. Los estudios de cribado y la optimización de compuestos pueden incluir ensayos *in vitro* e *in vivo* para evaluar la citotoxicidad, la inhibición del crecimiento celular, la inducción de apoptosis y otros parámetros relevantes. En este contexto, los compuestos obtenidos mediante la síntesis enantioselectiva por reacciones de α-amidoalquilación intermolecular catalizada por ácidos de Brønsted (SERAICAB) pueden representar una fuente de moléculas con propiedades específicas que pueden influir en su actividad biológica contra el CCR, y que se desea estudiar.

La reacción de formación de enlaces carbono-carbono asimétricos mediante SERAICAB se utiliza para obtener compuestos quirales enantioméricamente puros, que son moléculas quirales con alta selectividad y enriquecidas. Estos compuestos quirales pueden tener propiedades físico-químicas y estereoquímicas particulares que los hacen adecuados para la interacción selectiva con objetivos biológicos específicos. En el contexto del CCR, la síntesis enantioselectiva puede permitir la obtención de compuestos quirales que sean más efectivos y selectivos en la inhibición de las vías moleculares relevantes para la proliferación y la supervivencia de las células cancerígenas en el colon y recto. Al sintetizar compuestos quirales con alta enantioselectividad y evaluar su actividad contra células cancerosas, se puede llegar a identificar estructuras químicas que presenten una mayor actividad anticancerígena y una menor toxicidad para las células normales, lo que es fundamental en el desarrollo de agentes terapéuticos más efectivos y selectivos para cualquier tipo de tratamiento.

Al igual que sucedió con el modelo previo, la aplicación de modelos de ML necesita la caracterización del problema de laboratorio numéricamente. En este caso, como se comentó en el capítulo 2 de Fundamentos, la Teoría de la Perturbación (TP) provee un marco teórico para analizar y predecir los efectos de pequeñas perturbaciones en sistemas moleculares.

Además, la TP puede ser aplicada para analizar la influencia de perturbaciones externas, tanto estructurales como condicionantes de la reacción, como temperatura, presión o presencia de disolventes. Estas perturbaciones pueden influir significativamente en la selectividad y la eficiencia de este tipo de reacciones de síntesis enantioselectiva. En este sentido y por todo ello, la TP se revela como una herramienta valiosa para analizar y predecir los efectos de las perturbaciones introducidas en SERAICAB como catalizadores.

Esta combinación de la TP con los modelos de síntesis enantioselectiva proporciona un enfoque integral para el diseño racional de nuevos compuestos quirales y la optimización de dichas reacciones químicas asimétricas.

En la SERAICAB el catalizador ácido juega un papel fundamental al facilitar la formación de enlaces carbono-carbono asimétricos y conducir a la obtención de productos quirales. El mecanismo de reacción implica la generación de un par iónico formado por la base conjugada/*N*-aciliminio que puede ser atrapado con un nucleófilo enantioselectivamente, dando lugar al producto deseado. Sin embargo, resulta algo complicado poder predecir la enantioselectividad de estas reacciones en las que se emplean ácidos fosfóricos quirales porque depende de la naturaleza del sustrato, nucleófilo y catalizador, así como de las condiciones experimentales de reacción.

En este sentido, la aportación de la tesis doctoral se focalizada en el uso de la quimioinformática para la exploración del espacio químico de este tipo de reacciones en modelos *in silico*, reducción de tiempos y costes de las reacciones generadas en el *wet lab*.

Descriptores moleculares para el estudio de la reactividad química.

Los descritores utilizados por el modelo propuesto en la presente tesis doctoral contienen la siguiente información sobre condicionantes de la reacción: configuración del catalizador, aditivo (cloruro de tetrametil silano), temperatura, tiempo, dipolo del disolvente, carga del catalizador y agente secante.

Por otro lado, los factores estructurales moleculares son: molécula q^{th} en la reacción, sustrato, producto, disolvente, catalizador y nucleófilo. En la Tabla 4.1 se resumen

las variables de salida frente a las variables de entrada utilizadas en la generación de los diferentes modelos de ML.

Tab. 4.1.: Variables de salida frente a variables de entrada utilizadas en el modelo.

Tipo de variable	Variable	Detalles
Salida	$ee(\,\%)\,[R_{\mathrm{cat}}]$	EE usando el catalizador R
Entrada	$\mathrm{f_0 = (R = 1/S = -1)cat}$	Configuración del catalizador
Reacción	$\mathrm{f_1 = TMSCl(eq)}$	Aditivo TMSCl
Operación	$\mathrm{f_2 = T(K)}$	Temperatura
Variables	$\mathrm{f_3 = t(h)}$	Tiempo de reacción
	$\mathrm{f_4 = Ds}$	Dipoleo del solvente
	$\mathrm{f_5 = Load(\,\%)}$	Dipolo del catalizador
	$\mathrm{f_6 = Da}$	Agente secante
Entrada	$\mathrm{fV(w, g,}$ Sub $)$	Sub = Sustrato $(\mathrm{q} = 0)$
Químico	$\mathrm{SV(w, g,}$ Prod $)$	Prod = Producto $(\mathrm{q} = 1)$
Estructura	$\mathrm{SV(w, g,}$ Solv $)$	Solv = Disolvente $(\mathrm{q} = 2)$
Variables	$\mathrm{SV(w, g,}$ Cat $)$	Cat = Catalizador $(\mathrm{q} = 3)$
	$\mathrm{SV(w, g,}$ Nuc $)$	Nuc = Nucleófilo $(\mathrm{q} = 4)$

$\mathrm{SV\,(w, g, m_q)}$ = Max Singular Values $(\mathrm{SV_{máx}})$ for molecule $(\mathrm{m_q})$ con grupo químico g y papel q-ésimo (sustrato, producto, disolvente, etc.) en la reacción.

El modelo de síntesis enantioselectiva por reacción de α-amidoalquilación catalizada por ácidos de Brønsted final propuesto se implementará, del mismo modo, en una aplicación web para su validación por los investigadores del campo, proporcionando así una rápida y potente herramienta que permite el diseño de modelos de ML lineales aditivos para la predicción del nivel de respuesta biológicamente activa en cáncer de colon. En este caso, se desarrollará una aproximación experimental incremental generando modelos cada vez más complejos a partir del inicial. En el siguiente capítulo se mostrará la fase de experimentación, resultados y la herramienta web desplegada.

5 Pruebas y resultados

En este capítulo se expondrán y se analizarán los resultados obtenidos para los dos modelos propuestos. Se estructuran los siguientes capítulos de la siguiente manera, primero se indicarán los datos utilizados y su procedencia, después se comentará la experimentación realizada y se finalizará con los resultados obtenidos.

Experimentación y resultados de modelos quimioinformáticos.

Datos

Para la generación del modelo quimioinformático se obtuvieron del repositorio ChEMBL 5644 ensayos preclínicos de compuestos activos frente al CCR. El resultado de cada ensayo se expresó mediante un parámetro experimental ε_{ij}, usado para cuantificar la actividad biológica de las i^{th} moléculas (m_i) sobre las j^{th} diana.

Los valores de ε_{ij} dependen de la estructura del fármaco y de una serie de condiciones del entorno que delimitan las características del ensayo $c_j = (c_0, c_1, c_2, ...c_n)$. En donde c_0 = la actividad biológica de v_{ij} (inhibición, GI_{50}, IC_{50}, etc), c_1 = proteína diana, c_2 = organismo del ensayo, etc. Se discretizan los valores como sigue:

- $f(v_{ij})_{obs} = 1$ apunta a un fuerte efecto del compuesto sobre la diana. Esto se produce cuando v_{ij} >nivel deseado de actividad biológica $d(c_0) = 1$ y cuando v_{ij} <nivel deseado $d(c_0) = -1$, en función del *cutoff*.

- $f(v_{ij})_{obs} = 0$ indica que no hay efecto del compuesto sobre la diana y se produce en los casos contrarios a los anteriores.

- El nivel de actividad $d(c_0) = 1$ o -1, indica que el parámetro medido aumenta o disminuye directamente con un efecto biológico deseado o no deseado respectivamente.

Experimentación

Para realizar la experimentación del modelo propuesto en la tesis doctoral y como se comentó con anterioridad, se ha desarrollado un paquete de R llamado RMarkovTI (disponible en GitHub en: `https://github.com/muntisa/RMarkovTI`) con el objetivo de implementar, en el lenguaje de programación R, un paquete que permita calcular los siguientes descriptores: Markov Mean Properties (MMPs) y *Markov Singular Values of Transition Probabilities* (MMSV).

A partir de los descriptores generados por RMarkovTI se han aplicado diferentes modelos de ML para regresión. En concreto se han utilizado redes de neuronas artificiales, máquinas de vector soporte, random forest y mínimos cuadrados parciales (PLS). Los resultados obtenidos se mostrarán a continuación.

Resultados

Se utilizaron como variables de entrada Valores SV_{max} de todas las moléculas implicadas en la reacción (sustrato, nucleófilo, catalizador y producto). Los resultados, en R^2 y RMSE como promedio de 10 divisiones aleatorias de datos, se muestran en la Tabla 5.5. No se han eliminado características correlacionadas. Las diez divisiones aleatorias de los datos (75 % entrenamiento y 25 % de prueba) determinaron el mejor modelo y se finalizó la experimentación con un conjunto de 100 ejecuciones de aleatorización para la validación del mejor modelo, un RF. El fin de la última validación es asegurar que el modelo se desarrolló sin ningún tipo de sesgo debido a los datos utilizados.

A continuación para explorar nuevos modelos de ML y, como se muestra en la Figura 5.1, se ha utilizado el paquete RRegrs [192] para probar modelos Lasso, regresión lineal múltiple (LM) y el modelo lineal generalizado (GLM). Como se puede observar, los resultados obtenidos no mejoran a los mostrados en la tabla anterior.

Los métodos PLS y RNA tienen valores de R^2 inferiores o similares (R^2 = 0,775 y R^2 = 0,829) al valor del método LM (R^2 = 0,828) descrito anteriormente durante la

Tab. 5.1.: Resultados de RRegrs para el exceso enantiomérico utilizando la predicción del catalizador R.

Método	Entrenamiento		Test	
	R	RMSE	R	RMSE
RF	0.907	0.101	0.926	0.093
SVM	0.868	0.122	0.866	0.128
NN	0.849	0.132	0.829	0.143
PLS	0.801	0.155	0.775	0.167

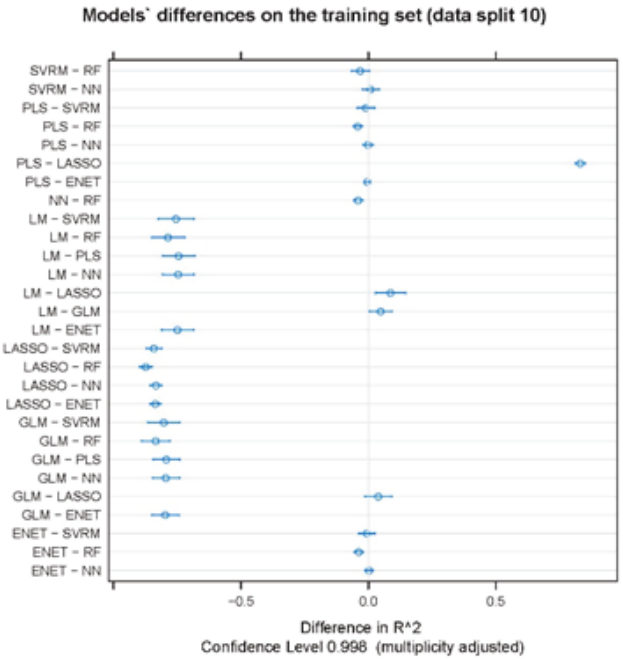

Fig. 5.1.: Diferencias de rendimiento obtenidas por los modelos para el conjunto de entrenamiento (10 divisiones aleatorias). Se presentan las diferencias de R^2 entre los modelos de ML. Se representa el rendimiento medio con límites de confianza de dos caras, obtenido mediante la prueba T de Student con corrección de multiplicidad de Bonferroni.

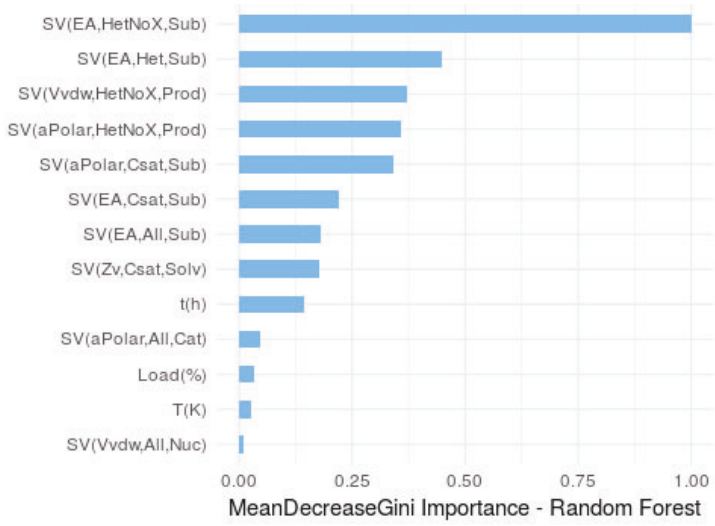

Fig. 5.2.: Media de la disminución de la importancia de Gini de las principales variables seleccionadas por RF. Este valor oscila entre cero y uno para comprender de forma sencilla la influencia de cada variable en el modelo final.

fase de prueba. El método SVM-radial (R^2 = 0,866) tiene un valor de R^2 ligeramente superior al del método LM. El mejor resultado en la prueba lo obtuvo el método RF con R^2 = 0,926. Este valor implica que el modelo explica más del 90 % de la varianza, lo que supone un 10 % más que el método LM.

Además, se encontraron tres modelos que superaron a LM: RF, SVM y RNA. En la Figura 5.2, se resumen las principales variables del mejor modelo (RF) mediante la denominada importancia de Gini. Este índice puede calcularse para evaluar la importancia de cada variable en el modelo final.

Comparación con otros modelos de la literatura

Bediaga *et al.* y Speck-Planche y Cordeiro *et al.* ya publicaron anteriormente diferentes modelos para el descubrimiento de compuestos biológicamente activos en diferentes tipos de cáncer [70-78]. En la Tabla 5.2 se resumen los resultados obtenidos utilizando estos modelos comparativos.

Todos estos modelos tienen en cuenta perturbaciones (variaciones) en la estructura del fármaco y en múltiples condiciones de ensayo simultáneamente, como las proteínas diana, líneas celulares, organismos, etc. Se excluyen los modelos clásicos de la comparación porque sólo son útiles para un conjunto de condiciones específicas.

Debido a las diferencias en el conjunto de datos la comparación se centró únicamente en los modelos y no en el rendimiento de los descriptores. Podemos observar que casi todos los modelos se centran en otros tipos de cáncer. Sin embargo, el modelo publicado por Speck-Planche *et al.* en 2012 es específico para compuestos activos en el CCR. Se puede observar en la Tabla 5.2 que nuestro modelo tiene valores más bajos de Sn(%) y Sp(%) pero es capaz de ajustarse a un conjunto de datos de entrenamiento más de tres veces mayor que el modelo anterior, 4233 frente a 1237 ensayos preclínicos. En este sentido, se espera que el presente modelo sea capaz de predecir una gama más amplia de compuestos y ensayos preclínicos gracias al conjunto de datos más amplio y actualizado utilizado.

Tab. 5.2.: **Comparación con otros modelos**

Tipo de cáncer	Cancers	PT[a]	ML[b]	NV[b]	Casos [c]	Sn(%)[d]	Sp(%)[d]	Refs.
Colorrectal	1	MMA	LDA	mit	4233(entrenamiento)	> 70	> 80	This work
Colorrectal	1	MA	LDA	> 10	1237(entrenamiento)	> 90	> 90	[74]
Mama	1	MA	LDA	> 10	24285 (total)	> 90	> 90	[70 71]
Vejiga	1	MA	LDA	n.a.	n.a.	> 90	> 90	[72]
Cerebral	1	MA	LDA	n.a.	n.a.	> 90	> 90	[73]
Mama	1	MA	LDA	> 10	2272 (total)	> 85	> 95	[75]
Próstata	1	MA	LDA	> 10	1250 (train)	> 85	> 95	[78]

[a] operadores PT utilizados, MA = Media móvil, MMA = Media móvil multicondición. [b] Método ML utilizado y NV = Número de variables de entrada, n.d. = no disponible para los autores de este trabajo. [c] Número total de casos en las series de entrenamiento y/o validación. ^{mathrmd} Valores aproximados de las series de entrenamiento y validación..

Experimentación y resultados de modelos de síntesis enantioselectiva por reacción de α-amidoalquilación catalizada por ácidos de Brønsted

Datos

Para la generación del modelo se ha utilizado un conjunto de datos de referencia amplio de reacciones de α-amidoalquilación. Este conjunto de datos incluye reacciones de α-amidoalquilación de múltiples procedencias [193, 194, 195, 196, 197, 198, 199]. Se trata de reacciones relacionadas con diferentes tipos de sustratos (hidroxilactamas cíclicas y bicíclicas), nucleófilos (enamidas, indoles, etc.) y catalizadores quirales (ácidos fosfóricos, fosforamidas, etc.), en diferentes condiciones experimentales de reacción.

La reacción de α-amidoalquilación [200, 201, 202, 203] es uno de los métodos más atractivos para la formación estereoselectiva de enlaces C-C y se ha utilizado ampliamente en la síntesis de una gran variedad de moléculas orgánicas complejas, incluyendo productos naturales y farmacéuticos [204, 194, 205]. Este método posee varias ventajas distintivas. Se ha informado de que la reacción tiene una amplia gama de nucleófilos y sustratos susceptibles a reaccionar. Además, la reacción de iones *N*-aciliminio es altamente diastereoselectiva [206, 207, 208] cuando están implicados cíclicos y bicíclicos, que pueden generarse *in situ* a partir de las hidroxilactamasas correspondientes, utilizando ácidos próticos y ácidos de Lewis.

Esta estrategia es aplicable a la construcción de estereocentros terciarios y cuaternarios de forma asimétrica [209]. Estas variantes enantioselectivas [210, 211, 212] se han desarrollado utilizando principalmente ácidos de Brønsted quirales como ácidos fosfóricos y fosforamidas [213, 214, 214, 215], así como ureas y tioureas [216, 217, 218, 219] como catalizadores. Además, el procedimiento funciona bien con enamidas aromáticas y heteroaromáticos (reacciones de Friedel-Crafts) [220, 221, 222, 223], éteres de sililenol [224], etc.

La química computacional ha ayudado a comprender el mecanismo de estas reacciones de α-amidoalquilación. En la Figura, se representa la idea general que subyace a

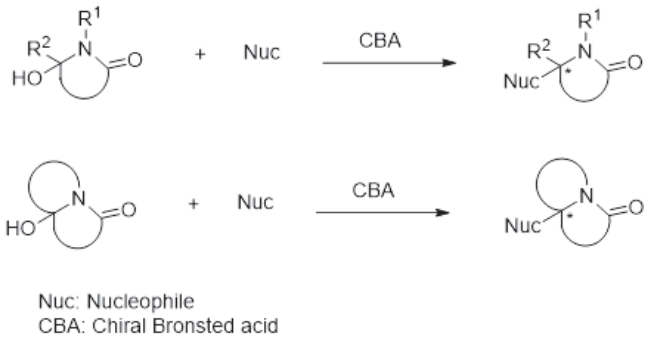

Fig. 5.3.: Reacciones catalíticas de α-amidoalquilación intermolecular enantioselectiva.

las reacciones catalíticas enantioselectivas intermoleculares de α-amidoalquilación estudiadas.

Finalmente, 332 reacciones (324 se obtienen de la literatura y 8 son estudiadas en la tesis doctoral por primera vez) conforman el conjunto de datos de entrada del modelo. Las reacciones pueden agruparse en 34 familias, de acuerdo a diferencias estructurales, patrones de los substratos, nucleófilos o catalizadores como se observa en la Figura 5.4.

A partir de las reacciones, se ha estudiado el parámetro de exceso enantiomérico $eeR(\,\%)_{obs}$. El valor $eeR(\,\%)_{obs}$ cuantifica el exceso enantiomérico obtenido utilizando un catalizador (*R*). Por otra parte, Sign(Cat*R*) = 1 para todas las reacciones llevadas a cabo con un catalizador (*R*), independientemente del producto obtenido, mientras que el signo se cambió a Sign(Cat*R*) = -1 para las reacciones llevadas a cabo con un catalizador (*S*). De este modo, todas las reacciones se estudiaron como si se hubieran realizado con un catalizador (*R*). Por consiguiente, los valores de $ee_R(\,\%)_{obs}$ y $eeR(\,\%)_{calc}$ predichos no dependen de la quiralidad del catalizador original.

Además de los descriptores moleculares, se utilizaron diferentes variables de condiciones de reacción $V_k(c_{qi})$ como variables de entrada para cuantificar una k^{th} propiedad (k = 1, 2, 3) relacionada con una condición de reacción general (c_q). En este conjunto de datos, las variables consideradas para las i^{th} reacciones de consulta fueron:

- $V_1(c_{qi})$ = T(^{o}C) = temperatura.
- $V_2(c_{qi})$ = t(h) = tiempo de reacción.
- $V_3(c_{qi})$ = L(%) = carga de catalizador.

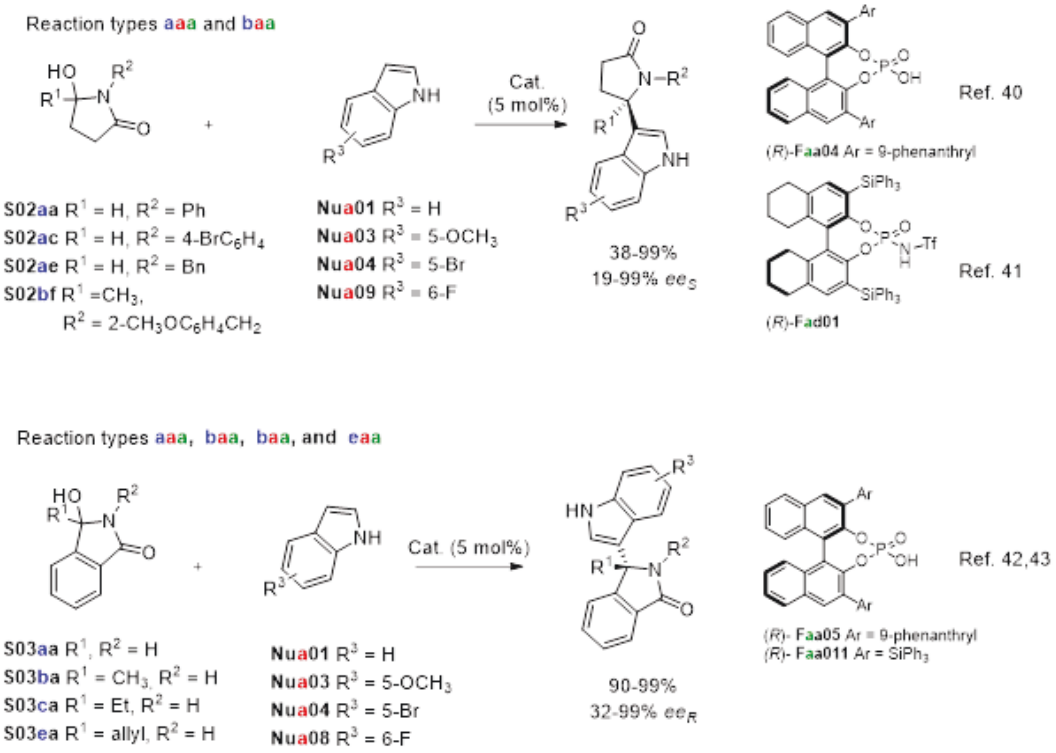

Fig. 5.4.: Ejemplos seleccionados de tipos de reacción enantioselectiva intermolecular de α-amidoalquilación.

Por analogía, los valores de las variables consideradas para cada th reacción de referencia fueron:

- $V_1(c_{rj})$ = T(^{o}C) = temperatura.
- $V_2(c_{rj})$ = t(h) = tiempo de reacción.
- $V_3(c_{rj})$ = L(%) = carga de catalizador.

Concretando, el conjunto de datos utilizado incluyó 332 reacciones que en total contienen:

- 55 sustratos diferentes (hidroxilactamas cíclicas y bicíclicas).
- 53 nucleófilos (enamidas, indoles, etc.).
- 39 catalizadores quirales (ácidos fosfóricos, fosforamidas, etc.).
- 17 disolventes diferentes bajo diferentes condiciones experimentales

El número de posibles combinaciones de sustratos, catalizadores y condiciones de reacción a explorar es potencialmente muy alta para ser cubierta por prueba y error.

Por ejemplo, si las reacciones se realizan de forma independiente cambiando un reactivo a la vez, se deben ejecutar un total de $N_{comb} = N(Subs_{qi})$ x $N(Nuc_{qi})$ x $N(Cat_{qi})$ x $N(Solv_{qi})$=55 x 53 x 39 x 17 = 1.932.645 combinaciones únicas de subtipos moleculares.

Por otro lado, también hay variaciones importantes en las tres principales variables de condiciones experimentales $V_k(c_{qi})[T(^oC), t(h) y L(\%)]$. La Tabla 5.4 muestra diferentes estadísticas de estas variables para las reacciones reportadas.

Se incluyen los valores enteros para el máximo (T_{max}, t_{max} y L_{max}), el mínimo (T_{min}, t_{min} y L_{min}) y step o incremento (T_{step}, t_{step} y L_{step}). Esto es importante porque la expresión $Rango[V_k(c_{qi})] = V_k(c_{qi})_{max} \breve{} V_k(c_{qi})_{min}$ nos da el rango de valores para esta variable que se pueden cubrir en el laboratorio en la práctica real. En consecuencia, cuando este rango se divide por el valor mínimo, se decide cambiar el Paso $[V_k(c_{qi})]$, se puede obtener el número de experimentos $N(c_{qi}) = Rango[V_k(c_{qi}))/Paso(V_k(c_{qi})]$ que se pueden ejecutar para explorar esta variable. Cuando las reacciones se ejecutan de forma independiente cambiando una condición experimental a la vez, se debe ejecutar un total de N_{exp} experimentos.

Esto será igual a $N_{exp} = N(c_1)$ x $N(c_2)$ x $N(c_3) = N(T)xN(t)xN(L)$ = [Rango(T)/Paso(T)] x [Rango(t)/Paso(t)] x [Rango (L)/Paso(L)] = [144/10] x [(239/1] x [(28/1] = 96.365 experimentos de optimización para cada combinación única de subtipos de moléculas dando como resultado un producto específico del reacciones R_{qi} 5.4. La multiplicación de ambas partes de la ecuación da una estimación del gran número de reacciones accesibles en este espacio químico $N(Rqi)_{max} = N_{comb} \cdot N_{exp} \approx$ 1011. Las ecuaciones utilizadas para realizar los cálculos del número de reacciones en este espacio químico se muestran a continuación [225]:

$$N\left(R_{qi}\right)_{\text{máx}} = N\left(\text{Sub}_{qi}\right) \cdot N\left(Nuc_{qi}\right) \cdot N\left(\text{Cat}_{qi}\right) \cdot N\left(\text{Solv}_{qi}\right) \cdot \frac{\text{Range}(T)}{\text{Step}(T)} \cdot \frac{\text{Range}(t)}{\text{Step}(t)} \cdot \frac{\text{Range}(L)}{\text{Step}(L)}$$

$$N\left(R_{qi}\right)_{\text{máx}} = \prod_{s=1}^{s=4} \left[N\left(m_{sqi}\right)\right] \cdot \prod_{k=1}^{k=3} \left[\frac{V_k\left(c_{qi}\right)_{\text{máx}} - V_k\left(c_{qi}\right)_{\text{mín}}}{\text{Step}\left(V_k\left(c_{qi}\right)_{\text{máx}}\right)}\right]$$

$$N\left(R_{qi}\right)_{\text{máx}} = \prod_{s=1}^{s=4} \left[N\left(m_{sqi}\right)\right] \cdot \prod_{k=1}^{k=3} \left[N\left(c_{qi}\right)\right]$$

$$N\left(R_{qi}\right)_{\text{máx}} = N_{\text{comb}} \cdot N_{\text{exp}}$$

El cálculo anterior da una idea de cuán amplio es el espacio de reacción química para las reacciones de α-amidoalquilación intermoleculares enantioselectivas. Es poco práctico verificar en el laboratorio todas estas reacciones por razones de tiempo y coste en recursos materiales y humanos. Es, por lo tanto, un campo abierto que se aborda en la presente tesis doctoral.

Experimentación

Sin embargo, no existen modelos quimioinformáticos para esta reacción utilizando descriptores. Por lo tanto, se intentaron desarrollar métodos computacionales para predecir la enantioselectividad de este tipo de reacciones de α-amidoalquilación intermolecular.

Para ello se utilizaron los índices anteriores calculados con RMarkovTI. El resultado del modelo es el parámetro $ee(\,\%)[R_{cat}]$. Este parámetro es igual al exceso enantiomérico de la reacción utilizando un catalizador de configuración *R*. Por consiguiente, en los casos de reacciones en la literatura con catalizadores *R*, $ee(\,\%)[R_{cat}] = ee(\,\%)$. Por el contrario, en los casos de reacciones con exceso enantiomérico $ee(\,\%)$ comunicado para un catalizador *S* $ee(\,\%)[R_{cat}] = -ee(\,\%)[S_{cat}] = -ee(\,\%)$. Por lo tanto, todos los valores de exceso enantiomérico predicho con este modelo son para reacciones que utilizan un catalizador *R*. Es decir, para el cálculo de los descriptores MCDs a partir de las secuencias, se utiliza la misma aproximación comentada en el capítulo anterior.

Se desarrollará una aproximación experimental incremental generando modelos cada vez más complejos a partir del inicial. Para ello, se tomará un modelo lineal ML como punto de partida y se modificará siguiendo la teoría de la perturbación (PTML). A partir de ese punto, se tratará de mejorar el resultado aplicando diferentes heurísticas al modelo (HPTML). Además, se aplicará simulación de Monte Carlo el muestreo aleatorio para simular y estimar propiedades químicas de sistemas complejas, generando gran cantidad de muestras de manera aleatoria. A continuación, se comentará la experimentación de cada uno de los incrementos.

Modelo Lineal ML. A partir de los valores de $D_k(m_{sqi})_g$ se busca un modelo de aprendizaje automático lineal, en el que cada línea de entrada del conjunto de

datos hace referencia a una sola reacción de consulta (R_{qi}). El exceso enantiomérico $ee_R(\%)_{qicalc}$ de la reacción de consulta (R_{qi}) se predice utilizando como entrada tanto las variables $V_k(c_{qi})$ relacionadas con las condiciones experimentales como los descriptores moleculares $D_k(m_{sqi})_g$ de las moléculas involucradas en la reacción. Usando ambos conjuntos de variables como entradas, se puede buscar un modelo aditivo de ML lineal. Idealmente, $eeR(\%)calcqi \approx eeR(\%)qiobs$ si la hipótesis lineal aditiva es correcta. La fórmula general de este tipo de modelo se muestra en la siguiente ecuación:

$$ee_R(\%)_{\text{cald}}\,\hat{k}_j = \sum_{k=1}^{k_{\text{máx}}}\sum_{s=1}^{s_{\text{máx}}} a_{k,s}\cdot V_k\left(c_{qi}\right) + \sum_{k=1}^{k_{\text{máx}}}\sum_{s=1}^{s_{\text{máx}}}\sum_{g=1}^{g_{\text{máx}}} b_{k,s,g}\cdot D_k\left(m_{sqi}\right)_g + e_0$$

Modelo lineal PTML. El modelo PTML sirve para predecir una propiedad de un nuevo caso (reacción) haciendo una comparación con otras reacciones conocidas. La salida de nuestro modelo también es $ee_R(\%)_{calcqi}$. Sin embargo, en este caso el $ee_R(\%)qicalc$ se calcula para una reacción de consulta (R_{qi}) ya que ya se conoce el exceso enantiomérico observado $ee_R(\%)_{rjobs} = ee_R(\%)_{refj}$ de una reacción (R_{rj}) utilizada como reacción de referencia.

En consecuencia, en el conjunto de datos utilizado para entrenar el modelo PTML, cada línea de entrada hace referencia a un par de reacciones: una reacción de consulta comparada con una reacción de referencia (R_{qi} frente a R_{rj}). El modelo aditivo lineal PTML predice $ee_R(\%)_{calci}$ comenzando con el valor experimental de $ee_R(\%)_{refj}$ de una reacción de referencia. A continuación, el modelo agrega los efectos de cambios en las condiciones estructurales, operativas o experimentales (perturbaciones) en la consulta con respecto a la reacción de referencia. Los parámetros utilizados para cuantificar estos cambios o perturbaciones se denominan Operadores PT (PTOs). Se utilizan PTOs con la forma $\Delta V_k(c_{qi}, c_{rj})$ para cambios estructurales y $\delta D_k(m_{sqi}, m_{srj})_g$ para cambios en las condiciones experimentales. La fórmula del modelo PTML utilizado aquí se muestra en las ecuaciones siguientes:

$$\begin{aligned} ee_R(\%)_{\text{calcqi}} = & ee_R(\%)_{refj} + \sum_{k=1}^{k_{\text{máx}}}\sum_{i=1}^{i_{\text{máx}}} a_{k,i}\cdot\Delta V_k\left(c_{qi}, c_{rj}\right) \\ & + \sum_{k=1}^{k_{\text{máx}}}\sum_{s=1}^{s_{\text{máx}}}\sum_{g=1}^{g_{\text{máx}}} b_k\cdot\Delta D_k\left(m_{sqi}, m_{srj}\right)_g + e_0 \end{aligned}$$

$$ee_R(\,\%)_{\text{calcqi}} = ee_R(\,\%)_{refj} + \sum_{k=1}^{k_{\text{máx}}} \sum_{i=1}^{i_{\text{máx}}} a_{k,i} \cdot [V_k\,(c_{qi}) - V_k\,(c_{rj})] \quad (4)$$

$$+ \sum_{k=1}^{k_{\text{máx}}} \sum_{i=1}^{i_{\text{máx}}} \sum_{g=1}^{g_{\text{máx}}} b_k \cdot \left[D_k\,(m_{sqi})_g - D_k\,(m_{srj})_g\right] + e_0$$

En el caso de este modelo aditivo lineal, se utiliza como entrada una función de referencia $ee_R(\,\%)_{robs}$ y dos conjuntos de PTO indicados por $\delta V(c_{qi}, c_{rj})$ y $\delta V(m_{sqi}, m_{srj})_g$. La función de referencia $eeR(\,\%)_{robs}$ es igual a los valores observados de exceso enantiomérico $ee(\,\%)$ para la reacción de referencia usando un catalizador con configuración *R*. El primer tipo de PTO tiene la fórmula $\delta V_k(c_{qi}, c_{rj}) = [V_k(c_{qi})\breve{}V_k(c_{rj})]$. Da cuenta de las perturbaciones/desviaciones en los valores de las k^{th} variables/condiciones de las reacciones $V(c_{qi})$ de la q^{th} reacción de consulta frente a los valores originales de las mismas variables $V_k(c_r)$ para la r^{th} reacción de referencia. Por analogía, el segundo tipo de PTO tiene la fórmula: $\Delta D_k(m_{sqi}, m_{srj}) = [D_k(m_{sqi})\breve{}D_k(m_{srj})]_g$. Da cuenta de las perturbaciones/desviaciones en los valores de los descriptores moleculares de la consulta frente a las moléculas de referencia. En consecuencia, las variables de entrada para la reacción de la referencia $V_k(c_{rj})$ están conectadas a una propiedad k^{th} (k = 1, 2, 3) relacionada con las condiciones generales de reacción (c_{rj}) y/o reactivos específicos:

- $V_1(c_{qi})$ = T(^{o}C) = temperatura.
- $V_2(c_{qi})$ = t(h) = tiempo de reacción.
- $V_3(c_{qi})$ = L(%) = carga de catalizador.

para la reacción de referencia (R_{rj}).

Las variables de entrada $D_k(m_{ri})_g$ son descriptores moleculares de tipo k^{th} para las i^{th} moléculas (m_{sri}) de tipo qth implicadas en la reacción de referencia (R_{rj}). Por analogía, los tipos de moléculas m_{ri} involucradas en la reacción de referencia son $m_{r1j} = Substrate_{rj}$, $m_{r2j} = Nucleofilo_j$, $m_{r3j} = Catalizador_{rj}$ y $m_{4rj} = disolvente_{rj}$. Los k^{th} tipos de descriptores moleculares considerados son los mismos que para la reacción de consulta:

- D1 = Número de electrones de valencia (Zv).

- D2 = Volumen de Van der Waals (Vvdw).
- D3 = Electronegatividad de Sanderson (χ).
- D4 = Polarizabilidad (α).
- D5 = Afinidad Electrónica (EA).

La Tabla 5.3 muestra las definiciones de todos los PTO utilizados como variables de entrada en los modelos PTML.

Tab. 5.3.: Definición de variables usadas como entrada del modelo lineal PTML.

Condiciones experimentales (c_ε)	Operadores de perturbación [b]	Tipo de operador
Temperatura de la reacción (T)	$\Delta V(T) = \Delta T = T_\Omega - T_{lr}$	Desviación de temperatura
Tiempo de reacción (t)	$\Delta V(t) = \Delta t = t_a - t_r$	Desviación de tiempo
Catalizador Loading [Carga(%)]	$\Delta V(\underline{Load}(\,\%)) = Load(\,\%)_q - Load(\,\%)_r$	Conc. difference
Moléculas (m_Θ)[a]	Perturbation terms	Tipo de operador [a]
Sustrato (Sub)	$\Delta D_k\left(Sub_{\Omega i}, Sub_i\right)_g = \left[D_k\left(Sub_{gi}\right)_g - D_k\left(Sub_{rj}\right)\right]_g$	
Producto(Prod)		
Nucleofilo (Nuc)	$\Delta D_b\left(Nuc_{si}, Nuc\right)_g = \left[D_k(Nuc)_g - D_k\left(NucNri^2\right)\right]_g$	Variación estructural
Catalizador (Cat)	$\Delta D_k\left(Cat_{\circ i,}, Catc_j\right)_g = \left[D_k\left(Cat_a\right)_g - D_k\left(Cat_s\right)\right]_g$	
disolvente (Solv)		

[a] $Moléculas$ m implicadas en la reacción con roles distinguibles: mosi = Sustrato (Sub&), Producto (Brodg), Estos PTOs miden la variación del valor de la propiedad molecular/variable estructural (V) en las moléculas de consulta maCon respecto al valor para la molécula ms con el mismo rol en la reacción de referencia. Los valores de $Y_k(m_q)_g$ son valores medios de las propiedades V_k Electronegatividades de Sanderson χ, Polarizabilidades, etc., para todos los átomos del grupo g y todos sus átomos vecinos situados a una distancia topológica $k \leq 5$. En consecuencia, estas propiedades se han calculado para todos los átomos de la molécula (Tot) o para subconjuntos de átomos (grupo g). Los grupos de átomos estudiados son g = carbonos insaturados (C_{cuns}), carbonos saturados (C_{sat}) , heteroátomos (Het), heteroátomos no halógenos (HetNoX).

Para buscar los modelos lineales ML y PTML, se utilizaron métodos LM y RNA lineal. En los modelos de regresión PTML, los valores de exceso enantiomérico observado (experimental) $ee_R(\%)_{obsqi}$ *vs.* diferentes valores de referencia $eeR(\%)_{refj}$ tienen que ser ajustados, generando artefactos en la distribución normal de los datos [39]. La ecuación para los modelos de regresión lineal PTML se ajustan de acuerdo a lo mostrado en:

$$\Delta ee_R(\%)_{qi} = \sum_{k=1}^{k_{\text{máx}}} \sum_{s=1}^{s_{\text{máx}}} a_{k,s} \cdot \Delta V_k \left(c_{qi}, c_{rj}\right) + \sum_{k=1}^{k_{\text{máx}}} \sum_{s=1}^{s_{\text{máx}}} \sum_{g=1}^{g_{\text{máx}}} b_{k,s,g} \cdot \Delta D_k \left(m_{sqi}, m_{srj}\right)_g + e_0$$

Modelo lineal HPTML. El modelo lineal PTML predice diferentes salidas para la misma reacción dependiendo de las reacciones de referencia seleccionadas. Así, se pueden utilizar diferentes Heurísticas (H) para determinar la mejor reacción o conjunto de reacciones a utilizar como referencia. Para el modelo propuesto en la tesis doctoral, se seleccionaron dos heurísticas.

La primera heurística (H_1) calcula el valor previsto final $ee_R(\%)_{qrpred} = eeR(\%)_{qrmin}$. Este valor se obtiene tomando como referencia la reacción con valor mínimo (Min) de los PTOs (desviación mínima), es decir, la reacción con diferencia mínima (Δ) sobre las variables de entrada $\Delta V(m_{qsi}, m_{rsj})$ y $\Delta V(c_{qi}, c_{rj})$ para todos los ($\forall$) pares de reacciones se utilizó como referencia.

La heurística (H_2) calcula el valor $eeR(\%)_{qrpred} = eeR(\%)_{qravg} = Avg(eeR(\%)_{qrcalc})$, es decir, los valores de las variables $\Delta V(m_{qi}, m_{rj})$ y $\Delta V(c_{qi}, c_{rj})$ para todos los pares de reacciones se utilizan como entrada. Primero, se calcularon los 331 valores diferentes de $eeR(\%)qrcal$ (excluyendo la consulta). A continuación, se calculan los valores finales como la media de todas las referencias.

Estas dos heurísticas se pueden escribir como se muestra en las siguientes ecuaciones.

$$H_1: \quad ee_R(\%)_{\text{qrpred}} = ee_R(\%)_{\text{qrmin}} \overset{\forall q,r}{\Longrightarrow} \text{Min}\left\{\text{PTOs}\left[\Delta\text{V}\left(\text{m}_{qi},\ \text{m}_{rj}\right), \Delta\text{V}\left(\text{c}_{qi}, \text{c}_{rj}\right)\right]\right\}$$

$$H_2: \quad ee_R(\%)_{\text{qrpred}} = ee_R(\%)_{\text{qravg}} \overset{\forall q,r}{\Longrightarrow} \text{Avg}\left\{\text{PTOs}\left[\left(\Delta\text{V}\left(\text{m}_{qi},\ \text{m}_{rj}\right), \Delta\text{V}\left(\text{c}_{qi}, \text{c}_{rj}\right)\right]\right\}$$

Simulación Monte Carlo. Los métodos similares a Monte Carlo (MC) se encuentran entre los muestreos de datos más útiles en quimioinformática [226, 227, 228]. Dado que el conjunto de datos de 332 reacciones iniciales no era totalmente representativo

de algunos subtipos de reacciones, se propone el uso de simulación MC tanto para el enriquecimiento de datos sintéticos como para el modelo.

A partir del cálculo de los valores mínimo $V_k(c_{qi})_{min}$, máximo $V_k(c_{qi})_{max}$ y paso $V_k(c_{qj})$ para todas las condiciones operativas, ver Tabla 5.4, se propone utilizar un modelo MC basado en el siguiente sistema de ecuaciones para generar los nuevos datos sintéticos.

Tab. 5.4.: Resumen de las estadísticas básicas de las reacciones del conjunto de datos.

Stat.[a]	Dataset reaction conditions $(c_{qi})^b$		
	T(^oC)	t(h)	Load(%)
Nreacc	12	53	7
Avg.	11.59	35.87	9.10
S.D.	26.80	33.94	5.72
Min.	-78.00	1.00	2.00
Max.	66.00	240.00	30.00
Range	144	239	28
Step	10	1	1
Nexpr	14	239	28

[a] $Stat.$ = Parámetros estadísticos para los parámetros de entrada (condiciones operativas) de todas las reacciones presentes en el conjunto de datos: N_{reacc} = Número de reacciones presentes en nuestro conjunto de datos, Avg. = valor promedio, S.D. = desviación estándar, Max. = valor máximo, Min. = valor mínimo, Range = Max. - Mín., Paso = cambio mínimo permitido en una condición experimental, N_{expr}. = Número de experimentos (reacciones) cambiando una condición y manteniendo las demás constantes. [b] $Condicionesoperativas$: T(oC) = temperatura, t(h) = tiempo de reacción, Load(%) = carga de catalizador.

$$V_k\left(c_{qi}\right)_{new} = \left(V_k\left(c_{qi}\right)_{\text{mín}} + \text{Rnd}\left(0, n_{\text{máx}}\right) \cdot V_k\left(c_{qi}\right)_{\text{step}}\right)$$
$$V_k\left(c_{qi}\right)_{\text{synth}} = if\left[V_k\left(c_{qi}\right)_{new} > V_k\left(c_{qi}\right)_{\text{máx}}; V_k\left(c_{qi}\right)_{\text{máx}}; V_k\left(c_{qi}\right)_{new}\right]$$

A partir de dichas ecuaciones, se obtienen nuevos valores de datos sintéticos $V_k(c_{qi})_{synth}$ después de aplicar una condición de contorno. Esta condición de contorno mantiene los valores sintéticos $V_k(c_{qi})_{synth}$ dentro del rango $[V_k(c_{cqi})_{min}, V_k(c_{cqi})_{max}]$. Es usado en la generación de valores sintéticos para las variables de condición experimental $V_1(c_{qi}) = T(^oC), V_2(c_{qi}) = t(h), V_3(c_{qi}) = L(\%)$.

Significa que los nuevos valores de datos sintéticos son iguales a $V_(c_k)_{synth} = V(c_k)_{min} + rnd(0, Nmax)$ůů$V(c_k)_{step}$ si y solo si son menores que $V_k(c_{qi})_{max}$; de lo contrario, son iguales a $V_k(c_{qi})_{max}$. La función $Rnd(0, n_max)$ es un generador de

números naturales pseudoaleatorios ($n = 0, 1, 2, \ldots N_{max}$) basado en el algoritmo Mersenne-Twister MC (MT19937).

Resultados

A partir de la experimentación presentada en la sección anterior, se mostrarán a continuación los resultados obtenidos por el modelo y sus sucesivos refinamientos.

Modelo lineal ML. En las reacciones de α-amidoalquilación, no existe una relación obvia entre la quiralidad de los catalizadores y la notación Cahn, Ingold, Prelog (CIP) del producto. De hecho, en el conjunto de datos de la literatura consultada se puede observar el ratio de quiralidad Catalizador/Producto, (*R*)/(*R*) 140 reacciones (43,2 %), (*S*)/(*R*) 102 reacciones (31,5 %), (*R*)/(*S*) 72 reacciones (22,2 %) y (*S*)/(*S*) 9 reacciones (2,8 %) de un total de 324 reacciones. Solo hay una reacción en todo el conjunto de datos con un catalizador de configuración (*S*) y un exceso enantiomérico igual a cero. Por lo tanto, es muy importante contar con un modelo computacional para predecir la estereoquímica absoluta y el exceso enantiomérico del producto de la reacción.

Este tipo de modelos podrían ser muy útiles para diseñar nuevos catalizadores y/o seleccionar a priori las condiciones óptimas de reacción. El modelo fue entrenado con las 324 reacciones originales. La ecuación de este modelo resultó ser:

$$
\begin{aligned}
ee_R(\,\%)_{qrpred} = 912{,}48 \cdot \mathrm{Loq}_q\, d(\,\%) + 21{,}90 \cdot T\,(^{\circ}\mathrm{C}) - 194{,}76 \cdot t(\,\mathrm{h}) \\
-13{,}21 \cdot \alpha\,(\,\mathrm{Cat}_{\,qi})_{Cuns} - 45{,}02 \cdot \propto (\mathrm{Prod}_q)_{\mathrm{HetNoX}} \\
+830{,}06 \cdot \Delta EA\,(\mathrm{Prod}_q)_{\mathrm{Csat}} - 0{,}34 \cdot EA\,(\,\mathrm{Cat}_{\,qi})_{\mathrm{HetNoX}} \\
+0{,}22 \cdot \chi\,(\,\mathrm{Nuc}_{\,qi})_{Het} - 2024{,}05 \cdot \chi\,(\,\mathrm{Cat}_{\,q})_{HetNoX} \\
-178{,}69 \cdot V\,(\,\mathrm{Sub}_{\,qi})_{Tot} - 1678{,}05 \cdot \mathrm{Zv}\,(\,\mathrm{Cat}_{\,qi})_{\mathrm{Cuns}} \\
-34{,}41 \cdot Zv\,(\,\mathrm{Solv}_{\,q})_{\mathrm{Cuns}} - 0{,}70
\end{aligned}
$$

$$
n = 332(reaciones) R^2 = 0{,}74 F = 59{,}2 pv 0{,}5
$$

El modelo lineal ML no utiliza reacciones de referencia para la comparación. Los parámetros estadísticos del modelo son $n = 332$, coeficiente de regresión $R2 = 0{,}74$, relación de Fisher $F = 59{,}2$, error estándar de estimación $SEE = 37{,}1$ y $p - valor < 0{,}05$. Los detalles sobre los coeficientes y las variables del modelo, incluidos los

símbolos y los nombres de las variables, el error estándar, los valores de la T de Student y los p-valores, se muestran en la Tabla 5.5.

Tab. 5.5.: Resultados del modelo de regresión PTML.

Model	Input Varsa	Symbol	Coeff	eer(%))^{b}	S.E. c	t^{d}	p-valor ^{e}
[12*ML]	Load(%)	$V_3\,(c_{Gi})$	a_1	912.48	258.3259	3.53227	$< 0{,}05$
	$T\,(^{\circ}C)_{ex}$	$V_1\,(c_{cil})$	a_2	21.90	18.7647	1.16732	0.24
	$t(h)_{ar}$	$V_2\,(c_{aij})$	a3	-194.76	55.2274	-3.52654	$< 0{,}05$
	$\alpha\,(Cat_{ci})_{Cups}$	$D_4\,(\,m_{3q})_2$	b_1	-13.21	2.7499	-4.80465	$< 0{,}05$
	$\alpha\,(\,$ Prod $_{ai})$ HetNox	$D_4\,(\,m_{5q})_4$	b_2	-45.02	21.8624	-2.05905	$< 0{,}05$
		$D_5\,(\,m_{5q})_1$	b_3	830.06	163.3526	5.08139	$< 0{,}05$
	EA(Catsii)Hettoox	$D_5\,(\,m_{3q})_4$	b_4	-0.34	0.0949	-3.62574	$< 0{,}05$
	$\chi\,(NucNai^{Het}$	$D_3\,(\,m_{2qi})\,3$	b_5	0.22	0.0742	3.01949	$< 0{,}05$
	$\chi\,(\,$ Catais $^{Hetad\ 8}$	$D_3\,(\,m_{3qi})_4$	b_6	-2024.05	484.4448	-4.17809	$< 0{,}05$
	$V\,(Sub_{sij})_{Tot}$	$D_2\,(\,m_{1q})\,5$	b_7	-178.69	43.4355	-4.11390	$< 0{,}05$
	Zv(Catai) Cugns	$D_1\,(\,m_{3q})_2$	b_8	-1678.05	468.7747	-3.57965	$< 0{,}05$
	Zv $(Sely_{si})$ Cumser	$D_1\,(\,m_{4q})_2$	b9	-34.41	11.6896	-2.94399	$< 0{,}05$
	Intercept	-	eo	-0.70	0.5150	-1.35948	0.18

propiedades: $\alpha =$ polarizabilidad atómica media, EA = Electroafinidad atómica media, $\chi =$ Electronegatividad Sanderson atómica media, Zx = número atómico medio. ^{b} Coeficientes de las variables del modelo, la variable de salida es el Δ en exceso enantiomérico $ge(\,\%)^{*}$ de la reacción q con respecto a la reacción r cuando ambas reacciones se han llevado a cabo con *R*-catalizador. ' Error estándar de los coeficientes. ' Valor t de Student. $^{mathrmp} - valordelerror$.

En cualquier caso, el $SEE = 37{,}1$ podría considerarse relativamente alto [225]. Otro inconveniente importante de este modelo lineal clásico de ML es que no proporciona pistas sobre las reacciones más similares reportadas en la literatura. Esto puede limitar la capacidad para deducir posibles mecanismos y/o comparar los resultados obtenidos con otros ya conocidos.

Por lo tanto, este modelo se utilizará junto con otra estrategia de búsqueda de moléculas similares para obtener pistas de reacciones similares para un caso de estudio determinado. Una opción es acoplar este modelo con estrategias de búsqueda de similitud basadas en índices de similitud de Tanimoto [229]. De hecho, hay interesantes trabajos que reportan el acoplamiento de modelos quimioinformáticos con estrategias de búsqueda basadas en similitud [230, 231, 232]. Un ejemplo bien conocido de una herramienta de búsqueda en línea es la plataforma Scifinder [233, 234].

Modelo PTML. Los modelos de reactividad PTML pueden estudiar reacciones por pares [225]. El modelo infiere la reactividad de una reacción de consulta (q) comparándola con una reacción de referencia previamente conocida (r). La ecuación del modelo PTML lineal resultante fue:

$$\begin{aligned}\Delta e_R(\,\%)_{qr} = &-0{,}82 \cdot \Delta\,\mathrm{Load}(\,\%) - 0{,}34 \cdot \Delta T\,(^{\circ}\mathrm{C}) + 0{,}21 \cdot \Delta t(h) \\ &- 174{,}37 \cdot \Delta \propto (\ \mathrm{Cat}_{\,q},\ \mathrm{Cat}_{\,r})_{\mathrm{Cuns}} - 1534{,}17 \cdot \Delta \propto (\ \mathrm{Prod}_{\,q},\ \mathrm{Prod}_{\,r})_{\mathrm{HetNoX}} \\ &- 215{,}98 \cdot \Delta EA\,(\mathrm{Prod}_q, \mathrm{Prod}_r)_{\mathrm{Csat}} - 1747{,}12 \cdot \Delta EA\,(\ \mathrm{Cat}_{\,q},\ \mathrm{Cat}_{\,r})_{\mathrm{HetNoX}} \\ &- 42{,}49 \cdot \Delta\chi\,(\ \mathrm{Nuc}_{\,q},\ \mathrm{Nuc}_{\,r})_{\mathrm{Het}} + 750{,}76 \cdot \Delta\chi\,(\ \mathrm{Cat}_{\,q},\ \mathrm{Cat}_{\,r})_{\mathrm{HetNoX}} \\ &- 34{,}19 \cdot \Delta V\,(\ \mathrm{Sub}_{\,q},\ \mathrm{Sub}_{\,r})_{\mathrm{Tot}} + 22{,}04 \cdot \Delta Zv\,(\ \mathrm{Cat}_{\,q},\ \mathrm{Cat}_{\,r})_{\mathrm{Cuns}} \\ &- 12{,}46 \cdot \Delta Zv\,(\ \mathrm{Solv}_{\,q},\ \mathrm{Solv}_{\,r})_{\mathrm{Cuns}} - 0{,}91\end{aligned}$$

rtain =0.84 F=15238.7 p<0.5

El modelo se entrenó utilizando un total de $n_{train} = 78.732$ pares de reacciones seleccionados al azar. El modelo presentó un valor de coeficiente de regresión de $R_{train} = 0{,}84$, un $SEE = 51,67$ y una razón de Fisher de $F = 15{,}238,7$ con $p - valor < 0{,}05$ en entrenamiento. Esto indica una relación significativa entre los valores relativos observados de $\Delta eeR(\,\%)_{qrobs}$ *vs.* los valores predichos $\Delta eeR(\,\%)_{qrobs}$.

Además, se utilizó otro subconjunto de $n_{val} = 28{,}836$ pares de reacciones para validar el modelo. Para esta serie de validación se encontró un coeficiente de regresión $R_{val} = 0,77$ y $SEE = 60{,}225$. La salida del modelo es $ee_R(\%)_{qrcalc}$, que representa el valor del exceso enantiomérico calculado utilizando una única reacción de referencia. El valor $eeR(\%)_{calc}$ cuantifica el exceso enantiomérico obtenido utilizando un catalizador (*R*), de manera que:

- Si $eeR(\%)_{calc}$ >0, se predice que el producto tendrá notación (*R*).
- Si $eeR(\%)_{calc}$ <0, se predice que el producto tendrá notación (*S*);
- Si $eeR(\%)_{calc}$ = 0 se predice mezcla racémica.

El p-valor del modelo es $< 0{,}05$ y todas las variables introducidas en el modelo son estadísticamente significativas, resultados en la Tabla 5.6. Las tres primeras variables de entrada cuantifican el efecto de factores no estructurales sobre el parámetro de enantioselectividad, $eeR(\%)_{calc}$. Las variables de entrada restantes cuantifican la contribución de las variaciones estructurales en el Catalizador (Cat), Producto (Prod), Nucleófilo (Nuc) y Disolvente (Solv).

Tab. 5.6.: Resultados del modelo de regresión PTML.

Model	Input Varsa	Symbol	Coeff [b]	$\mathrm{Aeger}(\,\%)^{\mathrm{b}}_{\mathrm{own}}$	SEE	t^{d}	p-levele
[12* PTML]	$\Delta\mathrm{Load}(\,\%)_{\propto}\quad I$	$\Delta V_3\,(C_{ij;}, C_{cj})$	a_1	-0.82	0.03243	-25.3154	$< 0{,}005$
	$\Delta\mathrm{T}\,(\mathrm{QC})_{\mathrm{s}}$	$\Delta V_l\,(C_{\mathrm{ligi}}\,, \mathrm{C}_{rj})$	a_2	-0.34	0.00594	-57.8960	$< 0{,}005$
	$\Delta t(h)_{\propto+\mathrm{c}}$	$\Delta\mathrm{V}_2\,(C_{ij;}, C_{cj})$	a_3	0.21	0.00476	44.4627	$< 0{,}005$
	$\underline{\alpha}\,(\mathrm{Cat}_8\mathrm{Cat})_{\mathrm{cuns}}$	$\Delta\mathrm{D}_4\,(\,\mathrm{m}_{3\mathrm{qi}}, \mathrm{m}_{3\mathrm{rj}})_2$	b_1	-174.37	2.67954	-65.0741	$< 0{,}005$
	$\underline{\alpha}\,(\,\mathrm{Prod}\,_a,\ \mathrm{Prod}\,)_{\mathrm{Heetax}}$	$\Delta\mathrm{D}_4\,(\,\mathrm{m}_{5\mathrm{qi}}, \mathrm{m}_{5\mathrm{rj}})_4$	b_2	-1534.17	26.38185	-58.1525	$< 0{,}005$
	$\mathrm{EA(Prod\ Prod}\,_2)_{\mathrm{cast}}$	$\Delta\mathrm{D}_5\,(\,\mathrm{m}_{5\mathrm{q}i},\ \mathrm{m}_{5\mathrm{rr}})_1$	b_3	-215.98	3.38484	-63.8086	$< 0{,}005$
	$\mathrm{EA}\,(\mathrm{Cat}_0\,,\mathrm{Catt}\,)\,\mathrm{Hetax}$	$\Delta\mathrm{D}_5\,(\,\mathrm{m}_{3\mathrm{q}i},\ \mathrm{m}_{3\mathrm{r}})_4$	b_4	-1747.12	26.48292	-65.9715	$< 0{,}005$
	$y(\mathrm{Nuc},\ \mathrm{Nuc}\,)_{\mathrm{Het}}$	$\Delta\mathrm{D}_3\,(\,\mathrm{m}_{2\mathrm{q}i},\ \mathrm{m}_{2\mathrm{rj}})_3$	b_5	-42.49	1.33694	-31.7788	$< 0{,}005$
		$\Delta\mathrm{D}_3\,(\,\mathrm{m}_{3\mathrm{qi}}, \mathrm{m}_{3\mathrm{rj}})_4$	b_6	750.76	8.98832	83.5259	$< 0{,}005$
		$\Delta\mathrm{D}_2\,(\,\mathrm{m}_{1\mathrm{qi}}, \mathrm{m}_{1\mathrm{r}})_5$	b_7	-34.19	0.70023	-48.8225	$< 0{,}005$
	$\mathrm{Zx}\,(\mathrm{Cat}_0, \mathrm{Cat})_{\mathrm{Cuna}}$	$\Delta\mathrm{D}_1\,(\,\mathrm{m}_{3\mathrm{q}}, \mathrm{m}_{3\mathrm{r}})_2$	b_8	22.04	1.16398	18.9356	$< 0{,}005$
	$\mathrm{Zx}\,(\ \mathrm{Solva,\ Solva}\ _{\mathrm{cuna}}$	$\Delta\mathrm{D}_1\,(\,\mathrm{m}_{4\mathrm{q}i},\ \mathrm{m}_{4\mathrm{r}})_2$	b_9	-12.46	0.16653	-74.8101	$< 0{,}005$
	Intercept	-	e_0	-0.91	0.18257	-5.0065	$< 0{,}005$

[a] variables de entrada con coeficiente $\mathrm{b_k}$ son los valores del desplazamiento (Δ) en q-reacys. r-reac para diferentes propiedades: $\alpha =$ polarizabilidad atómica media, EA = Electroafinidad atómica media, $\chi =$ Electronegatividad Sanderson atómica media, Zx = número atómico medio. [b] Coeficientes de las variables en el modelo, la variable de salida es el Δ en exceso enantiomérico $ge(\,\%)^*$ de la q-reacción con respecto a la r-reacción cuando ambas reacciones se han llevado a cabo con *R*-catalizador. [c] Error estándar de los coeficientes. [d] Valor T del test de Student. [e] p-valor del error.

Modelo PTML con una única reacción de referencia. Como se comentó anteriormente, este modelo de reactividad PTML estudia las reacciones por pares. Para evitar distorsiones en la distribución de las variables, el modelo PTML utiliza la variable $\Delta ee_R(\%)_{qrobs}$ como función objetivo [225]. Esta función objetivo es la función a ajustar y es igual a $\Delta ee_R(\%)_{qrobs} = ee_R(\%)_{qobs}$ - $ee_R(\%)_{robs}$. Como resultado, la salida del nuevo modelo es $\Delta ee_R(\%)_{qrcalc} = ee_R(\%)_{qcalc}$ - $eeR(\%)_{rcalc}$. Para modelos no precisos $\Delta ee_R(\%)_{qrcalc} \neq \Delta eeR(\%)_{qrobs}$ (donde $\neq$ indica no $\approx$). Por el contrario, para un predictor preciso no aleatorio, como éste, se puede aproximar $\Delta ee_R(\%)_{qrcalc} \approx \Delta eeR(\%)_{qrobs}$, lo que presupone que $eeR(\%)_{qcalc} \approx eeR(\%)_{qobs}$ y $eeR(\%)_{rcalc} \approx eeR(\%)_{robs}$.

Por lo tanto, con fines prácticos, se usa el modelo para predecir el exceso enantiomérico de las nuevas reacciones de consulta $eeR(\%)_{qcalc}$, en función del exceso enantiomérico observado de una reacción de referencia $eeR(\%)_{qrobs}$. La aproximación solo es válida para predictores precisos no aleatorios y tiene en cuenta que $eeR(\%)_{rcalc} \approx eeR(\%)_{robs}$ es siempre una reacción de referencia conocida, por lo que es necesario reordenar las variables en la ecuación, como se muestra a continuación:

$$
\begin{aligned}
\mathrm{Iee}_R(\%)_{\mathrm{qrcalc}} = & ee_R(\%)_{\mathrm{rjobs}} - 0{,}82 \cdot \Delta\,\mathrm{Load}(\%) - 0{,}34 \cdot \Delta T\,(^{\circ}\mathrm{C}) + 0{,}21 \cdot \Delta t(h) \\
& - 174{,}37 \cdot \Delta \propto (\ \mathrm{Cat}_{\ q},\ \mathrm{Cat}_{\ r})_{\mathrm{Cuns}} - 1534{,}17 \cdot \Delta \propto (\ \mathrm{Prod}_{\ q'}\ \mathrm{Prod}_{\ r})_{\mathrm{HetNoX}} \\
& - 215{,}98 \cdot \Delta EA\,(\mathrm{Prod}_q,\ \mathrm{Prod}_{\ r})_{\mathrm{Csat}} - 1747{,}12 \cdot \Delta EA\,(\ \mathrm{Cat}_{\ q'}\ \mathrm{Cat}_{\ r})_{\mathrm{HetNox}} \\
& - 42{,}49 \cdot \Delta\chi\,(\ \mathrm{Nuc}_{\ q'},\ \mathrm{Nuc}_{\ r})_{\mathrm{Het}} + 750{,}76 \cdot \Delta\chi\,(\ \mathrm{Cat}_{\ q'},\ \mathrm{Cat}_{\ r})_{\mathrm{HetNoX}} \\
& - 34{,}19 \cdot \Delta V\,(\ \mathrm{Sub}_{\ q'}\ \mathrm{Sub}_{\ r})_{\mathrm{Tot}} + 22{,}04 \cdot \Delta Zv\,(\ \mathrm{Cat}_{\ q'}\ \mathrm{Cat}_{\ r})_{\mathrm{Cuns}} \\
& - 12{,}46 \cdot \Delta Zv\,(\ \mathrm{Solv}_{\ q'}\ \mathrm{Solv}_{\ r})_{\mathrm{Cuns}} - 0{,}91
\end{aligned}
$$

El modelo calcula diferentes valores de $eeR(\%)_{qcalc}$ para la misma reacción dependiendo del valor experimental $eeR(\%)_{rjobs}$ de la reacción utilizada como referencia en el par [225].

La Figura 5.5 muestra los valores observados de $\Delta eeR(\%)qrobs$ frente a los valores previstos (calculados) de $\Delta ee_R(\%)_{qrcalc}$ para 10.000 pares de reacciones. Se observa una clara tendencia lineal (puntos con $\Delta ee_R(\%)_{qrcalc} \approx \Delta eeR(\%)qrobs$). Sin embargo, a pesar de ser un predictor con buen valor de ajuste, existen muchos puntos con mayor dispersión (puntos con $\Delta ee_R(\%)_{qrcalc} \neq \Delta eeR(\%)qrobs$).

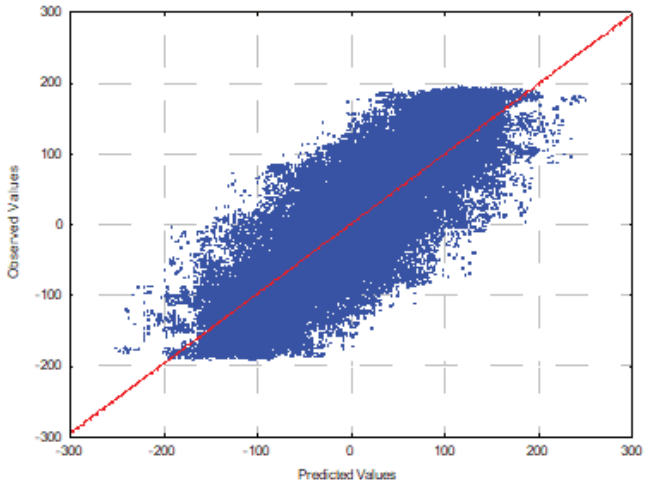

Fig. 5.5.: Observado *vs.* previsto ($\Delta ee_R(\%)_{qrobs}$*vs* .$\Delta ee_R(\%)_{qrcalc}$) para 10.000 pares de reacciones.

Modelo HPTML

Como se mencionó anteriormente, es necesario definir la mejor reacción o conjunto de reacciones a utilizar. Definir una reacción de referencia adecuada también puede ayudar a reducir la dispersión y a aumentar el valor del coeficiente de regresión, ya que cada reacción de consulta tendrá un único valor predicho. Para ello, se puede utilizar una regla heurística acoplada al modelo PTML para seleccionar la mejor referencia. Los métodos basados en heurística han sido ampliamente utilizados en quimioinformática para resolver problemas prácticos [235, 236, 237].

Como se indicó previamente, se probaron dos heurísticas (H_1 y H_2) calculando los valores $eeR(\%)_{qrpred}$ para las 332 reacciones del conjunto de datos, utilizando el PTML entrenado en la fase previa. La Figura 5.6 muestra una ilustración esquemática de las diferencias entre los modelos ML, PTML y HPTML, así como los procedimientos de enriquecimiento de datos de MC utilizados.

La Tabla 5.7 muestra los resultados obtenidos. Cabe señalar que ambos modelos HPTML que usan Heurística dan buenos resultados con un coeficiente de regresión en el rango 0.64 – 0.81 y $p - valor < 0{,}05$. En concreto, el modelo HPTML H_1 tiene el coeficiente de regresión más alto ($R2 = 0{,}81$ frente a 0,55) y un SEE más bajo ($SEE = 29{,}5$ frente a 37,1) que el modelo ML clásico. Sin embargo, este valor SEE sigue siendo relativamente alto.

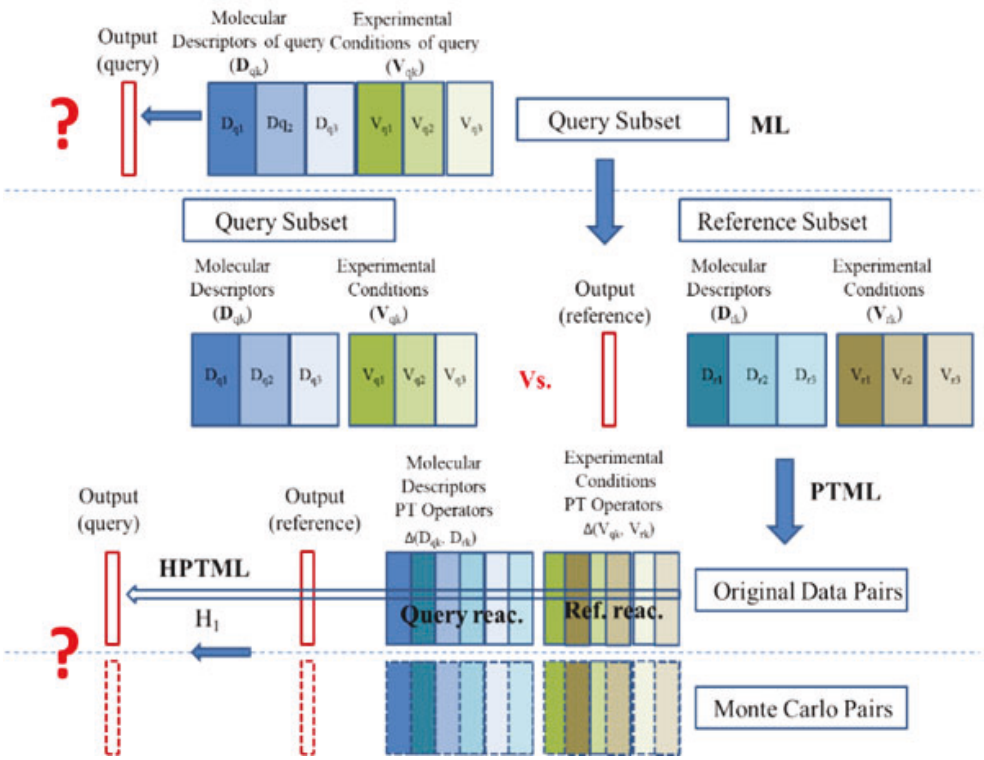

Fig. 5.6.: Esquema de reordenación de datos HPTML y enriquecimiento de datos MC

Tab. 5.7.: Modelos HPTML obtenidos con diferentes conjuntos de datos frente a heurísticas alternativas.

Modelo	Datos	Heu.	$N._{reac}$	$N._{pares}$	R^2	SEE	F	p-valor
ML		H_0	332	0	0.55	37.1	59.2	$< 0{,}05$
HPTML		H_1	332	107626	0.81	29.5	1332.2	$< 0{,}05$
		H_2	332	107626	0.64	39.3	603.0	$< 0{,}05$
HPTML	MC	H_1	332	109298	0.96	13.5	7560.6	$< 0{,}05$
	MC	H_2	332	109298	0.66	38.7	631.4	$< 0{,}05$

Aplicar MC mejoró los valores $R2 = 0{,}96$ y $SEE = 13{,}5$ del modelo HPTML H_1. Además, el modelo HPTML proporciona automáticamente la reacción de referencia más similar del conjunto de datos de referencia, lo que podría brindar algunas pistas sobre el posible mecanismo de reacción. Por el contrario, el modelo ML clásico no brinda información sobre el mecanismo de reacción plausible o reacciones similares en la literatura. En general, estos resultados justifican el uso del algoritmo HPTML en lugar del clásico algoritmo ML.

La estrategia por pares puede aumentar drásticamente el número de casos, a medida que pasa de conjuntos de datos con n elementos (reacciones) a n x n elementos (pares de reacciones). En este caso, pasamos de $n_{reacc} = 332$ reacciones a $n_{pairs} =$

107.626 pares de reacciones, lo que podría ser una ventaja del modelo PTML, ya que aumentar la cantidad de elementos para entrenar el modelo ML puede mejorar el aprendizaje. En este sentido, se pueden utilizar técnicas de generación de datos sintéticos para paliar la escasez de subconjuntos de datos poco poblados. Aunque, en cualquier caso, la abundancia total de cada subconjunto de datos enriquecidos debe permanecer esencialmente constante para evitar la creación de artefactos en los datos.

Entorno de simulación

Finalmente, además de poner disponible en un repositorio público el paquete RMarkovTI, los dos modelos propuestos en esta tesis doctoral también se encuentran disponibles en una herramienta web pública `https://cptmltool.rnasa-imedir.com/CPTMLTools-Web/` para su uso y validación por la comunidad científica.

La herramienta web que se presenta se ha desarrollado integrando varias tecnologías. Para la implementación del *frontend* se ha utilizado *Thymeleaf*, motor de plantillas de código abierto para aplicaciones web escritas en Java construído sobre estándares HTML5. Para el *backend* se utilizaron, por un lado el framework *Spring Boot*, que facilita el desarrollo rápido y sencillo de aplicaciones Java y por otro lado el lenguaje de programación R, que dispone de potentes librerías como ChemmineR y su expansión ChemmineOBE, diseñadas para el análisis y la visualización de datos químicos y bioquímicos, así como la ya mencionada RMarkovTI. Estos paquetes son ampliamente utilizados por investigadores en el campo de la química para el análisis de datos, identificación de compuestos bioactivos y la generación de conocimiento en el descubrimiento de fármacos. Ofrece múltiples funcionalidades entre las que destacan la de manipular y representar moléculas, calcular descriptores moleculares, realizar agrupamiento y análisis de similitud molecular, y construir modelos predictivos, además de realizar análisis de estructura-actividad.

El entorno de simulación está desplegado sobre un servidor con sistema operativo CentOS version 8, con Docker, Docker Compose, Apache Tomcat y NGINX como proxy inverso como herramientas de apoyo principales.

Docker

Es una plataforma de código abierto diseñada para facilitar la creación, implementación y ejecución de aplicaciones utilizando contenedores, que son entornos ligeros y aislados que encapsulan una aplicación y todas sus dependencias. Esto permite ejecuciones de manera consistente en cualquier entorno, ya sea en una computadora local, un servidor en la nube o un clúster de servidores. Además proporciona una forma fácil y eficiente de empaquetar una aplicación junto con todas, las librerías y dependencias necesarias en un contenedor, que puede ser desplegado y ejecutado en cualquier sistema operativo compatible, lo que garantiza que la aplicación se ejecute de manera confiable y sin problemas, sin importar las diferencias en los entornos de desarrollo o producción. También proporciona herramientas y APIs para gestionar la creación, distribución y ejecución de contenedores. Esto facilita el despliegue y la escalabilidad de aplicaciones, ya que los contenedores pueden ser fácilmente creados, compartidos y desplegados en diferentes entornos sin preocuparse por las diferencias de configuración y dependencias [238]. En la Figura 5.7 se muestra la arquitectura de Docker.

Docker Compose

Es una herramienta que se utiliza en conjunto con Docker y que facilita la gestión y el despliegue de aplicaciones compuestas por múltiples servicios en contenedores Docker. Proporciona una forma declarativa de definir y administrar la configuración de la aplicación, lo que simplifica el proceso de desarrollo, despliegue y escalado de aplicaciones basadas en contenedores. Permite describir la configuración de una aplicación (dependencias, comunicaciones) que consta de varios servicios, que pueden incluir contenedores individuales que representan diferentes componentes de la aplicación, como una base de datos, un servidor web, etc, en un archivo YAML [238].

Open Babel

Es una biblioteca de código abierto diseñada para la química computacional y la modelización molecular. Proporciona una plataforma para la conversión de datos químicos entre diferentes formatos, así como para el análisis y manipulación de estructuras moleculares. Es capaz de leer y escribir una amplia gama de formatos de

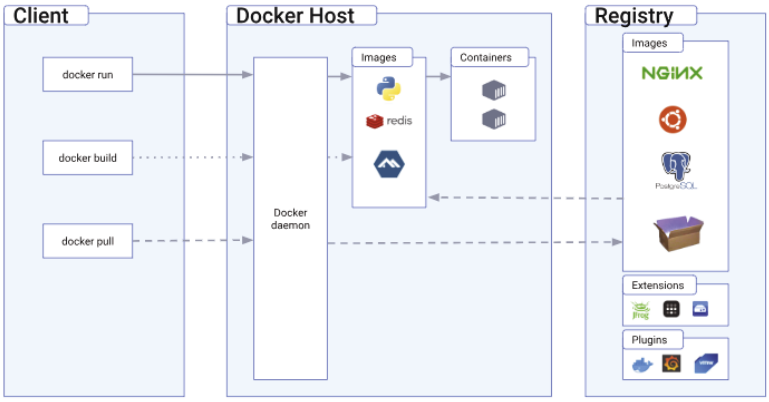

Fig. 5.7.: Arquitectura Docker. Imagen tomada de referencia principal.

archivo químico como SMILES, Mol2, PDB, SDF, etc. Esto permite a los investigadores y desarrolladores leer datos químicos de diferentes fuentes y exportarlos a los formatos requeridos por sus herramientas y aplicaciones. Además ofrece una amplia gama de funciones y algoritmos para el procesamiento y análisis de estructuras moleculares y permite realizar operaciones como la generación de conformaciones, el cálculo de propiedades físico-químicas, la búsqueda de subestructuras, la determinación de similitud molecular y la predicción de propiedades biológicas [239].

Nginx

Es un servidor web ligero de código abierto, así como un servidor proxy inverso y un servidor de correo electrónico proxy. Fue creado por Igor Sysoev en 2004 y desde entonces ha ganado popularidad debido a su rendimiento, escalabilidad y capacidad para manejar altas cargas de tráfico. Está diseñado para manejar de manera eficiente múltiples solicitudes concurrentes, lo que lo hace especialmente adecuado para servir contenido estático, así como para actuar como un proxy inverso, actuando así como una capa intermedia entre clientes y servidores de aplicaciones, equilibrando la carga de las solicitudes y mejorando el rendimiento y la disponibilidad de los servicios [240].

MDCalc Web Server

Los resultados principales de la propuesta de modelo bioinformático de la presente tesis doctoral está disponible de forma abierta y se ha llamado *MDCalc Web Server*. Hasta donde conoce la autora de la presente Tesis Doctoral, es la primera herramienta

pública para el cálculo en línea de MCD. El servidor web está disponible en línea en el siguiente enlace `https://cptmltool.rnasa-imedir.com/CPTMLTools-Web/`. En la Figura 5.8, se muestra la interfaz gráfica en la que prima la usabilidad y facilidad de uso.

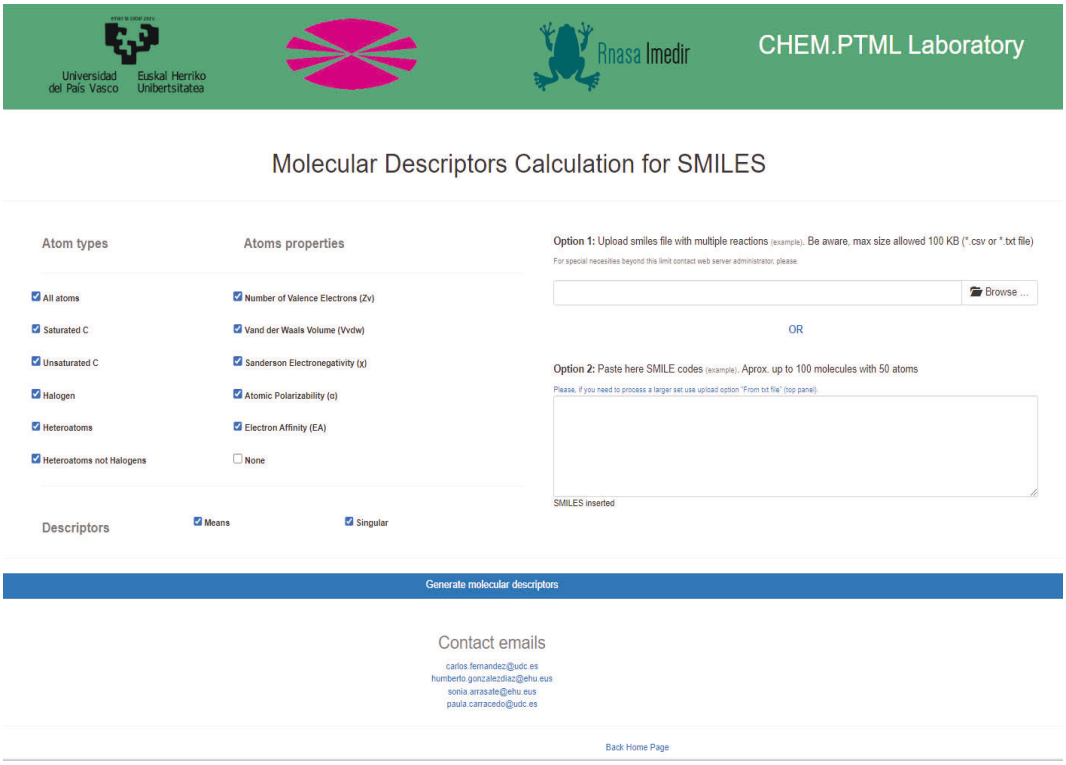

Fig. 5.8.: Interfaz Web de MDCalc.

Además, se ha desarrollado el paquete RMarkovTI (disponible en GitHub en: `https://github.com/muntisa/RMarkovTI`) con el objetivo de implementar, en el lenguaje de programación R, un paquete que permita calcular los siguientes descriptores: promedios de Markov y autovalores de Markov.

Además, se han desarrollado sendas aplicaciones de escritorio para uso fuera de línea que están disponibles, previa solicitud razonada a los autores y cuyo uso se recomienda en caso de que no tener conexión a Internet o si se producen fallos en el servidor. En ese caso, el requisito para el uso de la versión de escritorio es que el usuario debe tener instalada la máquina virtual Java en su propio equipo.

MATEO Web Server

Los resultados principales de la propuesta de modelo de síntesis enantioselectiva por reacción de α-amidoalquilación catalizada por ácidos de Brønsted de la presente

tesis doctoral está disponible de forma abierta y se ha llamado *MATEO Web Server*. El modelo modelo HPTML H_1 está disponible para uso público en el siguiente enlace: `https://cptmltool.rnasa-imedir.com/CPTMLTools-Web/mateo`.

La interfaz gráfica se muestra en la Figura 5.9. Los usuarios pueden cargar sus propios conjuntos de reacciones de consulta para predecir los valores de $ee_R(\%)_{qrcalc}$ en diferentes condiciones experimentales (disolvente, tiempo, temperatura, carga del catalizador) .

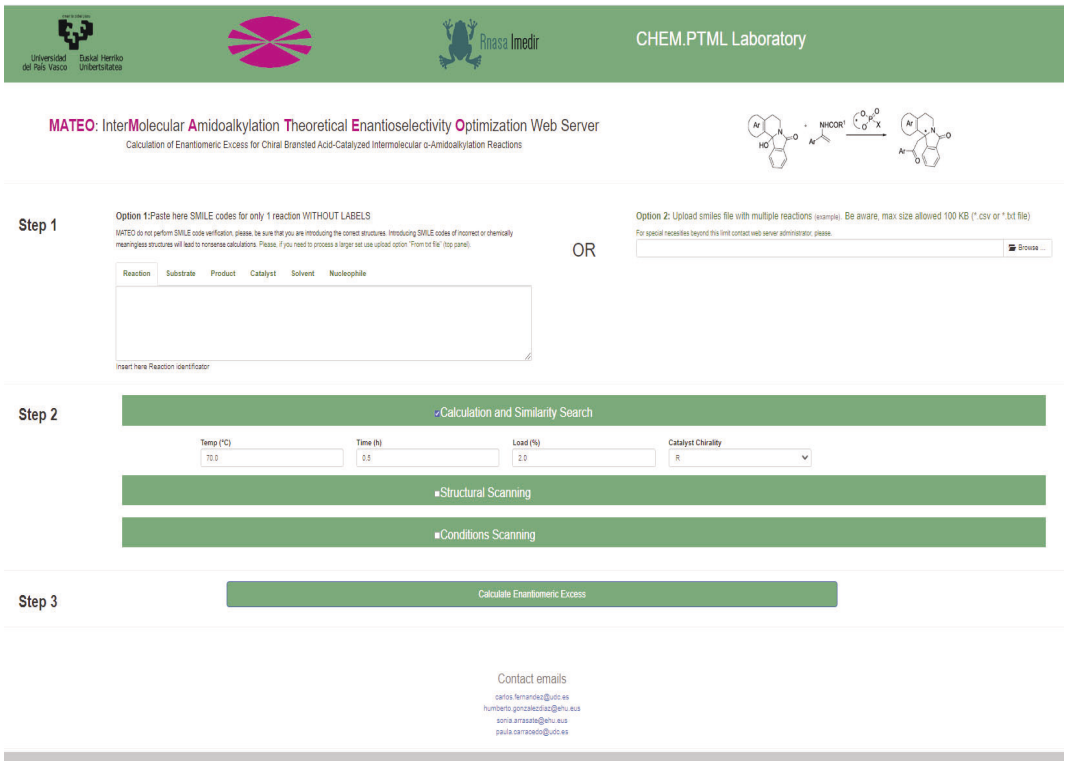

Fig. 5.9.: Interfaz Web de MATEO.

El procedimiento a seguir a la hora de utilizar MATEO se puede resumir en los siguientes pasos:

- Paso 1: se realiza la carga de las estructuras químicas de todas las moléculas involucradas en la reacción, que han de estar codificadas en formato SMILES (del inglés Simplified Molecular Input Line Entry System), que es una notación lineal y compacta utilizada para representar estructuras moleculares químicas de una manera legible por la máquina. El servidor permite cargar grandes colecciones de reacciones con diferentes combinaciones de sustrato, nucleófilo y catalizador. Esto podría ser útil para explorar grandes bibliotecas de moléculas

y/o para el diseño de nuevos catalizadores. El servidor también permite cargar la estructura del disolvente, lo que facilita la exploración.

- Paso 2: se pueden seleccionar varios tipos de cálculo:

  - **Búsqueda de similitud**: permite predecir los valores de exceso enantiomérico, además de obtener un informe de las reacciones más similares de las referencias en nuestro conjunto de datos.

  - **Exploración estructural**: permite cargar las estructuras específicas (sustrato, nucleófilo, catalizador y/o disolvente) y ejecutar un escaneo de estas moléculas en condiciones de reacción similares a las reportadas en la literatura.

  - **Exploración de condiciones**: permite mantener constantes los parámetros de la estructura (las mismas moléculas), mientras que el software realiza un escaneo de diferentes combinaciones de variables de entrada (temperatura, tiempo, carga del catalizador).

6 Conclusiones

Las conclusiones extraídas tras el desarrollo de la presente Tesis Doctoral se presentan en este capítulo. Las publicaciones de la tesis presentadas en el Anexo A (Producción), apoyaron la obtención de las mismas.

El principal objetivo de la presente tesis doctoral es el análisis de las posibles interacciones de compuestos biológicamente activos frente al cáncer de colon a partir de la estructura del compuesto para tratar de predecir así el resultado obtenido en ensayos clínicos, e identificar aquellos que mejor cumplen como potenciales dianas terapéuticas de la enfermedad. Además, implementando los mejores modelos de manera abierta en servidores web de altas prestaciones que permitan obtener nuevas predicciones sobre futuros compuestos no conocidos durante la generación del modelo. La investigación realizada ha permitido llegar a las siguientes conclusiones:

- A pesar de que los MCD se han utilizado previamente para resolver problemas en quimioinformática, algunos de ellos no están disponibles de forma abierta, únicamente en soluciones de pago. Esto dificulta mucho su validación por parte de la comunidad científica, así como su uso. Como parte del trabajo de esta Tesis Doctoral, se ha desarrollado la primera biblioteca en el lenguaje de programación abierto R que permite realizar el cálculo de nuevos MCD, llamados valores singulares de Markov $SV_k(w,g)$, junto con una clase de MCD llamados valores medios de Markov $D_k(w,g)$.

- También se ha implementado el primer servidor web público para el cálculo de MCD en línea. Esta herramienta en línea se llama MCDCalc.

- Se ha demostrado que $SV_k(w,g)$ es útil para predecir el exceso enantiomérico $ee(\,\%)[Rcat]$ para reacciones de α-amidoalquilación o para la predicción de la actividad de compuestos contra el cáncer colorrectal. En el caso de estudio de reactividad química, las reacciones tienen un mecanismo complejo dependiendo de varios factores. El modelo generado incluye MCD de el sustrato, disolvente, catalizador quiral, producto junto con valores de tiempo de reacción, temperatura, carga de catalizador, etc.

- Se validaron estos nuevos MCDs mediante diferentes modelos de ML para regresión, el mejor resultado se obtuvo con un modelo RF. El estudio de las propiedades biológicas puede conducir a una alternativa para el diseño rápido y racional del diseño de fármacos contra el cáncer colorrectal para diferentes organismos y líneas celulares.

- Se ha demostrado que los modelos ML lineales clásicos no son muy precisos para predecir la enantioselectividad de reacciones de α-amidoalquilación utilizando como entrada propiedades fisicoquímicas calculadas con un enfoque de cadenas de Markov. Además, estos modelos ML lineales no permiten detectar la reacción más similar directamente a partir del modelo. El modelo propuesto basado en teoría de la perturbación, PTML supera al modelo ML lineal clásico utilizando el mismo conjunto de datos y descriptores moleculares.

- Es más, incrementar la complejidad del modelo añadiendo una heurística, modelo HPTML, permite la detección directa de las reacciones de referencia más similares, respondiendo muy bien además en experimentos computacionales con series de validación.

- El modelo HPTML reproduce muy bien los valores experimentales de una nueva serie de reacciones generadas y, por lo tanto, estudiadas experimentalmente por primera vez en esta Tesis Doctoral.

- Finalmente, se ha implementado el mejor modelo HPTML en un servidor web llamado MATEO. De esta forma, el algoritmo esté disponible para uso público con una interfaz fácil de usar.

7 Futuros desarrollos

Se presentan en el último capítulo de esta Tesis Doctoral los futuros desarrollos que podrían ser abordados a partir del trabajo de investigación realizado a lo largo de estos años.

Los resultados obtenidos durante el desarrollo de esta Tesis Doctoral permiten que futuros investigadores puedan plantear nuevas preguntas y que, además, sean abordadas en futuras investigaciones sobre diferentes patologías. Parece claro además, que los modelos de ML pueden servir de apoyo en la toma de decisiones clínicas pero que es posible incrementar la complejidad de los mismos para mejorar los resultados, es un campo abierto en la actualidad. En pocos años, será posible ver como nuevos modelos/tecnologías compartirán el peso en la toma de decisiones de los expertos clínicos, automatizando además la gestión de los tratamientos en busca de aquellos que mejor respuesta podrían dar en cada caso en concreto.

Los modelos propuestos están disponibles de forma pública en forma de herramienta web o paquete de software y se ha podido comprobar la utilidad de MCDs, tanto como modelo quimioinformático como para predecir el exceso enantiomérico $ee(\%)[Rcat]$ para reacciones de α-amidoalquilación o para la predicción de la actividad de compuestos contra el cáncer colorrectal. Dado además el marcado carácter multidisciplinar de la tesis, partiendo de datos generados en entorno de laboratorio real, en el futuro se podrían incluir nuevas reacciones desconocidas. No solo eso, sino que nuevos procedimientos de laboratorio podrían ser validados mediante aproximaciones similares a las aquí propuestas.

En concreto, en lo que respecta al uso de técnicas de ML, se ha podido observar que muchos de los estudios actualmente publicados sufren de limitaciones en cuanto a la estandarización metodológica que proponen, además de no ser fácilmente reproducibles por hacer uso de descriptores que no están abiertos. Como posibilidad de trabajo a futuro, podrían implementarse más descriptores moleculares para que pudiesen ser utilizados y comprobados por la comunidad científica.

Si bien es cierto que los resultados obtenidos por los modelos propuestos son muy buenos, es necesario insistir en la necesidad de incrementar la complejidad de los modelos, en este caso, una posibilidad a estudiar en el futuro sería la inclusión de técnicas de extracción o selección de características que potencien el uso de aquellos descriptores moleculares que realmente aporten conocimiento a los modelos. Esto ayudará en el desarrollo de fármacos y a identificar nuevos potenciales tratamientos para enfermedades complejas y multifactoriales.

Referencias

[1] MJ Gunter, S Alhomoud, M Arnold, H Brenner, J Burn, G Casey, AT Chan, AJ Cross, E Giovannucci, R Hoover y col. «Meeting report from the joint IARC–NCI international cancer seminar series: a focus on colorectal cancer». En: *Annals of Oncology* 30.4 (2019), págs. 510-519 (vid. pág. 1).

[2] SVS Deo, Jyoti Sharma y Sunil Kumar. «GLOBOCAN 2020 report on global cancer burden: challenges and opportunities for surgical oncologists». En: *Annals of Surgical Oncology* 29.11 (2022), págs. 6497-6500 (vid. pág. 9).

[3] Hyuna Sung, Jacques Ferlay, Rebecca L Siegel, Mathieu Laversanne, Isabelle Soerjomataram, Ahmedin Jemal y Freddie Bray. «Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries». En: *CA: a cancer journal for clinicians* 71.3 (2021), págs. 209-249 (vid. págs. 10, 11).

[4] Freddie Bray, Jacques Ferlay, Isabelle Soerjomataram, Rebecca L Siegel, Lindsey A Torre y Ahmedin Jemal. «Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries». En: *CA: a cancer journal for clinicians* 68.6 (2018), págs. 394-424 (vid. pág. 10).

[5] NaNa Keum y Edward Giovannucci. «Global burden of colorectal cancer: emerging trends, risk factors and prevention strategies». En: *Nature reviews Gastroenterology & hepatology* 16.12 (2019), págs. 713-732 (vid. págs. 11, 16, 27).

[6] Yu Tian, Elham Kharazmi, Kristina Sundquist, Jan Sundquist, Hermann Brenner y Mahdi Fallah. «Familial colorectal cancer risk in half siblings and siblings: nationwide cohort study». En: *bmj* 364 (2019) (vid. pág. 11).

[7] Fay Kastrinos, N Jewel Samadder y Randall W Burt. «Use of family history and genetic testing to determine risk of colorectal cancer». En: *Gastroenterology* 158.2 (2020), págs. 389-403 (vid. págs. 11, 25).

[8] *Redecan*. `https://redecan.org/es`. Último acceso 01/05/2023 (vid. pág. 12).

[9] Valerie Gausman, David Dornblaser, Sanya Anand, Richard B Hayes, Kelli O'Connell, Mengmeng Du y Peter S Liang. «Risk factors associated with early-onset colorectal cancer». En: *Clinical Gastroenterology and Hepatology* 18.12 (2020), págs. 2752-2759 (vid. pág. 12).

[10] Paivi Peltomäki, Alisa Olkinuora y Taina T Nieminen. «Updates in the field of hereditary nonpolyposis colorectal cancer». En: *Expert Review of Gastroenterology & Hepatology* 14.8 (2020), págs. 707-720 (vid. pág. 13).

[11] Rebecca L Siegel, Christopher Dennis Jakubowski, Stacey A Fedewa, Anjee Davis y Nilofer S Azad. «Colorectal cancer in the young: epidemiology, prevention, management». En: *American Society of Clinical Oncology Educational Book* 40 (2020), e75-e88 (vid. pág. 13).

[12] Shailja C Shah y Steven H Itzkowitz. «Colorectal cancer in inflammatory bowel disease: Mechanisms and management». En: *Gastroenterology* 162.3 (2022), págs. 715-730 (vid. pág. 13).

[13] Konstantinos I Avgerinos, Nikolaos Spyrou, Christos S Mantzoros y Maria Dalamaga. «Obesity and cancer risk: Emerging biological mechanisms and perspectives». En: *Metabolism* 92 (2019), págs. 121-135 (vid. pág. 14).

[14] Samradhi Singh, Poonam Sharma, Devojit Kumar Sarma, Manoj Kumawat, Rajnarayan Tiwari, Vinod Verma, Ravinder Nagpal y Manoj Kumar. «Implication of Obesity and Gut Microbiome Dysbiosis in the Etiology of Colorectal Cancer». En: *Cancers* 15.6 (2023), pág. 1913 (vid. pág. 14).

[15] Heinz Freisling, Melina Arnold, Isabelle Soerjomataram, Mark George O'Doherty, José Manuel Ordóñez-Mena, Christina Bamia, Ellen Kampman, Michael Leitzmann, Isabelle Romieu, Frank Kee y col. «Comparison of general obesity and measures of body fat distribution in older adults in relation to cancer risk: meta-analysis of individual participant data of seven prospective cohorts in Europe». En: *British journal of cancer* 116.11 (2017), págs. 1486-1497 (vid. pág. 14).

[16] Md Sanower Hossain, Hidayah Karuniawati, Ammar Abdulrahman Jairoun, Zannat Urbi, Der Jiun Ooi, Akbar John, Ya Chee Lim, KM Kaderi Kibria, AKM Mohiuddin, Long Chiau Ming y col. «Colorectal cancer: a review of carcinogenesis, global epidemiology, current challenges, risk factors, preventive and treatment strategies». En: *Cancers* 14.7 (2022), pág. 1732 (vid. págs. 14, 31).

[17] Prashanth Rawla, Tagore Sunkara y Adam Barsouk. «Epidemiology of colorectal cancer: incidence, mortality, survival, and risk factors». En: *Gastroenterology Review/Przegląd Gastroenterologiczny* 14.2 (2019), págs. 89-103 (vid. pág. 15).

[18] Edoardo Botteri, Elisa Borroni, Erica K Sloan, Vincenzo Bagnardi, Cristina Bosetti, Giulia Peveri, Claudia Santucci, Claudia Specchia, Piet van den Brandt, Silvano Gallus y col. «Smoking and colorectal cancer risk, overall and by molecular subtypes: a meta-analysis». En: *Official journal of the American College of Gastroenterology| ACG* 115.12 (2020), págs. 1940-1949 (vid. pág. 15).

[19] Kathryn E Bradbury, Neil Murphy y Timothy J Key. «Diet and colorectal cancer in UK Biobank: a prospective study». En: *International journal of epidemiology* 49.1 (2020), págs. 246-258 (vid. pág. 15).

[20] Kathleen Yaus Wolin, Yan Yan, Graham A Colditz e IM Lee. «Physical activity and colon cancer prevention: a meta-analysis». En: *British journal of cancer* 100.4 (2009), págs. 611-616 (vid. pág. 15).

[21] Susan E Steck y E Angela Murphy. «Dietary patterns and cancer risk». En: *Nature Reviews Cancer* 20.2 (2020), págs. 125-138 (vid. pág. 15).

[22] Edoardo Botteri, Nathalie C Støer, Solveig Sakshaug, Sidsel Graff-Iversen, Siri Vangen, Solveig Hofvind, Thomas De Lange, Vincenzo Bagnardi, Giske Ursin y Elisabete Weiderpass. «Menopausal hormone therapy and colorectal cancer: a linkage between nationwide registries in Norway». En: *BMJ open* 7.11 (2017), e017639 (vid. pág. 16).

[23] Rebecca SY Wong. «Role of nonsteroidal anti-inflammatory drugs (NSAIDs) in cancer prevention and cancer promotion». En: *Advances in pharmacological sciences* 2019 (2019) (vid. pág. 16).

[24] Priyanka Kanth y John M Inadomi. «Screening and prevention of colorectal cancer». En: *Bmj* 374 (2021) (vid. pág. 16).

[25] Evelien Dekker, Pieter J Tanis, JL Vleugels, Pashtoon M Kasi y Michael Wallace. «Pure-AMC». En: *Lancet* 394 (2019), págs. 1467-80 (vid. págs. 16, 17).

[26] «The efficacy of chemopreventive agents on the incidence of colorectal adenomas: A systematic review and network meta-analysis». En: *Preventive Medicine* 162 (2022), pág. 107169 (vid. págs. 16, 17).

[27] Maaike Buskermolen, Dayna R Cenin, Lise M Helsingen, Gordon Guyatt, Per Olav Vandvik, Ulrike Haug, Michael Bretthauer e Iris Lansdorp-Vogelaar. «Colorectal cancer screening with faecal immunochemical testing, sigmoidoscopy or colonoscopy: a microsimulation modelling study». En: *Bmj* 367 (2019) (vid. pág. 17).

[28] The Lancet Gastroenterology Hepatology. *Colorectal cancer screening: is earlier better?* 2018 (vid. pág. 17).

[29] Veronika Fedirko, Elio Riboli, Anne Tjønneland, Pietro Ferrari, Anja Olsen, H Bas Bueno-de Mesquita y col. «Prediagnostic 25-hydroxyvitamin D, VDR and CASR polymorphisms, and survival in patients with colorectal cancer in western European populations». En: *Cancer Epidemiology Biomarkers & Prevention* 21.4 (2012), págs. 582-593. DOI: 10.1158/1055-9965.epi-11-1118 (vid. pág. 17).

[30] Enrique Quintero, Maria Carrillo, Maria-Luisa Leoz, Joaquín Cubiella, Carla Gargallo, Angel Lanas y col. «Risk of Advanced Neoplasia in First-Degree Relatives with Colorectal Cancer: A Large Multicenter Cross-Sectional Study». En: *PLOS Medicine* 13.5 (2016), e1002008. DOI: 10.1371/journal.pmed.1002008 (vid. pág. 18).

[31] Nilesh J Samadder, Karen Curtin, Therese M F Tuohy, Kerry G Rowe, Gerald P Mineau, Ken R Smith y col. «Increased Risk of Colorectal Neoplasia Among Family Members of Patients With Colorectal Cancer: A Population-Based Study in Utah». En: *Gastroenterology* 147.4 (2014), 814-821.e5. DOI: 10.1053/j.gastro.2014.06.001 (vid. pág. 18).

[32] J. Bernal, J. Sánchez y F. Vilariño. «Towards automatic polyp detection with a polyp appearance model». En: *Pattern Recognition* 45.9 (2012). Best Papers of Iberian Conference on Pattern Recognition and Image Analysis (IbPRIA'2011), págs. 3166-3182. ISSN: 0031-3203. DOI: https://doi.org/10.1016/j.patcog.2012.03.002 (vid. pág. 20).

[33] Bert Vogelstein, Eric R Fearon, Stanley R Hamilton, Scott E Kern, Ann C Preisinger, Mark Leppert, Alida MM Smits y Johannes L Bos. «Genetic alterations during colorectal-tumor development». En: *New England Journal of Medicine* 319.9 (1988), págs. 525-532 (vid. pág. 21).

[34] AuthorFirstName1 AuthorLastName1, AuthorFirstName2 AuthorLastName2 y AuthorFirstName3 AuthorLastName3. «The molecular characteristics of colorectal cancer: Implications for diagnosis and therapy». En: *Journal Name* Volume Number.Issue Number (Publication Year), Page Range. DOI: DOI (vid. pág. 21).

[35] Ahmed Malki, Rasha Abu ElRuz, Ishita Gupta, Asma Allouch, Semir Vranic y Ala-Eddin Al Moustafa. «Molecular mechanisms of colon cancer progression and metastasis: recent insights and advancements». En: *International journal of molecular sciences* 22.1 (2020), pág. 130 (vid. pág. 22).

[36] Shuko Harada y Diana Morlote. «Molecular pathology of colorectal cancer». En: *Advances in Anatomic Pathology* 27.1 (2020), págs. 20-26 (vid. pág. 22).

[37] Jesse Fischer, Logan C Walker, Bridget A Robinson, Frank A Frizelle, James M Church y Tim W Eglinton. «Clinical implications of the genetics of sporadic colorectal cancer». En: *ANZ Journal of Surgery* 89.10 (2019), págs. 1224-1229 (vid. pág. 22).

[38] Leah H Biller, Sapna Syngal y Matthew B Yurgelun. «Recent advances in Lynch syndrome». En: *Familial Cancer* 18 (2019), págs. 211-219 (vid. pág. 22).

[39] Guia Cerretelli, Ann Ager, Mark J Arends y Ian M Frayling. «Molecular pathology of Lynch syndrome». En: *The Journal of pathology* 250.5 (2020), págs. 518-531 (vid. pág. 22).

[40] Fatima Domenica Elisa De Palma, Valeria D'argenio, Jonathan Pol, Guido Kroemer, Maria Chiara Maiuri y Francesco Salvatore. «The molecular hallmarks of the serrated pathway in colorectal cancer». En: *Cancers* 11.7 (2019), pág. 1017 (vid. pág. 22).

[41] Dan Qiao, Xiao-Yan Liu, Lie Zheng, Ya-Li Zhang, Ren-Ye Que, Bing-Jing Ge, Hong-Yan Cao y Yan-Cheng Dai. «Clinicopathological features and expression of regulatory mechanism of the Wnt signaling pathway in colorectal sessile serrated adenomas/polyps with different syndrome types». En: *World Journal of Clinical Cases* 11.9 (2023), pág. 1963 (vid. pág. 22).

[42] Amirsaeed Sabeti Aghabozorgi, Amirhossein Bahreyni, Atena Soleimani, Afsane Bahrami, Majid Khazaei, Gordon A Ferns, Amir Avan y Seyed Mahdi Hassanian. «Role of adenomatous polyposis coli (APC) gene mutations in the pathogenesis of colorectal cancer; current status and perspectives». En: *Biochimie* 157 (2019), págs. 64-71 (vid. pág. 23).

[43] Peyman Dinarvand, Elizabeth P Davaro, James V Doan, Mary E Ising, Neil R Evans, Nancy J Phillips, Jinping Lai y Miguel A Guzman. «Familial adenomatous polyposis syndrome: an update and review of extraintestinal manifestations». En: *Archives of pathology & laboratory medicine* 143.11 (2019), págs. 1382-1398 (vid. pág. 23).

[44] Mingjing Meng, Keying Zhong, Ting Jiang, Zhongqiu Liu, Hiu Yee Kwan y Tao Su. «The current understanding on the impact of KRAS on colorectal cancer». En: *Biomedicine & pharmacotherapy* 140 (2021), pág. 111717 (vid. pág. 23).

[45] Lamei Huang, Zhixing Guo, Fang Wang y Liwu Fu. «KRAS mutation: from undruggable to druggable in cancer». En: *Signal transduction and targeted Therapy* 6.1 (2021), pág. 386 (vid. pág. 23).

[46] Hui Li, Jinglin Zhang, Joanna Hung Man Tong, Anthony Wing Hung Chan, Jun Yu, Wei Kang y Ka Fai To. «Targeting the oncogenic p53 mutants in colorectal cancer and other solid tumors». En: *International journal of molecular sciences* 20.23 (2019), pág. 5999 (vid. pág. 23).

[47] Javier Ros, Iosune Baraibar, Emilia Sardo, Nuria Mulet, Francesc Salvà, Guillem Argilés, Giulia Martini, Davide Ciardiello, José Luis Cuadra, Josep Tabernero y col. «BRAF, MEK and EGFR inhibition as treatment strategies in BRAF V600E metastatic colorectal cancer». En: *Therapeutic advances in medical oncology* 13 (2021), pág. 1758835921992974 (vid. pág. 24).

[48] Julien Taieb, Alexandra Lapeyre-Prost, Pierre Laurent Puig y Aziz Zaanan. «Exploring the best treatment options for BRAF-mutant metastatic colon cancer». En: *British journal of cancer* 121.6 (2019), págs. 434-442 (vid. pág. 24).

[49] Shui-Ming Wang, Bin Jiang, Youping Deng, Shu-Liang Huang, Ming-Zhi Fang y Yu Wang. «Clinical significance of MLH1/MSH2 for stage II/III sporadic colorectal cancer». En: *World journal of gastrointestinal oncology* 11.11 (2019), pág. 1065 (vid. pág. 24).

[50] R Kanwal y S Gupta. «Epigenetic modifications in cancer». En: *Clinical genetics* 81.4 (2012), págs. 303-311 (vid. pág. 24).

[51] Bruno Augusto Alves Martins, Gabriel Fonseca De Bulhões, Igor Norat Cavalcanti, Mickaella Michelson Martins, Paulo GonçAlves De Oliveira y Aline Maria Araújo Martins. «Biomarkers in colorectal cancer: the role of translational proteomics research». En: *Frontiers in Oncology* 9 (2019) (vid. pág. 26).

[52] Olorunseun O Ogunwobi, Fahad Mahmood y Akinfemi Akingboye. «Biomarkers in colorectal cancer: current research and future prospects». En: *International journal of molecular sciences* 21.15 (2020), pág. 5311 (vid. pág. 26).

[53] Gerhard Jung, Eva Hernández-Illán, Leticia Moreira, Francesc Balaguer y Ajay Goel. «Epigenetics of colorectal cancer: biomarker and therapeutic potential». En: *Nature reviews Gastroenterology & hepatology* 17.2 (2020), págs. 111-130 (vid. pág. 26).

[54] Maria Marcuello, Veronika Vymetalkova, Rui PL Neves, Saray Duran-Sanchon, Hege Marie Vedeld, Emma Tham, Guus van Dalum, Georg Fluegen, Vanesa Garcia-Barberan, Remond Ja Fijneman y col. «Circulating biomarkers for early detection and clinical management of colorectal cancer». En: *Molecular aspects of medicine* 69 (2019), págs. 107-122 (vid. pág. 26).

[55] Bing Chen, Zijing Xia, Ya-Nan Deng, Yanfang Yang, Peng Zhang, Hongxia Zhu, Ningzhi Xu y Shufang Liang. «Emerging microRNA biomarkers for colorectal cancer diagnosis and prognosis». En: *Royal Society Open Biology* 9.1 (2019), pág. 180212 (vid. pág. 26).

[56] Sunny H Wong y Jun Yu. «Gut microbiota in colorectal cancer: mechanisms of action and clinical applications». En: *Nature Reviews Gastroenterology & Hepatology* 16.11 (2019), págs. 690-704 (vid. pág. 26).

[57] Cynthia L Sears y Wendy S Garrett. «Microbes, microbiota, and colon cancer». En: *Cell host & microbe* 15.3 (2014), págs. 317-328 (vid. pág. 26).

[58] Wells A Messersmith. «NCCN guidelines updates: management of metastatic colorectal cancer». En: *Journal of the National Comprehensive Cancer Network* 17.5.5 (2019), págs. 599-601 (vid. pág. 27).

[59] Leah H Biller y Deborah Schrag. «Diagnosis and treatment of metastatic colorectal cancer: a review». En: *Jama* 325.7 (2021), págs. 669-685 (vid. pág. 27).

[60] Yuan-Hong Xie, Ying-Xuan Chen y Jing-Yuan Fang. «Comprehensive review of targeted therapy for colorectal cancer». En: *Signal transduction and targeted therapy* 5.1 (2020), pág. 22 (vid. págs. 27, 30).

[61] Srinivas Patnaik y Anupriya. «Drugs targeting epigenetic modifications and plausible therapeutic strategies against colorectal cancer». En: *Frontiers in Pharmacology* 10 (2019), pág. 588 (vid. pág. 29).

[62] Christine M Parseghian, Stefania Napolitano, Jonathan M Loree y Scott Kopetz. «Mechanisms of Innate and Acquired Resistance to Anti-EGFR Therapy: A Review of Current Knowledge with a Focus on Rechallenge TherapiesInnate and Acquired Resistance to Anti-EGFR Therapy». En: *Clinical Cancer Research* 25.23 (2019), págs. 6899-6908 (vid. pág. 30).

[63] Sanjay Kumar Jain, Ankita Tiwari, Ankit Jain, Amit Verma, Shivani Saraf, Pritish Kumar Panda y Gaytri Gour. «Application potential of polymeric nanoconstructs for colon-specific drug delivery». En: (2018), págs. 22-49 (vid. pág. 31).

[64] Tolou Hosseinifar, Simin Sheybani, Majid Abdouss, Sayed Alireza Hassani Najafabadi y Mehdi Shafiee Ardestani. «Pressure responsive nanogel base on alginate-cyclodextrin with enhanced apoptosis mechanism for colon cancer delivery». En: *Journal of Biomedical Materials Research Part A* 106.2 (2018), págs. 349-359 (vid. pág. 31).

[65] Amit Sharma, Eun-Joong Kim, Hu Shi, Jin Yong Lee, Bong Geun Chung y Jong Seung Kim. «Development of a theranostic prodrug for colon cancer therapy by combining ligand-targeted delivery and enzyme-stimulated activation». En: *Biomaterials* 155 (2018), págs. 145-151 (vid. pág. 31).

[66] Samantha L Ginn, Anais K Amaya, Ian E Alexander, Michael Edelstein y Mohammad R Abedi. «Gene therapy clinical trials worldwide to 2017: An update». En: *The journal of gene medicine* 20.5 (2018), e3015 (vid. pág. 31).

[67] David Zahavi y Louis Weiner. «Monoclonal antibodies in cancer therapy». En: *Antibodies* 9.3 (2020), pág. 34 (vid. pág. 31).

[68] Lindsey Carlsen, Kelsey E Huntington y Wafik S El-Deiry. «Immunotherapy for colorectal cancer: mechanisms and predictive biomarkers». En: *Cancers* 14.4 (2022), pág. 1028 (vid. pág. 31).

[69] Anna Nappi, Massimiliano Berretta, Carmela Romano, Salvatore Tafuto, Antonino Cassata, Rossana Casaretti, Lucrezia Silvestro, C Divitiis, Lara Alessandrini, Francesco Fiorica y col. «Metastatic Colorectal Cancer: Role of Target Therapies and Future Perspectives.» En: *Current cancer drug targets* 18.5 (2018), págs. 421-429 (vid. pág. 31).

[70] Junyong Weng, Shanbao Li, Zhonglin Zhu, Qi Liu, Ruoxin Zhang, Yufei Yang y Xinxiang Li. «Exploring immunotherapy in colorectal cancer». En: *Journal of hematology & oncology* 15.1 (2022), págs. 1-28 (vid. pág. 31).

[71] Do-Youn Oh y Yung-Jue Bang. «HER2-targeted therapies—a role beyond breast cancer». En: *Nature Reviews Clinical Oncology* 17.1 (2020), págs. 33-48 (vid. pág. 31).

[72] Ryo Ejima, Hiroyuki Suzuki, Tomohiro Tanaka, Teizo Asano, Mika K Kaneko y Yukinari Kato. «Development of a Novel Anti-CD44 Variant 6 Monoclonal Antibody C44Mab-9 for Multiple Applications against Colorectal Carcinomas». En: *International Journal of Molecular Sciences* 24.4 (2023), pág. 4007 (vid. pág. 31).

[73] Heinz-Josef Lenz, Eric Van Cutsem, Maria Luisa Limon, Ka Yeung Mark Wong, Alain Hendlisz, Massimo Aglietta, Pilar García-Alfonso, Bart Neyns, Gabriele Luppi, Dana B Cardin y col. «First-line nivolumab plus low-dose ipilimumab for microsatellite instability-high/mismatch repair-deficient metastatic colorectal cancer: the phase II CheckMate 142 study». En: *Journal of Clinical Oncology* 40.2 (2022), págs. 161-170 (vid. pág. 32).

[74] Marwan Fakih, Jaideep Sandhu, Dean Lim, Xiaochen Li, Sierra Li y Chongkai Wang. «Regorafenib, Ipilimumab, and Nivolumab for Patients With Microsatellite Stable Colorectal Cancer and Disease Progression With Prior Chemotherapy: A Phase 1 Nonrandomized Clinical Trial». En: *JAMA oncology* (2023) (vid. pág. 32).

[75] Karuna Ganesh, Zsofia K Stadler, Andrea Cercek, Robin B Mendelsohn, Jinru Shia, Neil H Segal y Luis A Diaz Jr. «Immunotherapy in colorectal cancer: rationale, challenges and potential». En: *Nature reviews Gastroenterology & hepatology* 16.6 (2019), págs. 361-375 (vid. pág. 32).

[76] Amanda L Wooster, Lydia H Girgis, Hayley Brazeale, Trevor S Anderson, Laurence M Wood y Devin B Lowe. «Dendritic cell vaccine therapy for colorectal cancer». En: *Pharmacological research* 164 (2021), pág. 105374 (vid. pág. 32).

[77] Lewis Au, Annika Fendler, Scott TC Shepherd, Karolina Rzeniewicz, Maddalena Cerrone, Fiona Byrne, Eleanor Carlyle, Kim Edmonds, Lyra Del Rosario, John Shon y col. «Cytokine release syndrome in a patient with colorectal cancer after vaccination with BNT162b2». En: *Nature medicine* 27.8 (2021), págs. 1362-1366 (vid. pág. 32).

[78] Yvan Saeys, Iñaki Inza y Pedro Larrañaga. «A review of feature selection techniques in bioinformatics». En: *bioinformatics* 23.19 (2007), págs. 2507-2517 (vid. pág. 40).

[79] Sheng Tian, Junmei Wang, Youyong Li, Dan Li, Lei Xu y Tingjun Hou. «The application of in silico drug-likeness predictions in pharmaceutical research». En: *Advanced drug delivery reviews* 86 (2015), págs. 2-10 (vid. pág. 44).

[80] Michael E Burton. *Applied pharmacokinetics & pharmacodynamics: principles of therapeutic drug monitoring*. Lippincott Williams & Wilkins, 2006 (vid. pág. 44).

[81] Ke Liu, Xiangyan Sun, Lei Jia, Jun Ma, Haoming Xing, Junqiu Wu, Hua Gao, Yax Sun, Florian Boulnois y Jie Fan. «Chemi-Net: a molecular graph convolutional network for accurate drug property prediction». En: *International journal of molecular sciences* 20.14 (2019), pág. 3389 (vid. pág. 45).

[82] Samarjeet Prasad y Bernard R Brooks. «A deep learning approach for the blind logP prediction in SAMPL6 challenge». En: *Journal of Computer-Aided Molecular Design* (2020), págs. 1-8 (vid. pág. 45).

[83] Jing Lu, Dong Lu, Xiaochen Zhang, Yi Bi, Keguang Cheng, Mingyue Zheng y Xiaomin Luo. «Estimation of elimination half-lives of organic chemicals in humans using gradient boosting machine». En: *Biochimica et Biophysica Acta (BBA)-General Subjects* 1860.11 (2016), págs. 2664-2671 (vid. pág. 45).

[84] Ignacio Aliagas, Alberto Gobbi, Timothy Heffron, Man-Ling Lee, Daniel F Ortwine, Mark Zak y S Cyrus Khojasteh. «A probabilistic method to report predictions from a human liver microsomes stability QSAR model: a practical tool for drug discovery». En: *Journal of Computer-Aided Molecular Design* 29.4 (2015), págs. 327-338 (vid. págs. 45, 59).

[85] Haodi Li, Buzhou Tang, Qingcai Chen, Kai Chen, Xiaolong Wang, Baohua Wang y Zhe Wang. «HITSZ_CDR: an end-to-end chemical and disease relation extraction system for BioCreative V». En: *Database* 2016 (2016) (vid. pág. 46).

[86] Baofang Hu, Hong Wang, Lutong Wang y Weihua Yuan. «Adverse Drug Reaction Predictions Using Stacking Deep Heterogeneous Information Network Embedding Approach». En: *Molecules* 23.12 (2018), pág. 3193 (vid. pág. 46).

[87] Wen Zhang, Feng Liu, Longqiang Luo y Jingxia Zhang. «Predicting drug side effects by multi-label learning and ensemble learning». En: *BMC bioinformatics* 16.1 (2015), pág. 365 (vid. pág. 46).

[88] Yi-Wei Wang, Lei Huang, Si-Wen Jiang, Kan Li, Jun Zou y Sheng-Yong Yang. «CapsCarcino: A novel sparse data deep learning tool for predicting carcinogens». En: *Food and Chemical Toxicology* 135 (2020), pág. 110921 (vid. págs. 46, 64).

[89] Aman Sharma y Rinkle Rani. «BE-DTI': Ensemble framework for drug target interaction prediction using dimensionality reduction and active learning». En: *Computer methods and programs in biomedicine* 165 (2018), págs. 151-162 (vid. pág. 46).

[90] Lingwei Xie, Song He, Xinyu Song, Xiaochen Bo y Zhongnan Zhang. «Deep learning-based transcriptome data classification for drug-target interaction prediction». En: *BMC genomics* 19.7 (2018), pág. 667 (vid. págs. 46, 65).

[91] Ruolan Chen, Xiangrong Liu, Shuting Jin, Jiawei Lin y Juan Liu. «Machine learning for drug-target interaction prediction». En: *Molecules* 23.9 (2018), pág. 2208 (vid. pág. 46).

[92] Jieyao Deng, Qingjun Yuan, Hiroshi Mamitsuka y Shanfeng Zhu. «DrugE-Rank: Predicting Drug-Target Interactions by Learning to Rank». En: *Data Mining for Systems Biology*. Springer, 2018, págs. 195-202 (vid. pág. 46).

[93] Jocelyn Sunseri, Jonathan E King, Paul G Francoeur y David Ryan Koes. «Convolutional neural network scoring and minimization in the D3R 2017 community challenge». En: *Journal of computer-aided molecular design* 33.1 (2019), págs. 19-34 (vid. pág. 46).

[94] Munhwan Lee, Hyeyeon Kim, Hyunwhan Joe y Hong-Gee Kim. «Multi-channel PINN: investigating scalable and transferable neural networks for drug discovery». En: *Journal of cheminformatics* 11.1 (2019), pág. 46 (vid. págs. 46, 65).

[95] Li Li, Ching Chiek Koh, Daniel Reker, JB Brown, Haishuai Wang, Nicholas Keone Lee, Hien-haw Liow, Hao Dai, Huai-Meng Fan, Luonan Chen y col. «Predicting protein-ligand interactions based on bow-pharmacological space and Bayesian additive regression trees». En: *Scientific reports* 9.1 (2019), págs. 1-12 (vid. págs. 47, 57).

[96] Jason B Cross. «Methods for virtual screening of gpcr targets: Approaches and Challenges». En: *Computational Methods for GPCR Drug Discovery*. Springer, 2018, págs. 233-264 (vid. pág. 47).

[97] Ying Chen, JD Elenee Argentinis y Griff Weber. «IBM Watson: how cognitive computing can be applied to big data challenges in life sciences research». En: *Clinical therapeutics* 38.4 (2016), págs. 688-701 (vid. pág. 47).

[98] Gang Fu, Ying Ding, Abhik Seal, Bin Chen, Yizhou Sun y Evan Bolton. «Predicting drug target interactions using meta-path-based semantic network analysis». En: *BMC bioinformatics* 17.1 (2016), pág. 160 (vid. págs. 47, 59).

[99] Omar Kana y Michal Brylinski. «Elucidating the druggability of the human proteome with eFindSite». En: *Journal of computer-aided molecular design* 33.5 (2019), págs. 509-519 (vid. pág. 47).

[100] Ben Hu, Z Kun Kuang, Shi-Yu Feng, Dong Wang, Song-Bing He y DE Xin Kong. «Supplementary Materials: Three-Dimensional Biologically Relevant Spectrum (BRS-3D): Shape Similarity Profile Based on PDB Ligands as Molecular Descriptor». En: *Molecules* 21.11 (2016), págs. 1554-1565 (vid. pág. 47).

[101] Chang Lu, Zhe Liu, Enju Zhang, Fei He, Zhiqiang Ma y Han Wang. «MPLs-Pred: Predicting Membrane Protein-Ligand Binding Sites Using Hybrid Sequence-Based Features and Ligand-Specific Models». En: *International journal of molecular sciences* 20.13 (2019), pág. 3120 (vid. pág. 47).

[102] Xiao-Ying Yan, Shao-Wu Zhang y Chang-Run He. «Prediction of drug-target interaction by integrating diverse heterogeneous information source with multiple kernel learning and clustering methods». En: *Computational biology and chemistry* 78 (2019), págs. 460-467 (vid. págs. 47, 59).

[103] Jonathan M Stokes, Kevin Yang, Kyle Swanson, Wengong Jin, Andres Cubillos-Ruiz, Nina M Donghia, Craig R MacNair, Shawn French, Lindsey A Carfrae, Zohar Bloom-Ackerman y col. «A deep learning approach to antibiotic discovery». En: *Cell* 180.4 (2020), págs. 688-702 (vid. págs. 47, 55, 65).

[104] Sean Ekins, Adwait Anand Godbole, György Kéri, Lászlo Orfi, János Pato, Rajeshwari Subray Bhat, Rinkee Verma, Erin K Bradley y Valakunja Nagaraja. «Machine learning and docking models for Mycobacterium tuberculosis topoisomerase I». En: *Tuberculosis* 103 (2017), págs. 52-60 (vid. págs. 48, 57).

[105] Yasaman KalantarMotamedi, Richard T Eastman, Rajarshi Guha y Andreas Bender. «A systematic and prospectively validated approach for identifying synergistic drug combinations against malaria». En: *Malaria journal* 17.1 (2018), pág. 160 (vid. pág. 48).

[106] Birgit Viira, Thibault Gendron, Don Antoine Lanfranchi, Sandrine Cojean, Dragos Horvath, Gilles Marcou, Alexandre Varnek, Louis Maes, Uko Maran, Philippe M Loiseau y col. «In silico mining for antimalarial structure-activity knowledge and discovery of novel antimalarial curcuminoids». En: *Molecules* 21.7 (2016), pág. 853 (vid. págs. 48, 59).

[107] James Schuler, Matthew L Hudson, Diane Schwartz y Ram Samudrala. «A systematic review of computational drug discovery, development, and repurposing for Ebola virus disease treatment». En: *Molecules* 22.10 (2017), pág. 1777 (vid. pág. 48).

[108] Yu Wei, Wei Li, Tengfei Du, Zhangyong Hong y Jianping Lin. «Targeting HIV/HCV Coinfection Using a Machine Learning-Based Multiple Quantitative Structure-Activity Relationships (Multiple QSAR) Method». En: *International journal of molecular sciences* 20.14 (2019), pág. 3572 (vid. págs. 48, 56).

[109] Ambrose Plante, Derek M Shore, Giulia Morra, George Khelashvili y Harel Weinstein. «A machine learning approach for the discovery of ligand-specific functional mechanisms of GPCRs». En: *Molecules* 24.11 (2019), pág. 2097 (vid. pág. 49).

[110] Jing Xing, Rukang Zhang, Xiangrui Jiang, Tianwen Hu, Xinjun Wang, Gang Qiao, Junjian Wang, Fengling Yang, Xiaomin Luo, Kaixian Chen y col. «Rational design of 5-((1H-imidazol-1-yl) methyl) quinolin-8-ol derivatives as novel bromodomain-containing protein 4 inhibitors». En: *European Journal of Medicinal Chemistry* 163 (2019), págs. 281-294 (vid. pág. 49).

[111] Hongao Zhang, Wei Liu, Zhihong Liu, Yingchen Ju, Mengyang Xu, Yue Zhang, Xinyu Wu, Qiong Gu, Zhong Wang y Jun Xu. «Discovery of indoleamine 2, 3-dioxygenase inhibitors using machine learning based virtual screening». En: *MedChemComm* 9.6 (2018), págs. 937-945 (vid. págs. 49, 57).

[112] Turki Turki y Jason TL Wang. «Clinical intelligence: New machine learning techniques for predicting clinical drug response». En: *Computers in biology and medicine* 107 (2019), págs. 302-322 (vid. págs. 49, 59).

[113] Cong Fan y Yanxin Huang. «Identification of novel potential scaffold for class I HDACs inhibition: An in-silico protocol based on virtual screening, molecular dynamics, mathematical analysis and machine learning». En: *Biochemical and Biophysical Research Communications* 491.3 (2017), págs. 800-806 (vid. pág. 49).

[114] Miao Yu, Qiong Gu y Jun Xu. «Discovering new PI3Kα inhibitors with a strategy of combining ligand-based and structure-based virtual screening». En: *Journal of computer-aided molecular design* 32.2 (2018), págs. 347-361 (vid. págs. 49, 57).

[115] Trieu-Du Ngo, Thanh-Dao Tran, Minh-Tri Le y Khac-Minh Thai. «Computational predictive models for P-glycoprotein inhibition of in-house chalcone derivatives and drug-bank compounds». En: *Molecular diversity* 20.4 (2016), págs. 945-961 (vid. págs. 50, 59).

[116] Alan M Sandercock, Steven Rust, Sandrine Guillard, Kris F Sachsenmeier, Nick Holoweckyj, Carl Hay, Matt Flynn, Qihui Huang, Kuan Yan, Bram Herpers y col. «Identification of anti-tumour biologics using primary tumour models, 3-D phenotypic screening and image-based multi-parametric profiling». En: *Molecular cancer* 14.1 (2015), págs. 1-18 (vid. págs. 50, 62).

[117] Qingqing Guo, Yao Luo, Shiyang Zhai, Zhenla Jiang, Chongze Zhao, Jianrong Xu y Ling Wang. «Discovery, biological evaluation, structure–activity relationships and mechanism of action of pyrazolo [3, 4-b] pyridin-6-one derivatives as a new class of anticancer agents». En: *Organic & biomolecular chemistry* 17.25 (2019), págs. 6201-6214 (vid. pág. 50).

[118] Peter Man-Un Ung, Rayees Rahman y Avner Schlessinger. «Redefining the Protein Kinase Conformational Space with Machine Learning». En: *Biophysical Journal* 116.3 (2019), 58a-59a (vid. pág. 50).

[119] Prson Gautam, Alok Jaiswal, Tero Aittokallio, Hassan Al-Ali y Krister Wennerberg. «Phenotypic screening combined with machine learning for efficient identification of breast cancer-selective therapeutic targets». En: *Cell chemical biology* 26.7 (2019), págs. 970-979 (vid. pág. 50).

[120] Silvana Valdebenito, Emil Lou, John Baldoni, George Okafo y Eliseo Eugenin. «The novel roles of connexin channels and tunneling nanotubes in cancer pathogenesis». En: *International journal of molecular sciences* 19.5 (2018), pág. 1270 (vid. pág. 50).

[121] Minji Jeon, Sunkyu Kim, Sungjoon Park, Heewon Lee y Jaewoo Kang. «In silico drug combination discovery for personalized cancer therapy». En: *BMC systems biology* 12.2 (2018), págs. 59-67 (vid. págs. 51, 62).

[122] Adityanarayanan Radhakrishnan, Karthik Damodaran, Ali C Soylemezoglu, Caroline Uhler y GV Shivashankar. «Machine learning for nuclear mechano-morphometric biomarkers in cancer diagnosis». En: *Scientific reports* 7.1 (2017), págs. 1-13 (vid. págs. 51, 65).

[123] Peter Bloomingdale y Donald E Mager. «Machine learning models for the prediction of chemotherapy-induced peripheral neuropathy». En: *Pharmaceutical Research* 36.2 (2019), pág. 35 (vid. págs. 51, 59).

[124] David Romeo-Guitart, Joaquim Forés, Mireia Herrando-Grabulosa, Raquel Valls, Tatiana Leiva-Rodríguez, Elena Galea, Francisco González-Pérez, Xavier Navarro, Valerie Petegnief, Assumpció Bosch y col. «Neuroprotective drug for nerve trauma revealed using artificial intelligence». En: *Scientific reports* 8.1 (2018), págs. 1-15 (vid. pág. 51).

[125] Chuipu Cai, Qihui Wu, Yunxia Luo, Huili Ma, Jiangang Shen, Yongbin Zhang, Lei Yang, Yunbo Chen, Zehuai Wen y Qi Wang. «In silico prediction of ROCK II inhibitors by different classification approaches». En: *Molecular diversity* 21.4 (2017), págs. 791-807 (vid. pág. 52).

[126] Man Luo, Terry-Elinor Reid y Xiang Simon Wang. «Discovery of natural product-derived 5-HT1A receptor binders by cheminfomatics modeling of known binders, high throughput screening and experimental validation». En: *Combinatorial chemistry & high throughput screening* 18.7 (2015), págs. 685-692 (vid. pág. 52).

[127] Aytun Onay, Melih Onay y Osman Abul. «Classification of nervous system withdrawn and approved drugs with ToxPrint features via machine learning strategies». En: *Computer Methods and Programs in Biomedicine* 142 (2017), págs. 9-19 (vid. págs. 52, 59).

[128] Yoan Brito-Sánchez, Yovani Marrero-Ponce, Stephen J Barigye, Iván Yaber-Goenaga, Carlos Morell Perez, Huong Le-Thi-Thu y Artem Cherkasov. «Towards better BBB passage prediction using an extensive and curated data set». En: *Molecular informatics* 34.5 (2015), págs. 308-330 (vid. pág. 52).

[129] Lindsey Burggraaff, Paul Oranje, Robin Gouka, Pieter van der Pijl, Marian Geldof, Herman WT van Vlijmen, Adriaan P IJzerman y Gerard JP van Westen. «Identification of novel small molecule inhibitors for solute carrier SGLT1 using proteochemometric modeling». En: *Journal of cheminformatics* 11.1 (2019), pág. 15 (vid. págs. 52, 62).

[130] Benny Da'adoosh, David Marcus, Anwar Rayan, Fred King, Jianwei Che y Amiram Goldblum. «Discovering highly selective and diverse PPAR-delta agonists by ligand based machine learning and structural modeling». En: *Scientific reports* 9.1 (2019), págs. 1-12 (vid. pág. 53).

[131] Jonathan J Chen, Lyndsey N Schmucker y Donald P Visco. «Pharmaceutical machine learning: Virtual high-throughput screens identifying promising and economical small molecule inhibitors of complement factor C1s». En: *Biomolecules* 8.2 (2018), pág. 24 (vid. págs. 53, 60).

[132] Yonghyun Nam, Myungjun Kim, Hang-Seok Chang y Hyunjung Shin. «Drug repurposing with network reinforcement». En: *BMC bioinformatics* 20.13 (2019), pág. 383 (vid. pág. 53).

[133] Kai Zhao y Hon-Cheong So. «Using drug expression profiles and machine learning approach for drug repurposing». En: *Computational methods for drug repurposing*. Springer, 2019, págs. 219-237 (vid. pág. 53).

[134] Yunguan Wang, Jaswanth Yella y Anil G Jegga. «Transcriptomic data mining and repurposing for computational drug discovery». En: *Computational Methods for Drug Repurposing*. Springer, 2019, págs. 73-95 (vid. pág. 53).

[135] William Mangione y Ram Samudrala. «Identifying protein features responsible for improved drug repurposing accuracies using the CANDO platform: Implications for drug design». En: *Molecules* 24.1 (2019), pág. 167 (vid. pág. 53).

[136] Thomas Olson y Rahul Singh. «Predicting anatomic therapeutic chemical classification codes using tiered learning». En: *BMC bioinformatics* 18.8 (2017), pág. 266 (vid. págs. 53, 57).

[137] Corwin Hansch y Toshio Fujita. «p-σ-π Analysis. A method for the correlation of biological activity and chemical structure». En: *Journal of the American Chemical Society* 86.8 (1964), págs. 1616-1626 (vid. pág. 53).

[138] * Ajay, W Patrick Walters y Mark A Murcko. «Can we learn to distinguish between "drug-like" and "nondrug-like" molecules?» En: *Journal of medicinal chemistry* 41.18 (1998), págs. 3314-3324 (vid. págs. 54, 64).

[139] Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner y Gabriele Monfardini. «The graph neural network model». En: *IEEE Transactions on Neural Networks* 20.1 (2008), págs. 61-80 (vid. pág. 55).

[140] Steven Kearnes, Kevin McCloskey, Marc Berndl, Vijay Pande y Patrick Riley. «Molecular graph convolutions: moving beyond fingerprints». En: *Journal of computer-aided molecular design* 30.8 (2016), págs. 595-608 (vid. págs. 55, 65).

[141] Thomas Bayes. «LII. An essay towards solving a problem in the doctrine of chances. By the late Rev. Mr. Bayes, FRS communicated by Mr. Price, in a letter to John Canton, AMFR S». En: *Philosophical transactions of the Royal Society of London* 53 (1763), págs. 370-418 (vid. pág. 56).

[142] Trevor Hastie y Robert Tibshirani. *Generalized additive models*. Wiley Online Library, 1990 (vid. pág. 56).

[143] Neel S Madhukar, Prashant K Khade, Linda Huang, Kaitlyn Gayvert, Giuseppe Galletti, Martin Stogniew, Joshua E Allen, Paraskevi Giannakakou y Olivier Elemento. «A Bayesian machine learning approach for drug target identification using diverse data types». En: *Nature communications* 10.1 (2019), págs. 1-14 (vid. pág. 56).

[144] Shuaibing He, Xuelian Zhang, Shan Lu, Ting Zhu, Guibo Sun y Xiaobo Sun. «A computational toxicology approach to screen the hepatotoxic ingredients in traditional Chinese medicines: Polygonum multiflorum Thunb as a case study». En: *Biomolecules* 9.10 (2019), pág. 577 (vid. pág. 57).

[145] Alexander L Perryman, Jimmy S Patel, Riccardo Russo, Eric Singleton, Nancy Connell, Sean Ekins y Joel S Freundlich. «Naive Bayesian models for vero cell cytotoxicity». En: *Pharmaceutical research* 35.9 (2018), pág. 170 (vid. pág. 57).

[146] Vladimir Vapnik. *Estimation of dependences based on empirical data*. Springer Science & Business Media, 2006 (vid. pág. 57).

[147] Bernhard Schölkopf, Koji Tsuda y Jean-Philippe Vert. *Kernel methods in computational biology*. MIT press, 2004 (vid. pág. 57).

[148] Carlos Fernandez-Lozano, Marcos Gestal, Nieves Pedreira-Souto, Lucian Postelnicu, Julián Dorado y Cristian Robert Munteanu. «Kernel-based feature selection techniques for transport proteins based on star graph topological indices». En: *Current topics in medicinal chemistry* 13.14 (2013), págs. 1681-1691 (vid. pág. 57).

[149] Carlos Fernandez-Lozano, Marcos Gestal, Humberto González-Díaz, Julián Dorado, Alejandro Pazos y Cristian R Munteanu. «Markov mean properties for cell death-related protein classification». En: *Journal of theoretical biology* 349 (2014), págs. 12-21 (vid. pág. 57).

[150] Colin Campbell y Yiming Ying. «Learning with support vector machines». En: *Synthesis lectures on artificial intelligence and machine learning* 5.1 (2011), págs. 1-95 (vid. pág. 57).

[151] John Shawe-Taylor, Nello Cristianini y col. *Kernel methods for pattern analysis*. Cambridge university press, 2004 (vid. pág. 57).

[152] Nello Cristianini, John Shawe-Taylor y col. *An introduction to support vector machines and other kernel-based learning methods*. Cambridge university press, 2000 (vid. págs. 57, 58).

[153] Christopher JC Burges. «A tutorial on support vector machines for pattern recognition». En: *Data mining and knowledge discovery* 2.2 (1998), págs. 121-167 (vid. pág. 58).

[154] Olivier Chapelle, Patrick Haffner y Vladimir N Vapnik. «Support vector machines for histogram-based image classification». En: *IEEE transactions on Neural Networks* 10.5 (1999), págs. 1055-1064 (vid. pág. 58).

[155] LS Moulin, AP Alves Da Silva, MA El-Sharkawi y Robert J Marks. «Support vector machines for transient stability analysis of large-scale power systems». En: *IEEE Transactions on Power Systems* 19.2 (2004), págs. 818-825 (vid. pág. 58).

[156] Bernhard Scholkopf, Sebastian Mika, Chris JC Burges, Philipp Knirsch, K-R Muller, Gunnar Ratsch y Alexander J Smola. «Input space versus feature space in kernel-based methods». En: *IEEE transactions on neural networks* 10.5 (1999), págs. 1000-1017 (vid. pág. 58).

[157] James Mercer. «Xvi. functions of positive and negative type, and their connection the theory of integral equations». En: *Philosophical transactions of the royal society of London. Series A, containing papers of a mathematical or physical character* 209.441-458 (1909), págs. 415-446 (vid. pág. 58).

[158] Alex J Smola y Bernhard Schölkopf. «A tutorial on support vector regression». En: *Statistics and computing* 14.3 (2004), págs. 199-222 (vid. pág. 58).

[159] Jingang Che, Lei Chen, Zi-Han Guo, Shuaiqun Wang y col. «Drug target group prediction with multiple drug networks». En: *Combinatorial Chemistry & High Throughput Screening* 23.4 (2020), págs. 274-284 (vid. pág. 59).

[160] Ben Hu, Zheng-Kun Kuang, Shi-Yu Feng, Dong Wang, Song-Bing He y De-Xin Kong. «Three-dimensional biologically relevant spectrum (BRS-3D): shape similarity profile based on PDB ligands as molecular descriptors». En: *Molecules* 21.11 (2016), pág. 1554 (vid. pág. 59).

[161] Leo Breiman. «Random forests». En: *Machine learning* 45.1 (2001), págs. 5-32 (vid. pág. 60).

[162] Jose Liñares-Blanco, Cristian R Munteanu, Alejandro Pazos y Carlos Fernandez-Lozano. «Molecular docking and machine learning analysis of Abemaciclib in colon cancer». En: *BMC Molecular and Cell Biology* 21.1 (2020), págs. 1-18 (vid. pág. 61).

[163] Lei Chen, Yu-Hang Zhang, Mingyue Zheng, Tao Huang y Yu-Dong Cai. «Identification of compound–protein interactions through the analysis of gene ontology, KEGG enrichment for proteins and molecular fragments of compounds». En: *Molecular Genetics and Genomics* 291.6 (2016), págs. 2065-2079 (vid. pág. 61).

[164] Meiyue Song y Zhenran Jiang. «Inferring association between compound and pathway with an improved ensemble learning method». En: *Molecular informatics* 34.11-12 (2015), págs. 753-760 (vid. pág. 61).

[165] Jun Hu, Yang Li, Jing-Yu Yang, Hong-Bin Shen y Dong-Jun Yu. «GPCR–drug interactions prediction using random forest with drug-association-matrix-based post-processing procedure». En: *Computational biology and chemistry* 60 (2016), págs. 59-71 (vid. pág. 61).

[166] Michael P Menden, Dennis Wang, Mike J Mason, Bence Szalai, Krishna C Bulusu, Yuanfang Guan, Thomas Yu, Jaewoo Kang, Minji Jeon, Russ Wolfinger y col. «Community assessment to advance computational prediction of cancer drug combinations in a pharmacogenomic screen». En: *Nature communications* 10.1 (2019), págs. 1-17 (vid. pág. 62).

[167] Alexios Koutsoukas, Keith J Monaghan, Xiaoli Li y Jun Huan. «Deep-learning: investigating deep neural networks hyper-parameters and comparison of performance to shallow methods for modeling bioactivity data». En: *Journal of cheminformatics* 9.1 (2017), pág. 42 (vid. pág. 65).

[168] Jeffrey Mendenhall y Jens Meiler. «Improving quantitative structure–activity relationship models using Artificial Neural Networks trained with dropout». En: *Journal of computer-aided molecular design* 30.2 (2016), págs. 177-189 (vid. pág. 65).

[169] Maria Batool, Bilal Ahmad y Sangdun Choi. «A structure-based drug discovery paradigm». En: *International journal of molecular sciences* 20.11 (2019), pág. 2783 (vid. pág. 65).

[170] Rebeca Diez-Alarcia, Víctor Yáñez-Pérez, Itziar Muneta-Arrate, Sonia Arrasate, Esther Lete, J Javier Meana y Humbert González-Díaz. «Big data challenges targeting proteins in GPCR signaling pathways; combining PTML-ChEMBL models and [35S] GTPγS binding assays». En: *ACS chemical neuroscience* 10.11 (2019), págs. 4476-4491 (vid. pág. 67).

[171] Ricardo Santana, Robin Zuluaga, Piedad Gañán, Sonia Arrasate, Enrique Onieva y Humbert González-Díaz. «Designing nanoparticle release systems for drug–vitamin cancer co-therapy with multiplicative perturbation-theory machine learning (PTML) models». En: *Nanoscale* 11.45 (2019), págs. 21811-21823 (vid. pág. 67).

[172] Celine B Santiago, Jing-Yao Guo y Matthew S Sigman. «Predictive and mechanistic multivariate linear regression models for reaction development». En: *Chemical science* 9.9 (2018), págs. 2398-2412 (vid. pág. 67).

[173] Kaid C Harper, Elizabeth N Bess y Matthew S Sigman. «Multidimensional steric parameters in the analysis of asymmetric catalytic reactions». En: *Nature chemistry* 4.5 (2012), págs. 366-374 (vid. pág. 67).

[174] Kaid C Harper y Matthew S Sigman. «Using physical organic parameters to correlate asymmetric catalyst performance». En: *The Journal of Organic Chemistry* 78.7 (2013), págs. 2813-2818 (vid. pág. 67).

[175] Elizabeth N Bess, Amanda J Bischoff y Matthew S Sigman. «Designer substrate library for quantitative, predictive modeling of reaction performance». En: *Proceedings of the National Academy of Sciences* 111.41 (2014), págs. 14698-14703 (vid. pág. 67).

[176] Huayin Huang, Hua Zong, Guangling Bian y Ling Song. «Constructing a quantitative correlation between N-substituent sizes of chiral ligands and enantioselectivities in asymmetric addition reactions of diethylzinc with benzaldehyde». En: *The Journal of Organic Chemistry* 77.22 (2012), págs. 10427-10434 (vid. pág. 67).

[177] Huayin Huang, Hua Zong, Bin Shen, Huifeng Yue, Guangling Bian y Ling Song. «QSAR analysis of the catalytic asymmetric ethylation of ketone using physical steric parameters of chiral ligand substituents». En: *Tetrahedron* 70.6 (2014), págs. 1289-1297 (vid. pág. 67).

[178] Kaid C Harper y Matthew S Sigman. «Predicting and optimizing asymmetric catalyst performance using the principles of experimental design and steric parameters». En: *Proceedings of the National Academy of Sciences* 108.6 (2011), págs. 2179-2183 (vid. pág. 67).

[179] Kaid C Harper y Matthew S Sigman. «Three-dimensional correlation of steric and electronic free energy relationships guides asymmetric propargylation». En: *Science* 333.6051 (2011), págs. 1875-1878 (vid. pág. 67).

[180] Kaid C Harper, Sarah C Vilardi y Matthew S Sigman. «Prediction of catalyst and substrate performance in the enantioselective propargylation of aliphatic ketones by a multidimensional model of steric effects». En: *Journal of the American Chemical Society* 135.7 (2013), págs. 2482-2485 (vid. pág. 67).

[181] Cristian R Munteanu, Julian Dorado, Alejandro Pazos-Sierra, F Prado-Prado, LG Pérez-Montoto, Santiago Vilar, Florencio M Ubeira, Angeles Sánchez-Gonzaléz, Maykel Cruz-Monteagudo, Sonia Arrasate y col. «Markov entropy centrality: Chemical, biological, crime, and legislative networks». En: *towards an information theory of complex networks: statistical methods and applications* (2011), págs. 199-258 (vid. pág. 67).

[182] Chun Zhang, Celine B Santiago, Jennifer M Crawford y Matthew S Sigman. «Enantioselective dehydrogenative Heck arylations of trisubstituted alkenes with indoles to construct quaternary stereocenters». En: *Journal of the American Chemical Society* 137.50 (2015), págs. 15668-15671 (vid. pág. 67).

[183] Chun Zhang, Celine B Santiago, Lei Kou y Matthew S Sigman. «Alkenyl carbonyl derivatives in enantioselective redox relay Heck reactions: accessing α, β-unsaturated systems». En: *Journal of the American Chemical Society* 137.23 (2015), págs. 7290-7293 (vid. pág. 67).

[184] Anat Milo, Andrew J Neel, F Dean Toste y Matthew S Sigman. «A data-intensive approach to mechanistic elucidation applied to chiral anion catalysis». En: *Science* 347.6223 (2015), págs. 737-743 (vid. pág. 67).

[185] Yoonsu Park, Zachary L Niemeyer, Jin-Quan Yu y Matthew S Sigman. «Quantifying structural effects of amino acid ligands in Pd (II)-catalyzed enantioselective C–H functionalization reactions». En: *Organometallics* 37.2 (2018), págs. 203-210 (vid. pág. 67).

[186] C Blázquez-Barbadillo, E Aranzamendi, E Coya, E Lete, N Sotomayor y H González-Díaz. «Perturbation theory model of reactivity and enantioselectivity of palladium-catalyzed Heck–Heck cascade reactions». En: *RSC advances* 6.45 (2016), págs. 38602-38610 (vid. pág. 67).

[187] Sonia Aguado-Ullate, Manuel Urbano-Cuadrado, Isabel Villalba, Elísabet Pires, José I García, Carles Bo y Jorge J Carbó. «Predicting the Enantioselectivity of the Copper-Catalysed Cyclopropanation of Alkenes by Using Quantitative Quadrant-Diagram Representations of the Catalysts». En: *Chemistry–A European Journal* 18.44 (2012), págs. 14026-14036 (vid. pág. 67).

[188] Huayin Huang, Hua Zong, Guangling Bian, Huifeng Yue y Ling Song. «Correlating the effects of the N-substituent sizes of chiral 1, 2-amino phosphinamide ligands on enantioselectivities in catalytic asymmetric Henry reaction using physical steric parameters». En: *The Journal of Organic Chemistry* 79.20 (2014), págs. 9455-9464 (vid. pág. 67).

[189] Ricardo Santana, Robin Zuluaga, Piedad Ganan, Sonia Arrasate, Enrique Onieva Caracuel y Humbert Gonzalez-Diaz. «PTML model of ChEMBL compounds assays for vitamin derivatives». En: *ACS Combinatorial Science* 22.3 (2020), págs. 129-141 (vid. pág. 68).

[190] Pablo Riera-Fernandez, Raquel Martin-Romalde, Francisco J Prado-Prado, Manuel Escobar, Cristian R Munteanu, Riccardo Concu, Aliuska Duardo-Sanchez y Humberto Gonzalez-Diaz. «From QSAR models of drugs to complex networks: state-of-art review and introduction of new Markov-spectral moments indices». En: *Current Topics in Medicinal Chemistry* 12.8 (2012), págs. 927-960 (vid. pág. 68).

[191] Richard D Hull, Eugene M Fluder, Suresh B Singh, Robert B Nachbar, Simon K Kearsley y Robert P Sheridan. «Chemical similarity searches using latent semantic structural indexing (LaSSI) and comparison to TOPOSIM». En: *Journal of medicinal chemistry* 44.8 (2001), págs. 1185-1191 (vid. pág. 68).

[192] Georgia Tsiliki, Cristian R. Munteanu, José A. Seoane, Carlos Fernandez-Lozano, Haralambos Sarimveis y Egon L. Willighagen. «RRegrs: an R package for computer-aided model selection with multiple regression models». En: *J. Cheminformatics* 7 (2015), 46:1-46:16. DOI: 10.1186/s13321-015-0094-2 (vid. pág. 74).

[193] Abdul Rahman y Xufeng Lin. «Development and application of chiral spirocyclic phosphoric acids in asymmetric catalysis». En: *Organic & Biomolecular Chemistry* 16.26 (2018), págs. 4753-4777 (vid. pág. 79).

[194] Joey Merad, Claudia Lalli, Guillaume Bernadat, Julien Maury y Géraldine Masson. «Enantioselective Brønsted Acid Catalysis as a Tool for the Synthesis of Natural Products and Pharmaceuticals». En: *Chem. - A Eur. J.* 24.16 (2018), págs. 3925-3943. DOI: 10.1002/chem.201703556 (vid. pág. 79).

[195] Yunpeng Xie, Ying Zhao, Bingqiang Qian, Lili Yang, Chunhui Xia y Hongbin Huang. «Enantioselective N-H Functionalization of Indoles with». En: *Angewandte Chemie International Edition* 50.25 (2011), págs. 5682-5686. DOI: 10.1002/anie.201102046 (vid. pág. 79).

[196] Xiaojie Yu, Anji Lu, Yan Wang, Gang Wu, Hai-bin Song, Zhipeng Zhou y Chunyan Tang. «Chiral Phosphoric Acid Catalyzed Asymmetric Friedel-Crafts Alkylation of Indole with 3-Hydroxyisoindolin-1-One: Enantioselective Synthesis of 3-Indolyl-Substituted Isoindolin-1-Ones». En: *European Journal of Organic Chemistry* 5 (2011), págs. 892-897. DOI: 10.1002/ejoc.201001408 (vid. pág. 79).

[197] Cheng Guo, Jingjing Song, Jiazhen Huang, Pihui Chen, Shao-Wen Luo y Liu-Zhu Gong. «Core-Structure-Oriented Asymmetric Organocatalytic Substitution of 3-Hydroxyoxindoles: Application in the Enantioselective Total Synthesis of (+)-Folicanthine». En: *Angewandte Chemie International Edition* 51.4 (2012), págs. 1046-1050. DOI: 10.1002/anie.201107079 (vid. pág. 79).

[198] Qiang Yin, Sheng-Guo Wang y Shu-Li You. «Asymmetric Synthesis of Tetrahydro-β-Carbolines via Chiral Phosphoric Acid Catalyzed Transfer Hydrogenation Reaction». En: *Organic Letters* 15.11 (2013), págs. 2688-2691. DOI: 10.1021/ol400995c (vid. pág. 79).

[199] Thomas Courant, Sakharam Kumarn, Lin He, Pascal Retailleau y Géraldine Masson. «Chiral Phosphoric Acid-Catalyzed Enantioselective Aza-Friedel-Crafts Alkylation of Indoles with γ-Hydroxy-γ-Lactams». En: *Advanced Synthesis & Catalysis* 355.5 (2013), págs. 836-840. DOI: 10.1002/adsc.201201008 (vid. pág. 79).

[200] Ahmet Yazici y Stephen G. Pyne. «Intermolecular Addition Reactions of Acyliminium Ions (Part I)». En: *Synthesis (Stuttg)* 3 (2009), págs. 339-368. DOI: 10.1055/s-0028-1083325 (vid. pág. 79).

[201] Ahmet Yazici y Stephen G. Pyne. «Intermolecular Addition Reactions of Acyliminium Ions (Part II)». En: *Synthesis* 4 (2009), págs. 513-541. DOI: 10.1055/s-0028-1083346 (vid. pág. 79).

[202] Unai Martínez-Estibalez, Amaia Gõmez-Sanjuan, Óscar García-Calvo, Eunate Aranzamendi, Esther Lete y Nuria Sotomayor. «Strategies Based on Aryllithium and N-Acyliminium Ion Cyclizations for the Stereocontrolled Synthesis of Alkaloids and Related Systems». En: *European J. Org. Chem.* 20-21 (2011), págs. 3610-3633. DOI: 10.1002/ejoc.201100123 (vid. pág. 79).

[203] Thomas E. Nielsen y Morten Meldal. «Solid-Phase Synthesis of Complex and Pharmacologically Interesting Heterocycles». En: *Curr. Opin. Drug Discov. Dev.* 12.6 (2009), págs. 798-810. DOI: 10.1002/chin.201019251 (vid. pág. 79).

[204] Clara Avendaño López y Esteban de la Cuesta. «Synthetic Chemistry with N-Acyliminium Ions Derived from Piperazine-2,5-Diones and Related Compounds». En: *Curr. Org. Synth.* 6 (2009), págs. 143-168. DOI: 10.2174/157017909788167310 (vid. pág. 79).

[205] I. Osante, M. I. Collado, E. Lete y N. Sotomayor. «ChemInform Abstract: Stereodivergent Synthesis of Hetero-Fused Isoquinolines by Acyliminium and Metalation Methods». En: *ChemInform* 33.13 (2010), no-no. DOI: 10.1002/chin.200213151 (vid. pág. 79).

[206] Iñaki González-Temprano, Iker Osante, Esther Lete y Nuria Sotomayor. «Enantiodivergent Synthesis of Pyrrolo[2,1-α]Isoquinolines Based on Diastereoselective Parham Cyclization and α-Amidoalkylation Reactions». En: *J. Org. Chem.* 69.11 (2004), págs. 3875-3885. DOI: 10.1021/jo049672o (vid. pág. 79).

[207] Muhammad Naveed Abdullah, Sonia Arrasate, Esther Lete y Nuria Sotomayor. «Stereoselective Synthesis of Thiaerythrinanes Based on an -Amidoalkylation/RCM Approach». En: *Tetrahedron* 64.7 (2008), págs. 1323-1332. DOI: 10.1016/j.tet.2007.12.012 (vid. pág. 79).

[208] Yong Seok Lee, Md Maksudul Alam y Rangappa S. Keri. «Enantioselective Reactions of N-Acyliminium Ions Using Chiral Organocatalysts». En: *Chem. - An Asian J.* 8.12 (2013), págs. 2906-2919. DOI: 10.1002/asia.201300814 (vid. pág. 79).

[209] Teiji Akiyama, Benjamin List y Keiji Maruoka, eds. *Science of Synthesis: Asymmetric Organocatalysis Vol. 2: Bronsted Base and Acid Catalysts, and Additional Topics*. Georg Thieme Verlag, 2012, págs. 169-217 (vid. pág. 79).

[210] Masahiko Terada y Keiji Maruoka, eds. *Asymmetric Organocatalysis 2: Brønsted Base and Acid Catalysts, and Additional Topics*. Vol. 2. Georg Thieme Verlag, 2012, págs. 219-278 (vid. pág. 79).

[211] Renato Dalpozzo. «Strategies for the Asymmetric Functionalization of Indoles: An Update». En: *Chem. Soc. Rev.* 44.3 (2015), págs. 742-778. DOI: 10.1039/c4cs00209a (vid. pág. 79).

[212] Teiji Akiyama. «Stronger Brønsted Acids». En: *Chem. Rev.* 107.12 (2007), págs. 5744-5758. DOI: 10.1021/cr068373r (vid. pág. 79).

[213] Teiji Akiyama y Keiji Mori. «Stronger Brønsted Acids: Recent Progress». En: *Chem. Rev.* 115.17 (2015), págs. 9277-9306. DOI: 10.1021/acs.chemrev.5b00041 (vid. pág. 79).

[214] Drupad Parmar, Emmerich Sugiono, Seenivasan Raja y Magnus Rueping. «Complete Field Guide to Asymmetric BINOL-Phosphate Derived Brønsted Acid and Metal Catalysis: History and Classification by Mode of Activation; Brønsted Acidity, Hydrogen Bonding, Ion Pairing, and Metal Phosphates». En: *Chem. Rev.* 114.18 (2014), págs. 9047-9153. DOI: 10.1021/cr5001496 (vid. pág. 79).

[215] Yoshiji Takemoto. «Recognition and Activation by Ureas and Thioureas: Stereoselective Reactions Using Ureas and Thioureas as Hydrogen-Bonding Donors». En: *Org. Biomol. Chem.* 3.24 (2005), págs. 4299-4306. DOI: 10.1039/b511216h (vid. pág. 79).

[216] Abigail G. Doyle y Eric N. Jacobsen. «Small-Molecule H-Bond Donors in Asymmetric Catalysis». En: *Chem. Rev.* 107.12 (2007), págs. 5713-5743. DOI: 10.1021/cr068373r (vid. pág. 79).

[217] Robert R. Knowles y Eric N. Jacobsen. «Attractive Noncovalent Interactions in Asymmetric Catalysis: Links between Enzymes and Small Molecule Catalysts». En: *Proc. Natl. Acad. Sci.* 107.48 (2010), págs. 20678-20685. DOI: 10.1073/pnas.1006402107 (vid. pág. 79).

[218] Gergely Jakab y Peter R. Schreiner. «Comprehensive Enantioselective Organocatalysis». En: *Comprehensive Enantioselective Organocatalysis*. Ed. por Renato Dalpozzo. Vol. 2. Wiley-VCH, 2013, págs. 315-341 (vid. pág. 79).

[219] Vincent Terrasson, Ricardo M. De Figueiredo y Jean-Emmanuel Campagne. «Organocatalyzed Asymmetric Friedel-Crafts Reactions». En: *Eur. J. Org. Chem.* 14 (2010), págs. 2635-2655. DOI: 10.1002/ejoc.200901492 (vid. pág. 79).

[220] Mei Zeng y Shu-Li You. «Asymmetric Friedel-Crafts Alkylation of Indoles: The Control of Enantio- and Regioselectivity». En: *Synlett* 9 (2010), págs. 1289-1301. DOI: 10.1055/s-0029-1219929 (vid. pág. 79).

[221] Ricardo M. de Figueiredo y Jean-Emmanuel Campagne. «Comprehensive Enantioselective Organocatalysis». En: *Comprehensive Enantioselective Organocatalysis*. Ed. por Peter I. Dalko. Vol. 3. Wiley-VCH, 2013, págs. 1043-1066 (vid. pág. 79).

[222] I. P. Beletskaya y D. Averin. «Asymmetric Friedel-Crafts Reactions of Indole and Its Derivatives». En: *Curr. Organocatalysis* 3.1 (2015), págs. 60-83. DOI: 10.2174/2213337202666150505230013 (vid. pág. 79).

[223] R. Mazurkiewicz, A. Październiorniok-Holewa, J. Adamek y K. Zielińska. *-Amidoalkylating Agents: Structure, Synthesis, Reactivity and Application*. Vol. 111. Elsevier, 2014. DOI: 10.1016/B978-0-12-420160-6.00002-1 (vid. pág. 79).

[224] Y. Marrero-Ponce, D. Siverio-Mota, M. Gálvez-Llompart, M. C. Recio, R. M. Giner, R. García-Domnech, F. Torrens, V. J. Arán, M. L. Cordero-Maldonado, C. V. Esguera y col. «Discovery of Novel Anti-Inflammatory Drug-like Compounds by Aligning in Silico and in Vivo Screening: The Nitroindazolinone Chemotype». En: *Eur. J. Med. Chem.* 46.12 (2011), págs. 5736-5753. DOI: 10.1016/j.ejmech.2011.07.053 (vid. pág. 79).

[225] Lorena Simón-Vidal, Oihane García-Calvo, Uxue Oteo, Sonia Arrasate, Esther Lete, Nuria Sotomayor y Humberto Gonzalez-Diaz. «Perturbation-theory and machine learning (PTML) model for high-throughput screening of Parham reactions: experimental and theoretical studies». En: *Journal of Chemical Information and Modeling* 58.7 (2018), págs. 1384-1396 (vid. págs. 82, 93, 96).

[226] H. Liu, J. Deng, Z. Luo, Y. Lin, KM. Merz Jr y Z. Zheng. «Receptor–Ligand Binding Free Energies from a Consecutive Histograms Monte Carlo Sampling Method». En: *Journal of Chemical Theory and Computation* 16.11 (2020), págs. 6645-6655 (vid. pág. 88).

[227] I. Cabeza de Vaca, Y. Qian, JZ. Vilseck, J. Tirado-Rives y WL. Jorgensen. «Enhanced Monte Carlo methods for modeling proteins including computation of absolute free energies of binding». En: *Journal of Chemical Theory and Computation* 14.6 (2018), págs. 3279-3288 (vid. pág. 88).

[228] Daniel J Cole, Julian Tirado-Rives y William L Jorgensen. «Enhanced Monte Carlo sampling through replica exchange with solute tempering». En: *Journal of Chemical Theory and Computation* 10.2 (2014), págs. 565-571 (vid. pág. 88).

[229] Dávid Bajusz, Anita Rácz y Károly Héberger. «Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations?» En: *J Cheminformatics* 7 (2015), págs. 1-13. DOI: 10.1186/s13321-015-0069-3 (vid. pág. 93).

[230] Ceslav Škuta, Isidro Cortés-Ciriano, Wim Dehaen, Petr Kříž, Gerard J van Westen, Igor V Tetko, Andreas Bender y Daniel Svozil. «QSAR-derived affinity fingerprints (part 1): fingerprint construction and modeling performance for similarity searching, bioactivity classification and scaffold hopping». En: *J Cheminformatics* 12 (2020), págs. 1-16. DOI: 10.1186/s13321-020-0422-4 (vid. pág. 93).

[231] Isidro Cortes-Ciriano, Nicholas C Firth, Andreas Bender y Oliver Watson. «Discovering highly potent molecules from an initial set of inactives using iterative screening». En: *Journal of Chemical Information and Modeling* 58.9 (2018), págs. 2000-2014 (vid. pág. 93).

[232] Andreas Bender, Jeremy L Jenkins, Jörg Scheiber, Sai Chetan K Sukuru, Meir Glick y John W Davies. «How similar are similarity searching methods? A principal component analysis of molecular descriptor space». En: *J Chem Inf Model* 49 (2009), págs. 108-119. DOI: 10.1021/ci800249w (vid. pág. 93).

[233] Anna B Wagner. «SciFinder Scholar 2006: an empirical analysis of research topic query processing». En: *J Chem Inf Model* 46 (2006), págs. 767-774. DOI: 10.1021/ci050400b (vid. pág. 93).

[234] David D Ridley. «Strategies for chemical reaction searching in SciFinder». En: *J Chem Inf Comp Sci* 40 (2000), págs. 1077-1084. DOI: 10.1021/ci0002570 (vid. pág. 93).

[235] Gianmarc Grazioli, Saswata Roy y Carter T Butts. «Predicting Reaction Products and Automating Reactive Trajectory Characterization in Molecular Simulations with Support Vector Machines». En: *Journal of chemical information and modeling* 59.6 (2019), págs. 2753-2764 (vid. pág. 97).

[236] Antoine Charpentier, David Mignon, Sophie Barbe, Juan Cortes, Thomas Schiex, Thomas Simonson y David Allouche. «Variable neighborhood search with cost function networks to solve large computational protein design problems». En: *Journal of Chemical Information and Modeling* 59.1 (2018), págs. 127-136 (vid. pág. 97).

[237] Tigran M Abramyan, Yi An y Dmitri Kireev. «Off-pocket activity cliffs: a puzzling facet of molecular recognition». En: *Journal of Chemical Information and Modeling* 60.1 (2019), págs. 152-161 (vid. pág. 97).

[238] *Dockers*. https://docs.docker.com/get-started/overview/. Último acceso 01/05/2023 (vid. pág. 100).

[239] *openBabel*. https://openbabel.org/wiki/Main_Page. Último acceso 01/05/2023 (vid. pág. 101).

[240] *Nginx*. https://www.nginx.com/resources/glossary/nginx/. Último acceso 01/05/2023 (vid. pág. 101).

Producción

A

Se recopila en el anexo la producción científica y técnica generada durante el periodo de elaboración de la Tesis Doctoral. Además de dos publicaciones en revistas JCR y una tercera en revisión actualmente (disponible en abierto, con seguimiento del estado en *researchsquare*), una publicación en congreso internacional, se indican los repositorios de código en GitHub con las herramientas de las publicaciones y la dirección web en la que se encuentran desplegadas las mismas.

Artículos JCR (WOS)

- **Carracedo-Reboredo, P.**, Liñares-Blanco, J., Rodríguez-Fernández, N., Cedrón, F., Novoa, F.J., Carballal, A., Maojo, V., Pazos, A., and Fernandez-Lozano, C. A review on machine learning approaches and trends in drug discovery. Computational and Structural Biotechnology Journal. [Q1, 70/297 BIO-MB, 6.155 IF]. `http://dx.doi.org/10.1016/j.csbj.2021.08.011`

- **Carracedo-Reboredo, P.**, Corona, R., Martinez-Nunes, M., Fernandez-Lozano, C., Tsiliki, G., Sarimveis, H., Aranzamendi, E., Arrasate, S., Sotomayor, N., Lete, E., Munteanu, C.R., and González-Díaz, H. MCDCalc: Markov Chain Molecular Descriptors Calculator for Medicinal Chemistry. Current Topics in Medicinal Chemistry [Q3, 35/63 CH-ME, 3.295 IF]. `http://dx.doi.org/10.2174/156802662066191226092431`

- **Carracedo-Reboredo, P.**, Aranzamendi, E., He, S., Arrasate, S., Munteanu, C.R., Fernandez-Lozano, C., Sotomayor, N., González-Díaz, H. MATEO: InterMolecular α-Amidoalkylation Theoretical Enantioselectivity Optimization. Online Tool for Selection and Design of Chiral Catalysts and Products. Journal of Cheminformatics. **Under review**. El artículo está disponible online como preprint en la siguiente url: `https://www.researchsquare.com/article/rs-2642502/v1`. Situación actual: **Major revision desde 25 de marzo 2023**.

Artículos en congresos internacionales

- **Carracedo-Reboredo, P.**, Munteanu, C.R., González-Díaz, H., and Fernandez-Lozano, C. Web server and R library for Calculation of markov chains molecular descriptors. In Proc. of the 3rd XoveTIC Conference, pp. 1-3, XoveTIC, 54, MDPI - proceedings. http://dx.doi.org/10.3390/proceedings2020054028

Repositorios de código en GitHub

- MATEO: InterMolecular Amidoalkylation Theoretical Enantioselectivity Optimization Web Server. https://github.com/glezdiazh/MATEO

- MCDCalc: Calculation of Markov Singular Values Molecular Descriptors Online Tool. https://github.com/glezdiazh/MCDCALC

Herramientas desplegadas

- Ambas herramientas se encuentran actualmente desplegadas a disposición de la comunidad investigadora en la siguiente url: https://cptmltool.rnasa-imedir.com/CPTMLTools-Web/

Abreviaturas

B

- ADN - Ácido Desoxirribonucleico
- ADMET - Administración, Distribución, Metabolismo, Eliminación y Toxicidad
- AINE - Antiinflamatorios No Esteroideos
- APC - Adenomatous Polyposis Coli
- ATC - Anatomical Therapeutic Chemical
- BDP - Banco de Datos de Proteinas
- CCR - Cáncer Colorectal
- CCRm - Cáncer Colorrectal Metastásico
- CCNPH - Cáncer Colorrectal Hereditario No Polipósico
- CEA - Antígeno Carcinoembrionario
- CIN - Cromosomal Instability Pathway
- COX - Ciclooxigenasa
- CPA - Catalizadores de Ácido Fosfórico Quiral
- DCC - Delete in Colon Cancer
- DM - Descriptores Moleculares
- DM2 - Diabetes Mellitus tipo 2

- ECF - Extended Connectivity Fingerprint
- $ee_R(\%)_{obs}$ Exceso Enantiomérico observado (experimental) usando (R)-Catalizador
- $ee_R(\%)_{ref}$ Exceso Enantiomérico de referencia (experimental) usando (R)-Catalizador
- $ee_R(\%)_{calc}$ Exceso Enantiomérico calculado usando (R)-Catalizador de referencia
- $ee_R(\%)_{pred}$ Exceso Enantiomérico predicho por el modelo usando (R)-Catalizador
- $ee_R(\%)_{res}$ Exceso Enantiomérico residual usando (R)-Catalizador
- EII - Enfermedad Inflamatoria Intestinal
- EpCAM - Epithelial Cellular Adhesion Molecule
- GLOBOCAN - Observatorio Global del Cáncer
- HPTML Heuristic Perturbation-Theory and Machine Learning
- IA - Inteligencia Artificial
- IGF - Insulin-like Growth Factor
- IFP - Interacciones Fármaco-Proteína
- LOX - Lipooxigenasa
- MACCS - Molecular ACCess System
- MC - Monte Carlo
- MCDs - Markov Chain Molecular Descriptors
- ML - Machine Learning

- MMR - Mismatch Repair
- NB - Naïve Bayes
- PAF - Poliposis Adenomatosa Familiar
- PARP - Poli ADP Ribosa Polimerasa
- PD-1 - Programmed cell Death
- PT Perturbation Theory
- PTO Perturbation Theory Operator
- RAM - Reacciones Adversas a Medicamentos
- REDECAN - Red Española de Registros de Cáncer
- SEOM - Sociedad Española de Oncología Médica
- SE Standard Error
- SEE Standard Error Estimates
- SMILE Simplified Molecular Input Line Entry Specification
- SVM - Support Vector Machines
- THF Tetrahydrofuran
- THS - Terapia Hormonal Sustitutiva
- VHB - Virus de la Hepatitis B
- VIH - Virus de la Inmunideficiencia Humana

C

Listado de figuras

D

Listado de tablas