



FACULTAD DE INGENIERÍA
ESCUELA DE INGENIERÍA CIVIL EN BIOINFORMÁTICA

Minería de datos para la caracterización de mecanismos de acción de fungicidas utilizados en fruticultura

CHRISTIAN IGNACIO OLIVERA FUENTES

Profesor Tutor: JOSÉ ANTONIO REYES SUÁREZ.

Memoria para optar el título de Ingeniero Civil en Bioinformática.

Talca – Chile



FACULTAD DE INGENIERÍA
ESCUELA DE INGENIERÍA CIVIL EN BIOINFORMÁTICA

Minería de datos para la caracterización de mecanismos
de acción de fungicidas utilizados en fruticultura

CHRISTIAN IGNACIO OLIVERA FUENTES

Profesor Tutor: DR. JOSÉ ANTONIO REYES SUÁREZ.

Profesor Informante: DR. MAURICIO ANTONIO ARENAS SALINAS

Memoria para optar el título de Ingeniero Civil en Bioinformática.

Talca – Chile

TABLA DE CONTENIDOS

RESUMEN	9
1 INTRODUCCIÓN.....	10
1.1 Fruticultura en Chile.....	10
1.2 Fungicidas y control de hongos.....	11
1.3 Clasificación de fungicidas.	12
1.4 Mecanismos de Acción.	18
1.5 Ciencia de datos en biología.....	19
2 HIPÓTESIS	24
3 OBJETIVOS.....	24
3.1 Objetivo general	24
3.2 Objetivos específicos.....	24
4 MATERIALES Y MÉTODOS.....	25
4.1 Materiales	25
4.1.1 Identificación y obtención de estructuras proteicas y de fungicidas.....	25
4.1.2 Caracterización a nivel molecular de fungicidas	27
4.1.3 Preparación y caracterización de complejos proteína-ligando.....	27
4.1.4 Implementación de selección de atributos y técnicas de minería de datos.....	28
4.2 Metodología	29
4.2.1 Identificación y obtención de estructuras moleculares.	29
4.2.2 Caracterización de moléculas y complejos proteína-ligando.....	30
4.2.3 Generación de los sets de datos	34
4.2.4 Minería de datos.....	34
5 RESULTADOS	39
5.1 Resultados de estudios con complejos proteína-ligando.....	39
5.2 Resultado de predicción de los niveles en efectos colaterales.	46
6 DISCUSIÓN.....	51
6.1 Recolección de datos.....	51
6.2 Selección de atributos.....	52
6.3 Desempeño de los modelos	52

7	CONCLUSIÓN	54
8	ANEXOS	55
8.1	Anexos tarea de interacción a proteína	55
8.2	Anexos tarea de nivel de efecto colateral.....	67
	REFERENCIAS.....	83

ÍNDICE DE TABLAS

Tabla 1.1: Nombre De Los Tipos De Grupos Químicos Y Algunos Ejemplos De Fungicidas Según Su Tipo. Tabla Correspondiente A La Clasificación De Gupta, 2017 Y La Inclusión De Los Fungicidas Inorgánicos Y Metálicos De Gupta, 2018.	14
Tabla 1.2: Clasificación De Los Métodos De Acción Y Su Código. Tomado De Las Clasificaciones Realizadas Por FRAC.	19
Tabla 4.1: Representación De Una Matriz De Confusión. Las Columnas Corresponden A La Clasificación De Los Ejemplos Definida Por La Predicción Del Modelo Y Las Filas Corresponden A Su Clase Correcta. La Cantidad De Ejemplos Que Fueron Predichos Correctamente Se Denominan Verdaderos Positivos (VP) Y Verdaderos Negativos (VN). Los Ejemplos Que Fueron Clasificados Erróneamente Pasan A Ser Falsos Positivos (FP) O Falsos Negativos (FN).	38
Tabla 5.1: Proteínas Identificadas Y Los Fungicidas Con Los Que Interactúan, Según FRAC.	40
Tabla 5.2: Descripción De Los Sets De Datos Para El Análisis En Interacción A Proteína Con Atributos Calculados Por PLPG3D.	42
Tabla 5.3: Resultados Del Mejor Desempeño Según Número De Atributos.	43
Tabla 5.4: Número De Ejemplos Por Clase Posterior Al Proceso De Balance De Clases Con El Método SMOTE.....	43
Tabla 5.5: Resultados Del Mejor Desempeño Según Número De Atributos.	44
Tabla 5.6: Atributos Más Relevantes Para Las Tareas Relacionadas Con La Interacción A Proteína.....	45
Tabla 5.7: Número De Fungicidas Con Niveles Identificados En Cada Efecto Colateral Estudiado.	46
Tabla 5.8: Efectos Colaterales Del Uso De Fungicidas Y Sus Niveles Identificados En El Comité De Acción Al Riesgo A Resistencia (FRAC) Y La Base De Datos De Propiedades De Pesticidas (PPDB).	47
Tabla 5.9: Número De Fungicidas Y Atributos Posterior Al Proceso De Limpieza.	47
Tabla 5.10: Resultado Del Desempeño De Random Forest Para Diferenciar Entre Los Grados De Los Efectos Colaterales Por El Uso De Fungicidas.....	48
Tabla 5.11: Resultado Del Desempeño De Random Forest Con Set De Datos Balanceado Para Diferenciar Entre Los Grados De Los Efectos Colaterales Por El Uso De Fungicidas.....	49
Tabla 5.12: Atributos Más Relevantes Para Las Tareas Relacionadas Con El Nivel De Efecto Colateral.....	50
Tabla 8.1: Se Dan A Conocer Los Atributos Más Relevantes Para La Tarea De Interacción Con Rna Polimerasa I.	55
Tabla 8.2: Se Dan A Conocer Los Atributos Más Relevantes Para La Tarea De Interacción Con Beta Tubulina.....	58

Tabla 8.3: Se Dan A Conocer Los Atributos Más Relevantes Para La Tarea De Interacción Con Succinato Deshidrogenasa.....	61
Tabla 8.4: Se Dan A Conocer Los Atributos Más Relevantes Para La Tarea De Interacción Con Citocromo BC1.....	64
Tabla 8.5: Se Dan A Conocer Los Atributos Más Relevantes Para La Tarea De Diferenciar El Nivel De Riesgo A Resistencia De Los Fungicidas.....	67
Tabla 8.6: Se Dan A Conocer Los Atributos Más Relevantes Para La Tarea De Diferenciar El Nivel Ecotóxico De Los Fungicidas.....	71
Tabla 8.7: Se Dan A Conocer Los Atributos Más Relevantes Para La Tarea De Diferenciar El Grado De Daño Medioambiental De Los Fungicidas.....	75
Tabla 8.8: Se Dan A Conocer Los Atributos Más Relevantes Para La Tarea De Diferenciar El Grado De Daño En Salud Humana De Los Fungicidas.....	79

ÍNDICE DE FIGURAS

Figura 1.1: Ejemplos De Estructuras Químicas De Fungicida, Con Sus Respectivos Nombres.....	17
Figura 4.1: Diagrama De La Metodología General Aplicada En Este Estudio	29
Figura 4.2: Representación Gráfica De Los Parámetros Para El Cálculo Mediante PLPG3D..	32
Figura 4.3: Representación Gráfica Del Modelo De Aprendizaje Supervisado.	36
Figura 8.1: Boxplots De Los Atributos Más Relevantes Para La Tarea De Interacción Con RNA Polimerasa I.....	56
Figura 8.2: Boxplots De Los Atributos Más Relevantes Para La Tarea De Interacción Con Beta Tubulina.	59
Figura 8.3: Boxplots De Los Atributos Más Relevantes Para La Tarea De Interacción Con Succinato Deshidrogenasa.	62
Figura 8.4: Boxplots De Los Atributos Más Relevantes Para La Tarea De Interacción Con Citocromo Bc1.....	65
Figura 8.5: Boxplots De Los Atributos Más Relevantes Para La Tarea De Diferenciar Los Niveles De Riesgo A Resistencia De Fungicidas.	68
Figura 8.6: Boxplots De Los Atributos Más Relevantes Para La Tarea De Diferenciar Los Niveles De Ecotoxicidad De Los Fungicidas.	72
Figura 8.7: Boxplots De Los Atributos Más Relevantes Para La Tarea De Diferenciar El Grado De Daño Medioambiental De Los Fungicidas.	76
Figura 8.8: Boxplots De Los Atributos Más Relevantes Para La Tarea De Diferenciar El Grado De Daño En Salud Humana De Los Fungicidas.....	80

ÍNDICE DE FÓRMULAS

Fórmula 4.1: Normalización Con El Método Min-Max.....	35
Fórmula 4.2: Cálculo De Exactitud Del Modelo.....	38
Fórmula 4.3: Cálculo De Sensibilidad.....	38

RESUMEN

La fruticultura es una de las actividades más importante en Chile, formando parte de las 10 áreas más exportadas del país. No obstante, la infección provocada por hongos es una de las problemáticas que enfrenta esta área, produciendo importantes pérdidas. Con el fin de evitar estas enfermedades, los productores hacen uso de fungicidas, los cuales presentan problemas propios de su uso como resistencia a fungicidas por parte de organismo fungi y efectos dañinos al entorno donde se aplican, haciendo necesario el desarrollo de nuevos fungicidas que a la vez no sean nocivos.

En este trabajo se utilizaron técnicas de minería de datos para identificar características y atributos relevantes que permitan diferenciar distintos tipos de mecanismos de acción de fungicidas aplicados generalmente en la fruticultura. Para esto se generó información asociada a propiedades estructurales, geométricas y fisicoquímicas de fungicidas y de sitios de unión de estos con proteínas de patógenos fungi conocidos.

Se encontró información para 223 moléculas fungicidas y 20 proteínas dianas que son listadas en el Comité de Acción a Resistencia a Fungicida (FRAC). Las estructuras tridimensionales de las proteínas dianas fueron descargadas de la base de datos Protein Data Bank (PDB). Se realizó un estudio de acoplamiento molecular para identificar la mejor conformación molecular de fungicidas y fijarla en la estructura proteica. Un total de 186 atributos fueron calculados en base a las propiedades estructurales, geométricas y fisicoquímicas de las zonas de unión en complejos proteína-ligando. Adicionalmente, las moléculas fueron caracterizadas en base a sus propiedades estructurales utilizando el software Mordred, estimando un total de 1826 atributos para cada una de estas.

Del análisis realizado se determinó que atributos y características de las zonas de unión en complejos proteína-ligando permiten discriminar eficientemente entre los distintos tipos de interacciones entre fungicidas y proteínas estudiados. Encontrando modelos predictivos que alcanzan una exactitud superior al 80% en distintos escenarios evaluados. Adicionalmente, se generaron modelos para inferir sobre efectos colaterales del uso de fungicidas. Por ejemplo, para niveles de riesgo a resistencia, ecotoxicidad y el grado de daño al medio ambiente y salud humana.

1 INTRODUCCIÓN

1.1 Fruticultura en Chile

La fruticultura es el área de la agricultura que se dedica al cultivo de plantas y árboles que producen fruto. En Chile, este sector tiene un reconocimiento a nivel mundial, ya que es el principal país productor y exportador de frutas en el hemisferio sur. En el año 2019 se contaba para esta práctica con 342.654 hectáreas y un total de 571.894 trabajadores. Asimismo, existían cerca de 17.000 huertos frutales, de acuerdo con registros del año 2018, y 246 plantas procesadoras de productos hortofrutícolas, según un catastro del 2011 (Pefaur, 2020).

La producción frutícola en Chile se divide en tres áreas: (1) producción para frescos, (2) producción para procesamientos, considerando productos (i) congelados, (ii) conservas, (iii) deshidratados, (iv) jugos y (v) aceites, y (3) producción de frutos secos. Cada uno de estos tipos de productos se destinan principalmente para exportación. El país posee 29 acuerdos económicos que le permiten entrar a 65 países o bloques comerciales con cero o bajos aranceles. De esta forma, en el año 2019 se percibió un total de US \$5.540,45.- millones en exportaciones de vid de mesa, nogal, cerezo, palto, manzano rojo, avellano, olivo y arándano. Según un reporte de la ODEPA del mismo año, la actividad frutícola contribuyó un 39,2% al PIB sectorial y 34% a las exportaciones silvoagropecuarias (Apey, 2019).

A pesar de su destacada participación laboral y económica en la actividad frutícola, esta área no está exenta de desafíos. Uno de estos es extender la vida poscosecha para la exportación, la cual se ve limitada principalmente por la pudrición causada por hongos. La presencia de estos microorganismos ha afectado cultivos tales como cerezos (Auger et. al., 2021), arándanos (Rojo et. al., 2017), vides (Diaz et. al., 2013; Silva-Valderrama et. al., 2021), manzana, entre otros, provocando la pudrición del fruto (Diaz et. al., 2019). Estas infecciones se producen tanto en el fruto como en la madera del árbol, reduciendo los niveles de ofertas frutícolas y pérdidas en el área. Tales problemas se repiten en otros países sudamericanos como Uruguay (Delgado-Cerrone et. al., 2016; Sessa et. al., 2016) y en otros países alrededor del mundo (Bertetti et. al., 2020; Wood, 2017).

1.2 Fungicidas y control de hongos.

Los fungicidas son agentes utilizados para la prevención o erradicación de infecciones por hongos en plantas, árboles, frutos y semillas, constituidos por una amplia variedad de sustancias químicas. En sus inicios, los fungicidas estaban formados por materiales inorgánicos tales como cal, metales (p. ej. mercurio) y altos niveles de azufre, afectando a trabajadores agrícolas con dermatitis, discapacidad neurológica permanente y la muerte. Impactando, además, en la salud de los animales. Por ende, al transcurrir el tiempo se ha ido suprimiendo el uso de fungicidas de alta toxicidad. Actualmente, la gran mayoría de fungicidas utilizados son de origen sintético, los cuales en concentraciones normales no afectan a especies mamíferas, pero pueden producir efectos negativos a una alta exposición o ingesta de estos químicos. De esta forma, un fungicida efectivo debe poseer las siguientes características: (1) bajo nivel de toxicidad en plantas y animales, pero altamente tóxico para hongos, (2) la capacidad de transformarse, ya sea mediante enzimas vegetales o fúngicas, en un metabolito secundario tóxico, (3) la capacidad de penetrar esporas o micelios de hongos para llegar al sitio de acción, (4) bajos niveles de ecotoxicidad, y (5) que posea la capacidad de resistir en intemperie la luz solar, la lluvia y el viento (Gupta, 2018).

El desarrollo de nuevos fungicidas o técnicas de control de hongos se mantiene en el tiempo actual, enfocados principalmente en el desarrollo de bioagentes, microorganismos con propiedades antifúngicas, y uso de sustancias naturales; alternativa ante la resistencia de hongos a fungicidas sintéticos y la toxicidad que estos presentan en la salud humana y en el medio ambiente. La resistencia a fungicidas significa que no se logra eliminar al hongo con uno o más compuestos utilizados. En Chile, se evaluó la resistencia a fungicidas en hongos que afectan a la uva sin pepa (*Vitis vinifera* cv. ‘Thompson Seedless’), encontrando que la mayoría de estos agentes patógenos presentaba resistencia a uno o más fungicidas (Esterio et. al., 2017). Un estudio similar se aplicó a viñedos en Italia, donde se obtuvieron resultados variados. Por ejemplo, se mostró que la mitad de los viñedos presentaron resistencia a fungicidas benzimidazoles, pero si presentaba resultados positivos en fenilpirroles. No obstante, se cree que los resultados podrían haberse producido por la concentración aplicada de cada compuesto y podrían presentar resistencia posteriormente (Bertetti et. al. 2020).

Se han realizado diversos estudios para identificar patógenos fúngicos y estrategias para el control de estos. Evaluando la pudrición del acebo (*Ilex aquifolium*), cuya causa no era conocida, se encontraron una gran variedad de agentes patógenos que enferman al fruto y provocan pérdidas económicas en viveros del medio oeste y este de EE. UU. (Lin et. al., 2018), haciendo un llamado a estudiar nuevas estrategias para formular un tratamiento a estas enfermedades. En tomate (*Solanum lycopersicum*), se encontraron variaciones en su expresión génica entre la etapa premadura y madura, diferencia que aumenta la producción de la enzima pectato liasa e incrementa su susceptibilidad cuando está maduro; se sugiere apuntar a estos genes específicos para mayor resistencia del fruto ante enfermedades por hongos (Silva et. al., 2021). Otros estudios realizados en manzanas y en fruta cítrica, también promueven el control y modificación genética para disminuir la susceptibilidad del fruto (Souleyre et. al., 2019; Cheng et. al., 2020). Entre otras técnicas que se han evaluado está el uso de mallas fotoselectivas (Candian et. al. 2020), ozonización (Contigiani et. al., 2018), control por bioagentes basados en bacterias (Valdés-Gómez et. al., 2017; Kwon et. al., 2021) y otras especies fúngicas (De Curtis et. al, 2019; Silva-Valderrama et. al., 2021), uso de sustancias naturales (Ji et. al., 2020) y la mezcla de fungicidas que antes se usaban individualmente (Mendoza et. al., 2019). Sin embargo, conforme a estos artículos, los fungicidas sintéticos siguen siendo los más utilizados en fruticultura.

1.3 Clasificación de fungicidas.

A lo largo del tiempo, se han descubierto e implementado nuevos y distintos tipos de fungicidas. Estos, según Gupta, pueden ser clasificados conforme al grupo químico presente en su estructura (Gupta, 2018). Esta clasificación se puede observar en la Tabla 1.1, considerando los grupos químicos y algunos ejemplos de fungicidas relacionados.

Clase química	Ejemplos
Fungicidas Inorgánicos	Azida de potasio, tiocianato de potasio, azufre sublimado
Fungicidas metálicos	Fosfato de etilmercurio, cloruro de 2-metoxietil mercurio, cloruro de fenilmercurio
Aromáticos monocíclicos halogenados sustituidos	Clorotalonil, tecnazeno, dicloran, hexaclorobenceno, quintozeno, dinocap, diclorofeno, pentaclorofenol, cloroneb
Ftalimidas	Captan
Anilinopirimidinas	Mepanipyrim, pirimetanil, ciprodinil
Derivados del ácido carbámico	Ferbam, tiram, ziram, propamocarb, maneb, mancozeb, zineb, nabam
Benzimidazoles	Benomilo, tiofanato-metil, carbendazima, fuberidazol,
Conazoles	Ciproconazol, diniconazol, etridiazol, hexaconazol, penconazol, triadimefon
Morfolinas	Dodemorf, fenpropimorf, tridemorf
Amidas	Fenhexamida, benalaxil, acilanina, flutolanil, tolilfluanida, diclofuanida
Otros	Tiabendazol, cicloheximida, fludioxonil,

	dimetomorf, trifloxystrobin, fenpiroximato
--	--------------------------------------------

Tabla 1.1: Nombre de los tipos de grupos químicos y algunos ejemplos de fungicidas según su tipo. Tabla correspondiente a la clasificación de Gupta, 2017 y la inclusión de los fungicidas inorgánicos y metálicos de Gupta, 2018.

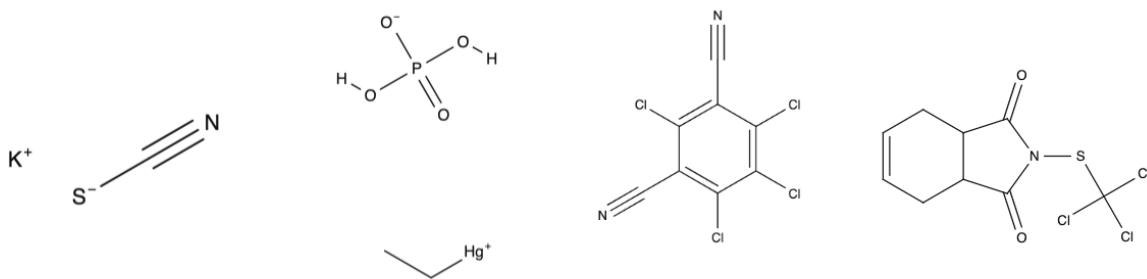
A continuación, se describen a mayor detalle cada uno de estos tipos (Gupta, 2018):

1. Fungicidas Inorgánicos: Formado principalmente por azufre (figura 1.1, a), que comenzó a utilizarse entre el siglo 19 y principios del siglo 20. Comúnmente se usan como fungicidas el azufre elemental y el azufre de cal crudo. La propiedad del azufre que destaca para esta tarea es su tendencia a oxidarse espontáneamente. Esto provoca efectos sobre los ojos, piel y tracto respiratorio, los cuales pueden ser mitigados evitando su aplicación a altas temperaturas ambiente. Usualmente, el azufre no presenta problemas toxicológicos, a menos que se esté tratando con azufre micronizado, cuyo veneno puede dejar secuelas en el tracto gastrointestinal, neurológico y efectos pulmonares.
2. Fungicidas metálicos: Cumplen una función protectora y preventiva ante hongos. En un principio, estos estaban compuestos por mercurio (figura 1.1, b) y eran aplicados en tratamientos de semillas para cereales y remolacha. Posteriormente, el mercurio y compuestos mercuriales han sido descontinuados debido a su toxicidad, teniendo efectos hepatotóxicos e inmunotóxicos. En la actualidad, se ha profundizado en el estudio de fungicidas basados en cobre, evaluando su efectividad sobre la mancha negra en cítricos (Franco et. al., 2020), su implicancia ambiental en el uso de fruta cítricas (Triantafyllidis et. al., 2020) y su impacto en la calidad antioxidante del lúpulo (Chrisfield et. al., 2021), obteniendo resultados variados.
3. Aromáticos monocíclicos halogenados sustituidos: Compuestos formados en su gran mayoría por al menos un anillo aromático y un elemento halógeno (ver figura 1.1, c), con algunas excepciones como el dinocap, que está ausente de elementos halógenos. Se han registrado problemas a la salud tales como irritación a la piel, lesiones en los

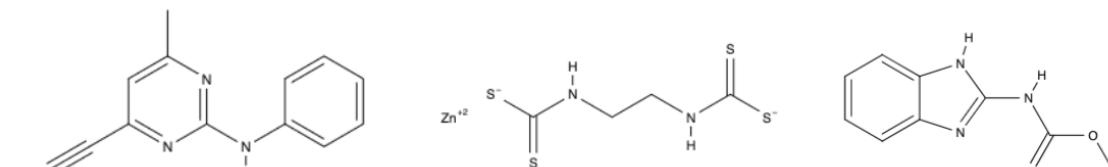
ojos y sensibilización a la piel. Químicos como cloroneb, quintozeno y dicloran, pertenecientes a este grupo, presentan bajos niveles en toxicidad.

4. Ftalimidas: Este tipo de fungicidas son imidas derivados del ácido ftálico. Usualmente, los tipos de químicos usados eran folpet, captafol y captan (figura 1.1, d), no obstante, los dos primeros han sido des registrados y solo captan se ocupa para tratamiento contra hongos. Este tipo de químicos son usados como protección superficial en muchos cultivos. Captan ha presentado efectos gastrointestinales en ratones, dilatación del intestino delgado e hiperplasia epitelial focal.
5. Anilinopirimidinas: Compuestos formados principalmente por un benceno y una pirimidina unidos por un grupo amino (ver figura 1.1, e). Este tipo de compuestos posee un bajo nivel de toxicidad, por lo que no debería presentar problemas en uso normal. A constante exposición, se han visto importantes afecciones relacionadas con la tiroides y riñón.
6. Derivados del ácido carbámico: Los fungicidas derivados del ácido carbámico incluyen ditiocarbamatos y bisditiocarbamatos de etileno (EDBCs). Estos compuestos forman complejos con elementos como manganeso, zinc y sodio formando fungicidas como maneb, zineb, nabam, entre otros (figura 1.1, f). Algunos signos de toxicidad incluyen hipocinesia, letargo, postura encorvada, temblores corporales y hemorragias nasales.
7. Benzimidazoles: Estos compuestos son hidrocarburos aromáticos, se representa como una fusión entre un imidazol y un benceno (figura 1.1, g). La mayoría de los benzimidazoles usados como fungicidas incluyen benomilo, carbendazima y fuberidazol. Estos son clasificados con toxicidad baja para benomilo y carbendazima, y media para fuberidazol. En altas dosis de ingesta oral, provoca efectos tóxicos en la reproducción masculina y femenina.

8. Conazoles: En su mayoría, derivados imidazólicos (figura 1.1, h) que poseen una toxicidad aguda que oscilan entre baja y moderada. Algunos de estos fungicidas no provocan irritación como el triadimenol, mientras otros provocan sensibilidad como el triadimefon. Por otro lado, propiconazol produce irritación a la piel. Se ha observado que el triadimenol también tiene incidencias en la toxicidad del desarrollo, afectando los órganos reproductores en mamíferos.
9. Morfolinas: Fungicidas que contienen una morfolina en su estructura química (figura 1.1, i). Dentro de este grupo hay compuestos con toxicidad leve, tales como dodemorf que puede presentar irritación moderadamente la piel y más severa en los ojos de conejos, y fenpropimorf, que solo irritan su piel. En cambio, tridemorf no produce irritación, pero está calificado con toxicidad moderada, produciendo malformaciones en el desarrollo de mamíferos, al igual que fenpropimorf.
10. Amidas: Componentes de baja toxicidad, excepto por el metalaxy que es ligeramente peligroso. Sus estructuras suelen tener al grupo amino acompañado de un anillo aromático (figura 1.1, j). Así como se ha visto en las clases de fungicidas descritas anteriormente, todo depende de la concentración y el nivel de exposición que se tiene sobre el fungicida. Por ejemplo, a una larga exposición del benalaxil, puede producir esteatosis hepática (acumulación excesiva de grasa en el hígado) y atrofia en túbulos seminíferos (producción de esperma) en ratas y perros, respectivamente. También se han visto otros efectos como irritación, en el uso de otros fungicidas de este tipo.
11. Otros: En esta clase se incluyen sustancias antibióticas, tiocarbonatos y derivados del ácido cinámico (ver figura 1.1, k). Poseen una toxicidad moderada, presentando irritación en la piel y ojos, en conejos. Algunos compuestos de esta lista han presentado toxicidad para el desarrollo y reproducción.



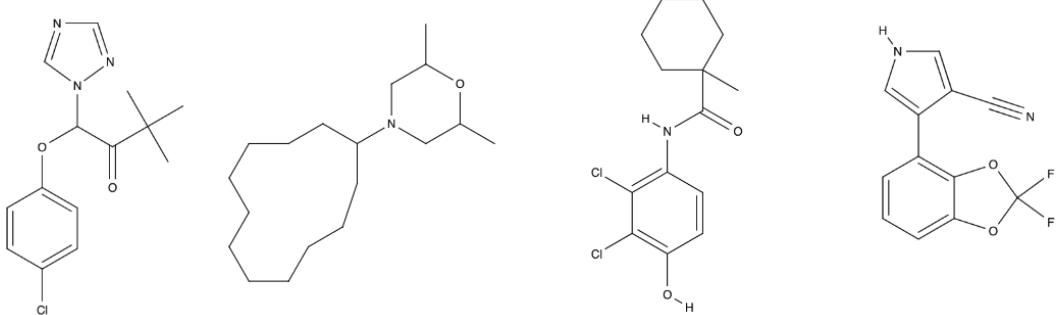
a) Tiocianato de potasio b) Fosfato de etilmercurio c) Clorotalonil d) Captan



e) Mepanipyrim

f) Zineb

g) Carbendazima



h) Triadimefon i) Dodemorf j) Fenhexamida k) Fludioxonil

Figura 1.1: Ejemplos de estructuras químicas de fungicida, con sus respectivos nombres. La clase correspondiente a cada compuesto sigue este orden: a) fungicidas inorgánicos, b) fungicidas metálicos, c) aromáticos monocíclicos halogenados sustituidos, d) ftalimidas, e) anilinopirimidinas, f) derivados del ácido carbámico, g) benzimidazoles, h) conazoles, i) morfolinas, j) amidas y k) fungicidas que poseen otros grupos químicos.

Esta clasificación muestra las propiedades de la estructura química de los fungicidas, permitiendo la interacción con organismos fungi, provocando su erradicación. Asimismo, se observaron algunas patologías o efectos en la salud de mamíferos, que se producen al contacto prolongado con estos químicos.

1.4 Mecanismos de Acción.

El modo de aplicación de los fungicidas se agrupa en foliares, suelo y apósito. Los foliares se aplican de forma líquida o en polvo a las partes aéreas de la planta (tallos, hojas, etc.), funcionando como una barrera protectora. Los de suelo son aplicados en forma de polvo, líquido o granulado, actuando mediante el vapor o propiedades sistémicas. Los apóstitos se aplican de forma líquida o polvo seco, utilizados principalmente cuando el almacenaje no posee las condiciones óptimas en temperatura y humedad necesaria (Gupta, 2018). Una vez aplicados, los fungicidas actúan de forma antagónica sobre los hongos a través de distintos mecanismos moleculares, interactuando con proteínas para interrumpir rutas biosintéticas vitales para estos organismos.

Existen 56 mecanismos de acción específicos diferentes entre fungicidas y bactericidas, los cuales son clasificados por el Comité de Acción de Resistencia a Fungicidas (FRAC) (Hermann et. al., 2019). FRAC pertenece a CropLife International, una asociación comercial de compañías agroquímicas, que funciona internacionalmente. Año tras año, se genera una lista actualizada de los mecanismos de acción con su respectiva clasificación. En este caso, la lista del año 2021 (CropLife, 2021). La clasificación de los mecanismos de acción se representa por letras, representadas de A a I, P, U, M y BM (Tabla 1.2), las cuales son denominadas códigos MOA (Mode-Of-Action Code). Cada una de estas letras representa un mecanismo de acción diferente. FRAC, además, incluye información sobre el sitio objetivo o de acción, el nombre del grupo o clase, grupo químico o biológico, nombre común de los fungicidas relacionados y comentarios sobre resistencias detectadas.

Código MOA	Mecanismo de Acción
A	Metabolismo de ácido nucleicos
B	Citoesqueleto y proteína motora
C	Respiración celular

D	Síntesis de aminoácidos y de proteínas
E	Transducción de señales
F	Síntesis o transporte de lípidos/Integridad o función de la membrana
G	Biosíntesis de esterol en membranas
H	Biosíntesis de la pared celular
I	Síntesis de melanina en la pared celular
P	Inducción de defensa en la planta hospedera
U	Desconocido
M	Químicos con multisitios de acción
BM	Componentes biológicos con múltiples modos de acción.

Tabla 1.2: Clasificación de los métodos de acción y su código. Tomado de las clasificaciones realizadas por FRAC.

También se han realizado otros estudios respecto al mecanismo de acción en fungicidas modernos (Baibakova et. al., 2019), que podría complementar la información de FRAC.

1.5 Ciencia de datos en biología.

La ciencia de datos es un campo interdisciplinario que tiene como objetivo convertir datos de distinto tipo en información que entrega un valor real, ya sean predicciones, decisiones automatizadas, modelos aprendidos o cualquier tipo de visualización que brinde información (Van der Aalst, 2016). Dentro de esta rama se encuentran áreas como la inteligencia artificial, aprendizaje automático, deep learning, big data y minería de datos.

En biología, ha tomado un rol importante, convirtiendo a la bioinformática en parte intrínseca a la investigación en ciencias de la vida. Provocando un llamado a entrenarse en esta área para satisfacer las necesidades científicas (Attwood et. al., 2019). Su rol radica en hacer evaluaciones con grandes cantidades de datos biológicos, característica que le destaca e inspira a formar profesionales en el medio (Greene et. al., 2016; Shaffer et. al., 2019). La biología del desarrollo, área que estudia la formación y desarrollo individual de un organismo, utiliza distintas técnicas de microscopía y secuenciación de alto rendimiento permitiendo estudiar embriones completos y la expresión génica en el desarrollo embrionario, abarcando el genoma completo, respectivamente. Dichas técnicas generan grandes cantidades de datos que imposibilitan analizarlos a mano, por este motivo, se utilizan herramientas de aprendizaje automático, específicamente deep learning, para el estudio de esta gran cantidad de datos (Villoutreix, 2021).

En biomedicina, el uso de la ciencia de datos ha permitido identificar factores a problemas de salud, asociándolo con poblaciones desfavorecidas, por ejemplo, se ha encontrado que la población de afroamericanos y nativos de Hawái y otras islas del Pacífico tienen una morbilidad y mortalidad por cáncer significativamente más altas que otros grupos étnicos, asimismo, se ha encontrado una serie de variables interrelacionadas como el nivel socioeconómico, el acceso a la atención médica, la genética, la obesidad, el consumo de tabaco y el consumo de alcohol que contribuyen a una salud dispar en diversas poblaciones étnicas (Canner et. al., 2017). En el desarrollo de drogas, se encuentran medicamentos obtenidos a partir de microorganismos, los cuales son denominados productos biológicos. Este tipo de fármaco ha tenido un gran éxito, habiendo 8 entre las 10 drogas más vendidas durante el año 2018. El fuerte de los productos biológicos es su alta especificidad y afinidad, una acción farmacocinética prolongada, baja toxicidad y efectos colaterales, en comparación a otras moléculas. Esto se debe a que sus ejemplares más exitosos satisfacen múltiples propiedades, entre las cuales se encuentran la actividad y características fisicoquímicas que se definen como su capacidad de desarrollo. El uso de inteligencia artificial y aprendizaje automático, permiten acelerar y mejorar la optimización de las propiedades proteicas, aumentando su actividad y seguridad, mientras disminuye el tiempo de desarrollo y costos asociados a su fabricación. El uso del aprendizaje automático permite el descubrimiento de nuevos productos biológicos, la

extracción de datos complejos y la reducción de esfuerzo experimental en el desarrollo de este tipo de drogas (Narayanan et. al., 2021).

La minería de datos es un concepto asociado al área denominada “Descubrimiento de Conocimiento en base de datos” o KDD, por sus siglas en inglés. Consiste en una serie de pasos como la preparación, post procesamiento y análisis de bases de datos (Piatetsky-Shapiro et. al., 2000). La minería de datos ha sido ampliamente utilizada para apoyar la investigación de problemas en biología y bioinformática, incluyendo predicción de estructura en proteínas, clasificaciones de genes modelados, estadísticas de interacciones proteína-proteína, entre otros (Yu et. al., 2003). El fundamento de esta área son las bases de datos en donde se alojan los elementos y atributos para analizar. Carugo & Eisenhaber, 2010 muestra que existen diversas bases de datos enfocadas en la biología: Estructura y secuencia de ácidos nucleicos, bases de datos y recursos genómicos, rutas metabólicas y, de secuencias y estructuras proteicas. En proteínas, existen bases de datos que almacenan características específicas como sus dominios, propiedades termodinámicas, bases de datos enfocadas a enzimas y, de interacciones proteína-proteína y sus complejos. Estos datos son evaluados con modelos de aprendizaje supervisado y no-supervisado, considerando algoritmos como clustering, Support Vector Machines y redes neuronales.

Existen varios trabajos que utilizan la minería de datos como base de sus investigaciones, entre estos se encuentra Schorn et. al., 2021 quienes han desarrollado la plataforma Paired Omics Data Platform (PoDP), repositorio que vincula datos genómicos y metabolómicos en un formato computacionalmente legible con el fin de desarrollar minería de datos con los datos recopilados en esta plataforma. El fin es estudiar a profundidad los metabolitos especializados, moléculas que entregan ventajas al organismo, como, por ejemplo, en la competencia de recursos nutricionales con otras especies. Los objetivos de esta plataforma son la vinculación entre esta molécula y sus productores, asociar a gran escala la relación genoma-metaboloma, usar datos genómicos para estudiar la estructura molecular y brindar una base de datos con conjunto de datos vinculados. Se espera, que este repositorio sea estudiado mediante técnicas de minería de datos para lograr los objetivos mencionados.

Por otra parte, en el área agrícola, se realizó un estudio para predecir múltiples perfiles ecotoxicológicos de fungicidas, bajo un enfoque quimioinformático. El set de datos contaba

con 81 tipos de agroquímicos fungicidas, descritos con un total de 124 atributos que relacionaban cuantitativamente su estructura y toxicidad, y clasificados como tóxicos y no tóxicos. El modelo, denominado como ms-QSTR, empleado para predecir la ecotoxicidad, tuvo un desempeño superior al 90%, tanto para la serie de entrenamiento como para la de predicción (Speck-Planche et. al., 2012). De esta forma, se consiguió correlacionar la estructura molecular de fungicidas con su impacto en el medio ambiente.

Estudios proteicos indican el análisis de 1.515 proteínas presentes en la saliva y 60 en la película dental. Mediante este análisis de minería de datos se obtuvieron las diferencias entre este tipo de proteínas que cohabitán en la cavidad oral, entre las cuales se encontró que las proteínas presentes en la película dental presentan un mayor fosforilación y glicosilación que las salivales, además, difieren significativamente en su actividad enzimática y unión iónica. Los hallazgos de este trabajo permiten la compartimentación del proteoma y así dar como resultado una función especializada, asimismo, se encontraron proteínas claves que participan en interacciones proteína-proteína o proteína-ligando las cuales podrían ser dianas para aplicaciones diagnósticas o terapéuticas (Schweigel et. al., 2016). Dentro de esta misma área, se evaluaron las interacciones catión- π , la cual se forma entre iones cargados positivamente y anillos aromáticos, y la influencia de residuos catiónicos en este tipo de interacciones. De esto se encontró que la arginina posee una mayor probabilidad de interactuar con ligandos aromáticos que la lisina, a pesar de que este supera en cantidad a la arginina (Kumar et. al., 2018). Para realizar los dos últimos trabajos mencionados, se utilizaron bases de datos públicas de proteínas, UniProt y RCSB PDB, respectivamente.

Ramakrishnan et. al, 2014 presenta el uso de métodos computacionales para caracterizar genes y proteínas a partir de su secuencia, permitiendo una mejor inferencia estructural y de la función proteica. La aplicación de esto ha permitido reconocer proteínas relacionadas con la evolución, aun cuando la similitud entre su secuencia es baja, y se han estudiado proteínas parasitarias del organismo *Trypanosoma brucei*, especie protista que afecta el sistema nervioso ocasionando tripanosomiasis africana o enfermedad del sueño en humanos y animales. Estas investigaciones han buscado relaciones secuencia-función y estructura-función en vías metabólicas, cuyo resultado ha sido la propuesta de nuevos fármacos para el tratamiento de esta enfermedad.

A modo de síntesis, en este capítulo se ha presentado la importancia de la fruticultura en Chile y el extranjero, las problemáticas que esta área presenta por la expresión de enfermedades causadas por hongos y el uso de fungicidas para suprimir la presencia de este tipo de patógenos. Compuestos que, a su vez, provocan daños en la salud de personas, animales y al medio ambiente, haciendo necesaria la invención y evaluaciones de nuevos componentes para atenuar dichas enfermedades. Asimismo, se ha presentado cómo el área de la minería de datos ha sido utilizada para distintos tipos de estudios biológicos, encontrando resultados variados para cada investigación, mostrándose como un área relevante y seria en este campo de estudio.

En base a lo anteriormente expuesto, este trabajo apunta recopilar y caracterizar datos de fungicidas y complejos que estos forman, utilizando repositorios que se describirán en secciones posteriores. Con el fin de encontrar relaciones entre este tipo de moléculas e interacciones, generando información que ayudará en el desarrollo o descubrimientos de nuevos agroquímicos o estrategias para prevenir patógenos fúngicos.

2 HIPÓTESIS

Existen numerosos fungicidas aplicados en fruticultura, algunos de estos interactúan con proteínas de patógenos conocidas. Tanto las moléculas fungicidas como los sitios de unión son posibles de caracterizar mediante sus propiedades estructurales, geométricas y fisicoquímicas. Estas propiedades permitirán identificar atributos relevantes para discriminar distintos tipos de mecanismos de acción, a través de técnicas en minería de datos. Teniendo esto en consideración, planteamos la siguiente hipótesis:

Es posible, mediante el análisis de minería de datos, encontrar atributos relevantes para discriminar entre los mecanismos de acción en la interacción molecular entre fungicidas y proteínas de patógenos fungi.

3 OBJETIVOS

3.1 Objetivo general

Utilizar técnicas de minería de datos para identificar propiedades y atributos relevantes para diferenciar distintos tipos de mecanismos de acción de fungicidas aplicados en la fruticultura.

3.2 Objetivos específicos

- Identificar fungicidas y proteínas dianas utilizados en estudios para prevención de hongos.
- Caracterizar propiedades fisicoquímicas, estructurales y geométricas de los fungicidas y complejos proteína-ligando identificados.
- Encontrar atributos relevantes de fungicidas para la predicción de su comportamiento e interacción con proteínas.

4 MATERIALES Y MÉTODOS

4.1 Materiales

4.1.1 Identificación y obtención de estructuras proteicas y de fungicidas

FRAC Code List

El Comité de Acción de Resistencia a Fungicidas (FRAC), es una organización conformada por el Grupo de Especialistas Técnicos de CropLife International. Su propósito es proporcionar pautas de manejo a resistencia a los fungicidas para prolongar su efectividad y limitar las pérdidas de cultivo en caso de resistencia.

Esta organización entrega un listado anual con los fungicidas utilizados para la protección de plantas, según su modo de acción y riesgo a resistencia. Este listado también incluye los bactericidas más importantes (CropLife, 2021).

Pesticide Properties Database (PPDB)

Base de datos online que almacena datos relacionados con propiedades de pesticidas como, por ejemplo, la velocidad a la que se disipan sobre o dentro de las plantas. Los datos que contiene son importantes para realizar evaluaciones de riesgo, ya sea de salud humana o ambientales. PPDB actualiza constantemente la lista de plaguicidas y los datos de cada uno (Lewis et. al., 2017). Las listas se dividen en cuatro tipos de pesticidas: (1) Insecticidas, (2) Herbicidas, (3) Fungicidas y (4) Otros tipos. Esta base de datos se puede visitar desde el siguiente enlace: <http://sitem.herts.ac.uk/aeru/ppdb/en/index.htm>

PubChem

PubChem es un repositorio de información química perteneciente al Centro Nacional de Información Biotecnológica de los Estados Unidos (NCBI). Esta plataforma tuvo sus inicios en el año 2004 y con el paso de los años ha ido creciendo en información incluyendo áreas como quimioinformática, biología química, química medicinal y descubrimiento de drogas (Kim et al., 2019). Junto esto, PubChem facilita la descarga de estructuras moleculares en 2D

y 3D en distintos formatos, permitiendo su estudio en la investigación. Este sitio web puede ser visitado siguiendo el enlace a continuación: <https://pubchem.ncbi.nlm.nih.gov>.

Universal Protein Resource (UniProt)

UniProt es una colección de bases de datos que permite encontrar una basta cantidad de información funcional y secuencial sobre proteínas. Las bases de datos UniProt son: UniProt Knowledgebase (UniProtKB), UniProt Reference Clusters (UniRef) y UniProt Archive (UniParc). UniProtKB es una combinación de Swiss-Prot (anotaciones revisadas) y TrEMBL (anotaciones automáticas no revisadas), UniRef proporciona conjuntos agrupados de secuencia incluyendo isoformas y UniParc entrega un set completo de secuencias conocidas, incluyendo secuencias obsoletas. Esta colección está formada por la colaboración entre el Instituto Europeo de Bioinformática (EMBL-EBI), el Instituto Suizo de Bioinformática (SIB) y el Recurso de información sobre proteínas (PIR) (Bateman et. al., 2017). A la fecha, UniProtKB contiene más de 550.000 secuencias con anotaciones revisadas y más de 200 millones con anotaciones no revisadas. Se puede acceder a través de: <https://www.uniprot.org/>.

Los formatos de archivos que se obtienen de UniProt son TSV (separado por tabulación), Excel, XML, RDF/XML, texto plano, GFF y FASTA. Este último se encuentra en formato canónico y canónico e isoforma.

RCSB Protein Data Bank

Base de datos que almacena estructuras tridimensionales de proteínas, obtenidas de experimentos tales como resonancia magnética nuclear (NMR), cristalografía de rayos X y criomicroscopía electrónica (Rose et. al., 2017). El formato en el que se almacenan estas estructuras es PDB, un formato de coordenadas atómicas, útil para determinados experimentos. Cada proteína se representa por un código de cuatro letras que las identifica. Esta base de datos se puede encontrar en el siguiente enlace: <https://www.rcsb.org/>.

4.1.2 Caracterización a nivel molecular de fungicidas

Mordred

Mordred es una herramienta basada en lenguaje Python, disponible en Windows, Linux y macOS, que describe la estructura molecular en formato 2D y 3D. Un descriptor molecular se define como el resultado final de un procedimiento lógico y matemático, el cual transforma la información química de una molécula en un número (Moriwaki et. al., 2018). Los atributos de Mordred son clasificados en 49 tipos o módulos que, durante el procedimiento de cálculo, van variando en sus distintos parámetros entregando un total de 1826 atributos. Este software trabaja mediante la línea de comandos o por paquetes de Python y se puede obtener mediante las plataformas de GitHub o Anaconda (<https://github.com/mordred-descriptor/mordred>, <https://anaconda.org/mordred-descriptor/mordred>).

4.1.3 Preparación y caracterización de complejos proteína-ligando

Protein Preparation Wizard (Schrödinger suite)

Protein Preparation Wizard es una suite de Schrödinger que permite llevar a la estructura proteica cristalizada y estructura molecular del ligando a una representación más exacta de la realidad. Esto, mediante la delección de moléculas de agua, asignación de cargas según el pH en el que se esté trabajando, asignación de estados rotámeros y cargas parciales, entre otras funciones. Esta suite, además, utiliza el algoritmo PROPKA para la predicción de estados de protonación en residuos (Bhachoo & Beuming, 2017).

AutoDock

AutoDock es una suite de herramientas que permiten evaluar el acoplamiento molecular (docking) de un ligando en un receptor. Para este estudio en particular se utilizará AutoDock Tools y AutoDock Vina. AutoDock Vina es un software altamente optimizado, realizando cálculos de forma paralela (Forli et. al., 2016). El score de cada cálculo está basado en el cálculo de energía libre de unión, con un enfoque en el aprendizaje automático utilizando la base de datos PDBbind (Trott & Olson, 2009).

Protein, Ligand, Geometrical and Physicochemical Properties 3D (PLGP3D)

PGLP3D es un software desarrollado por el equipo de investigación, que permite calcular diversas propiedades geométricas y fisicoquímicas que caracterizan los sitios de unión a ligando de los complejos que se deseen estudiar. Para esto, se debe entregar una lista con los nombres de todos los archivos de las estructuras proteicas que se encuentran en formato PDB. Este programa trabaja en conjunto con CGAL, Foldx y McVol, para cálculos geométricos, de energías y propiedades de las cavidades proteicas, respectivamente.

4.1.4 Implementación de selección de atributos y técnicas de minería de datos.

Python

Python, en especial Python 3, ha sido utilizado para el desarrollo de tareas científicas, incluyendo el análisis y visualización de grandes set de datos (VanderPlas, 2019). Esto último se debe a las distintas librerías que posee Python para el trabajo en el área de la ciencia de datos. Algunas las librerías utilizadas son: Numpy, Pandas, Scipy, Matplotlib, Imbalanced-Learn y Scikit-Learn; librerías que permiten el manejo, procesamiento y visualización de los datos.

4.2 Metodología

La metodología utilizada se visualiza en la figura 4.1, donde se muestran los pasos correlativos que se siguieron para alcanzar los objetivos estipulados:

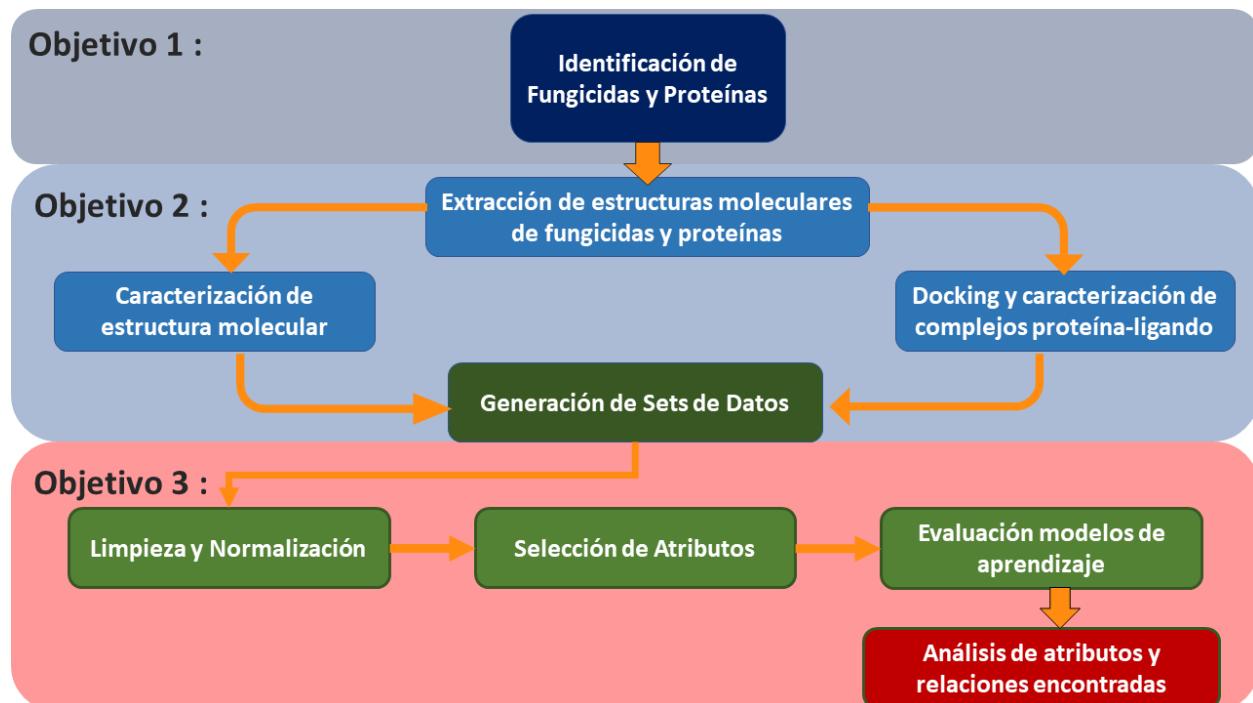


Figura 4.1: Diagrama de la metodología general aplicada en este estudio

4.2.1 Identificación y obtención de estructuras moleculares.

1. Identificación de fungicidas y proteínas

La identificación de los fungicidas y las proteínas se realizó en base al listado anual de FRAC (FRAC, 2021). En este listado se seleccionaron los fungicidas que tenían una clasificación conocida respecto a su riesgo a resistencia y a las proteínas que son parte del mecanismo de acción de estas moléculas. Se buscó, además, el grado de daño que producen estos compuestos en ecotoxicidad, medio ambiente y en la salud humana que se encuentran en el repositorio PPDB.

2. Obtención de estructuras moleculares

Las estructuras moleculares de los fungicidas se obtuvieron desde la base de datos PubChem, extrayendo las estructuras tridimensionales en formato SDF.

Las proteínas se buscaron utilizando el filtro de búsqueda avanzada de UniProt basado en los parámetros a continuación:

- Protein Name [DE]: El nombre específico de cada proteína a estudiar.
- Taxonomy [OC]: Eukaryota. Se utiliza este filtro para que solo aparezcan organismos del dominio Eucariota.
- Reviewed: Se realiza la búsqueda de secuencias aminoacídicas que estén con anotaciones revisadas y validadas en literatura.

Sus estructuras tridimensionales fueron descargadas desde la base de datos PDB.

4.2.2 Caracterización de moléculas y complejos proteína-ligando

1. Caracterización de fungicidas

Las estructuras moleculares de fungicidas fueron procesadas por Mordred, donde se calcularon los atributos bidimensionales y tridimensionales de las moléculas. Mordred entrega como resultado un archivo CSV (Comma Separated Values) con los atributos calculados y los valores obtenidos de los cálculos realizados.

2. Estudio de acoplamiento molecular.

El acoplamiento molecular o docking requiere una previa preparación de las moléculas fungicidas y de las proteínas, la cual se realizó mediante la suite de Schrödinger Protein Preparation Wizard. Este procedimiento consistió en la supresión de moléculas de agua en los archivos de coordenadas, creación de enlaces disulfuro, corrección de superposición atómica y la predicción de estados de protonación por el algoritmo PROPKA.

El estudio de acoplamiento molecular se realizó utilizando las estructuras preparadas del paso anterior, a través del software AutoDock Tools y AutoDock Vina. Con AutoDock Tools se

definió a la proteína de forma rígida y a la grilla donde evaluó la interacción del ligando con la proteína, cuyo tamaño y coordenadas se ajustaron según el sitio de unión de cada proteína. Se utilizó AutoDock Vina para realizar los cálculos. Para cada evaluación, se escogió la conformación con mejor score.

La formación de los complejos consistió en la combinación de la mejor conformación y la proteína estudiada, a través de la función Merge de MAESTRO de Schrödinger.

3. Caracterización de complejos proteína-ligando.

PLPG3D es el programa que se utilizó para generar los atributos estructurales, geométricos y fisicoquímicos que caracterizan los sitios de unión al ligando de los complejos estudiados. Los parámetros para la ejecución de este programa se describen como siguen:

- Nombre de archivo: nombre de la lista de archivos a evaluar. Los archivos de estudio pueden tener como nombre el código PDB de cuatro letras o el que el usuario asigne. En esta lista estarán los nombres de los archivos que contienen el complejo proteína-ligando descritos anteriormente.
- Esfera de selección: Corresponde a la distancia o área de selección del sitio que se quiere estudiar desde el centro de masa del ligando. Para este proyecto se utilizará una distancia de 7 Å.
- Número de capas: corresponde a la cantidad de capas en la cual se dividirá la esfera de selección, estas capas son concéntricas al centro de masa del ligando y poseen diferente radio. Con una selección de 7 Å, el radio de cada capa será de 1 Å (Figura 4.2).

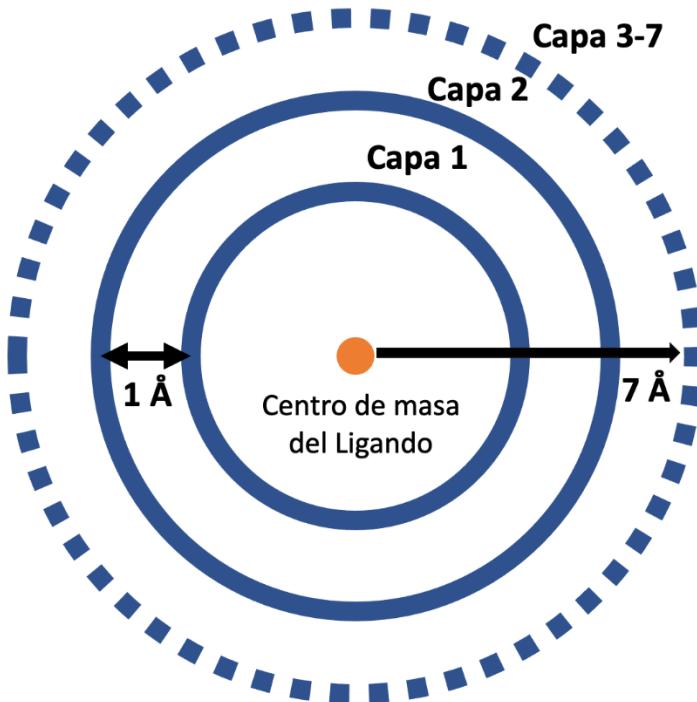


Figura 4.2: Representación gráfica de los parámetros para el cálculo mediante PLPG3D. El radio de la esfera es de 7 Å desde el centro de masa del ligando. Las líneas continuas representan las capas 1 y 2, las líneas segmentadas representan el conjunto de capas desde la 3 hasta la 7. La distancia entre capas es de 1 Å.

Los tipos de atributos que calcula PLPG3D son:

- *Distancias*: En base a los cuatro aminoácidos más cercanos al centro de masa del ligando, se calcula la distancia de carbonos alfa, carbonos beta y átomos con el centro de masa del ligando y entre estos mismos.
- *Ángulos*: En base a los cuatro aminoácidos más cercanos al centro de masa del ligando se calculan cuatro diferentes tipos de ángulos.
- *Torsiones*: Cada torsión o ángulo diedro corresponde al ángulo que se forma entre dos planos y pueden observarse al mirar lo largo de la línea de intersección. PLPG3D hace el cálculo de cuatro tipos de torsiones, basados en el ángulo Phi, Psi, Omega y una torsión formada entre los cuatro aminoácidos más cercanos al centro de masa del ligando.
- *Hidrofobicidad*: Propiedad fisicoquímica que cumple un rol fundamental en la comprensión de la estructura secundaria de la proteína. Existen diferentes escalas para

estimar la hidrofobicidad de los aminoácidos, PLPG3D considera la propuesta de Hessa (Hessa et al., 2005), Kyte & Doolittle (Kyte & Doolittle, 1982) y Wimley & White (Wimley & White, 1996).

- *Composición aminoacídica*: Se calculan las proporciones de los aminoácidos que se encuentran en el sitio de unión, resultando la cantidad de aminoácidos no polares, polares neutros, polares negativos, polares positivos y el total de aminoácidos del sitio.
- *Composición atómica por capa*: Se calcula la proporción de átomos de carbono, nitrógeno, azufre y oxígeno por cada capa, en base al total de átomos de cada una de estas zonas.
- *Interacciones*: Se consideran dos tipos de interacciones, aquellas que se producen con átomos de carbono, nitrógeno, azufre u oxígeno, sólo si estos se encuentran a una distancia máxima de 5 Å. Las otras interacciones son sólo de las cadenas laterales de los aminoácidos y se clasifican en hidrofilicas, hidrofóbicas, aromáticas, donadoras y aceptoras.
- *Energía*: El cálculo de energías se realizó de dos formas, una fue utilizando la función de energía FOLDEF, que está implementada en el software Foldx, entregando diversos valores energéticos de los complejos y propiedades fisicoquímicas. La otra forma consiste en calcular la energía de captación utilizando la escala propuesta por Sneddon, Morgan & Brooks (Sneddon et al., 1988).
- *Radios de sitio*: Cálculos del promedio de distancias, desviación estándar de las distancias, distancia mínima y máxima correspondiente al sitio de unión, con el fin de comprender mejor su forma y tamaño.
- *Aminoácidos cercanos*: Se contemplan 12 aminoácidos cercanos, según la distancia entre el su carbono alfa y el centro de masa del ligando. También se consideran los aminoácidos que se encuentran en la zona determinada (Figura 4.3).
- *Volumen*: Cálculo del volumen de todos los átomos incluidos en capa.
- *Alpha Shapes*: Técnica que representa a la estructura proteica en puntos, correspondiente a los carbonos alfa. Estos puntos se unen de manera única formando un conjunto no superponible de tetraedros irregulares, conocidos como triangulación de Delaunay, en el cual los residuos unidos por una arista son vecinos más cercanos. Se calcula con ayuda de CGAL.

- *Cavidades:* Con ayuda de McVol, se calcularon las cavidades o hendiduras presentes en las proteínas estudiadas. El cálculo consistió en la cantidad de cavidades, promedio de tamaño de las cavidades, desviación estándar del tamaño de las cavidades, tamaño mínimo y máximo de las cavidades.

Al igual que Mordred, PLPG3D entrega un archivo CSV con los atributos y los resultados de sus cálculos.

4.2.3 Generación de los sets de datos

En este estudio se trabajó con dos tipos de información por separado.

Un set de datos contiene los atributos calculados por PLPG3D en base a las propiedades fisicoquímicas, geométricas y estructurales de las zonas de unión a ligando de las proteínas. Se buscó la correlación entre estos atributos y de los tipos de interacciones que se producen.

El otro set de datos contiene atributos calculados por Mordred en base a las estructuras moleculares de los fungicidas. Con estos, se evaluó su correlación con las interacciones en complejos proteína-ligando y con los grados en los efectos colaterales de los fungicidas.

4.2.4 Minería de datos

1. Preprocesamiento de datos

La limpieza de datos es un proceso que se ocupa de detectar y eliminar los errores o inconsistencias de los datos, con el fin de mejorar la calidad de estos y evitar errores en los análisis aplicados. La corrección y estandarización de los datos verifica la exactitud de los datos analizados (Azeroual et. al., 2019).

A través de la librería Pandas de Python se trabajaron todos los datos. En primer lugar, se realizó una limpieza de los atributos y ejemplos que tenían un exceso de datos nulos. Los datos nulos son espacios vacíos producidos por error de cálculo u omisión de este, por ende, no entrega valor alguno para el análisis. Para su limpieza, primeramente, se calculó la cantidad promedio de datos nulos que tenían los atributos y ejemplos, para posteriormente suprimir los que tenían una cantidad superior al promedio. Los datos nulos restantes se reemplazaron por el valor promedio de cada atributo, según su clase.

2. Normalización

La estandarización o normalización busca la uniformidad de los datos para asegurar un correcto cálculo y análisis de cada set. En este trabajo se utilizó el método Min-Max (formula 4.1) para aplicar la normalización de los datos.

$$\text{Dato normalizado} = \frac{(\text{valor original} - \text{valor mínimo})}{(\text{valor máximo} - \text{valor mínimo})}$$

Fórmula 4.1: Normalización con el método Min-Max.

El valor mínimo y valor máximo corresponde al valor más pequeño y grande del atributo al que corresponde el valor original, respectivamente.

3. Balance de clase

En algunos escenarios evaluados se presenta el caso en donde existe una diferencia significativa entre el número de ejemplos disponible para distintas clases. Por este motivo, es posible que los modelos de clasificación generados tiendan a cometer mayor error en la predicción de la clase minoritaria. Para las tareas donde se encontraron errores relacionados a esta situación se aplicó el método SMOTE de la librería Imbalanced-Learn, igualando el número de ejemplos para las clases. SMOTE genera ejemplos artificiales en base a los datos de los ejemplos existentes, según cada categoría.

4. Encontrar atributos diferenciadores

La maldición de la dimensionalidad es un problema que surge cuando la cantidad de atributos o dimensiones es mayor que el número de ejemplos o casos a clasificar. Su efecto potencial es una disminución del desempeño en modelos de aprendizajes por aumentar la dispersión de los datos. Para evitar este potencial problema, se procedió a encontrar y seleccionar aquellos atributos más destacados, que permitan diferenciar de mejor manera las categorías o clases definidas en los diferentes escenarios evaluados.

Para encontrar estos atributos, se utilizó el método llamado “Mutual Information”. Este método mide el nivel de dependencia entre 2 variables, en este caso entre cada uno de los atributos y las clases definidas. Mutual-Information está disponible también en las librerías de Scikit Learn de Phyton.

5. Análisis con técnicas de minería de datos.

Los modelos de aprendizaje supervisado, también conocidos como modelos predictivos o modelos de clasificación, buscan la predicción de la clase o categoría de un nuevo ejemplo en base a sus atributos. Para cumplir esta tarea, es necesario un set de datos de entrenamiento con una serie de ejemplo etiquetados, es decir, cuya clase o categoría sea conocida. De esta forma, si el modelo logra relacionar los atributos con cada categoría, diferenciándolas entre sí, alcanzará un alto desempeño en sus predicciones. El funcionamiento de este tipo de método de aprendizaje se visualiza en la figura 4.3.

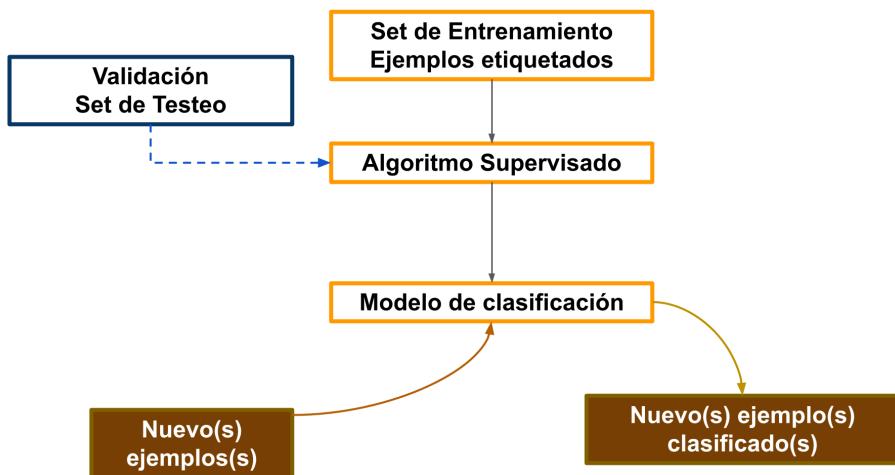


Figura 4.3: Representación gráfica del modelo de aprendizaje supervisado.

El modelo de clasificación que se utilizó en este trabajo fue Random Forest, un mecanismo que genera una serie de clasificadores de árboles de decisión en base a subconjuntos de datos desde el set de datos inicial. Random Forest finalmente entrega un consenso para predecir a qué clase o categoría pertenece un nuevo ejemplo. Si bien inicialmente se evaluaron otro tipo

de modelos de clasificación, Random Forest siempre fue el que presentaba los mejores resultados y por eso se decidió utilizarlo en todo el estudio.

En este trabajo se plantean diferentes tareas de clasificación que serán enfrentadas utilizando modelos de tipo Random Forest. Estas son las siguientes:

- Clasificar si una molécula fungicida interactúa o no con una proteína específica. Esto para cada uno de los tipos de interacción ligando-proteína identificados.
- Predecir el grado de efecto colateral para cada fungicida identificado. Considerando el riesgo a resistencia, ecotoxicidad, daño a la salud, y daño al medio ambiente.

6. Evaluación de desempeño

La performance en los modelos de aprendizaje se evaluó a través de medidas de desempeño para determinar cuán bien se desarrolló la tarea de clasificación.

Para esto se utilizó una metodología de validación cruzada llamada “Leave-One-Out cross validation” (LOOCV). Que consiste en realizar tantas evaluaciones como número de ejemplos disponibles. En cada evaluación uno de los ejemplos es extraído del set de datos disponible y por lo tanto no es utilizado para entrenar el modelo de clasificación. El modelo de clasificación se evalúa entonces sobre el caso extraído inicialmente para verificar si el modelo clasifica correctamente.

Todas estas evaluaciones son organizadas en una matriz de confusión, herramienta empleada en modelos de aprendizaje supervisado que visualiza el desempeño obtenido, como se presenta en la tabla 4.1.

		Predicción	
		Clase 1	Clase 2
Real	Clase 1	VP	FN
	Clase 2	FP	VN

Tabla 4.1: Representación de una matriz de confusión. Las columnas corresponden a la clasificación de los ejemplos definida por la predicción del modelo y las filas corresponden a su clase correcta. La cantidad de ejemplos que fueron predichos correctamente se denominan Verdaderos Positivos (VP) y Verdaderos Negativos (VN). Los ejemplos que fueron clasificados erróneamente pasan a ser Falsos Positivos (FP) o Falsos Negativos (FN).

Las medidas de desempeño utilizadas en cada evaluación se presentan a continuación:

- Exactitud: La exactitud o eficiencia global, indica el porcentaje de acierto que tuvo el modelo en relación con las predicciones realizadas. Su forma de cálculo se representa en la fórmula 4.2.

$$Exactitud = \frac{VP + VN}{VP + FN + FP + VN}$$

Fórmula 4.2: Cálculo de exactitud del modelo.

- Sensibilidad: Es la proporción entre la cantidad de ejemplos predichos correctamente y el total de ejemplos que pertenecen a esa clase (formula 4.3). Mientras más cercano al 100%, mejor es el desempeño.

$$Sensibilidad = \frac{VP}{VP + FN}$$

Fórmula 4.3: Cálculo de sensibilidad.

5 RESULTADOS

5.1 Resultados de estudios con complejos proteína-ligando

FRAC publica anualmente un listado que asocia a fungicidas con su modo de acción y su riesgo a resistencia. Desde este listado se identificaron 223 fungicidas que eran utilizados hasta el año 2021 y 20 proteínas que interactúan con algunos de estos compuestos. En la tabla 5.1 se presenta el listado de proteínas identificadas y los fungicidas con los que interactúan.

Proteínas	Fungicidas con los que interactúa		
RNA Polimerasa I	Benalaxyll	Metalaxyll	Oxadixyl
	Benalaxyll-M	Metalaxyll-M	Ofurace
	Furalaxyll		
Adenosin Deaminasa	Bupirimate	Dimethirimol	Ethirimol
Beta Tubulina	Benomyl	Thiabendazole	Diethofencarb
	Carbendazim	Thiophanate	Zoxamide
	Fuberidazole	Thiophanate-methyl	Ethaboxam
Espectrina	Fluopicolide	Fluopimonide	
Actina, Miosina o Fimbrina (3)	Phenamacril	Metrafenone	Pyriofenone
Complejo I: NADH deshidrogenasa	Diflumetorim	Tolfenpyrad	Fenazaquin
Complejo II: Succinato Deshidrogenasa	Benodanil	Thifluzamide	Penflufen
	Flutolanil	Benzovindiflupyr	Penthiopyrad
	Mepronil	Bixafen	Sedaxane
	Isofetamid	Fluindapyr	Isoflucypram
	Fluopyram	Fluxapyroxad	Pydiflumetofen
	Fenfuram	Furametpyr	Boscalid
	Carboxin	Inpyrfluxam	Pyraziflumid
	Oxycarboxin	Isopyrazam	
Complejo III: Citocromo BC1	Azoxystrobin	Pyraclostrobin	Metominostrobin
	Coumoxystrobin	Pyrametostrobin	Orysastrobin
	Enoxastrobin	Triclopyricarb	Famoxadone
	Flufenoxystrobin	Kresoxim-methyl	Fluoxastrobin

	Picoxystrobin Pyraoxystrobin Mandestrobin	Trifloxytrobin Dimoxytrobin Fenaminstrobin	Fenamidone Pyribencarb Metyltetraprole
ATP Sintasa	Fentin Acetate	Fentin Chloride	Fentin Hydroxide
Histidina quinasa	Fenpiclonil Fludioxonil Chlozolinate	Dimethachlone Iprodione	Procymidone Vinclozolin
Metiltransferasa	Edifenphos Iprobenfos	Pyrazophos Isoprothiolane	
C-14 desmetilasa	Triforine Pyrifenoxy Pyrisoaxazole Fenarimol Nuarimol Imazalil Oxoconazole Pefurazoate Prochloraz Triflumizole Azaconazole Bitertanol	Ipconazole Mefentrifluconazole Bromuconazole Cyproconazole Difenoconazole Diniconazole Epoxiconazole Etaconazole Fenbuconazole Fluquinconazole Flusilazole Flutriafol	Hexaconazole Imibenconazole Metconazole Myclobutanil Penconazole Propiconazole Simeconazole Tebuconazole Tetraconazole Triadimefon Triticonazole Prothioconazole
Delta-14 reductasa	Aldimorph Dodemorph Fenpropimorph	Tridemorph Fenpropidin	Piperalin Spiroxamide
3-keto reductasa	Fenhexamid	Fenpyrazamine	
Squalene monooxygenase	Pyributicarb	Naftifine	Terbinafine
Chitin Synthase	Polyoxin		
Celulosa sintasa	Dimethomorph Flumorph Pyrimorph	Benthiavalicarb Iprovalicarb	Valifenalate Mandipropamid

Tabla 5.1: Proteínas identificadas y los fungicidas con los que interactúan, según FRAC.

De estas proteínas se seleccionaron aquellas que interactúan con al menos 7 fungicidas, para así tener una cantidad de ejemplos suficientes en los análisis con los modelos de clasificación. En base a esta selección, se buscaron estructuras cristalizadas provenientes de organismos fungi o animal.

Bajo estas condiciones, se encontraron estructuras para las proteínas RNA Polimerasa I, Beta Tubulina, Succinato Deshidrogenasa y Citocromo BC1, las cuales fueron descargadas desde la base de datos PDB. El código PDB de estas estructuras son 4C2M (RNA Polimerasa I), 5S4N (Beta Tubulina), 1ZOY (Succinato Deshidrogenasa) y 1SQP (Citocromo BC1).

Debido a que estas estructuras no tenían incluidos a los fungicidas, se realizó un estudio de acoplamiento molecular. El resultado de este estudio entregó la mejor conformación de los compuestos y fijó al fungicida en la estructura proteica. Una vez formados los complejos proteína-ligando, fueron caracterizados por el software PLPG3D que calculó un total de 186 atributos para cada uno.

Este estudio buscó discriminar si un fungicida interactúa con una proteína o no, en un determinado sitio de unión. Para ello, fue necesario realizar un nuevo estudio de acoplamiento molecular entre las 4 proteínas encontradas y los fungicidas que se sabe no interactúan con ellas, se formaron los complejos proteína-ligando y fueron caracterizados por PLPG3D.

Esto resultó en sets de ejemplos positivos y negativos, siendo los positivos aquellos complejos de proteínas y ligandos que se sabe interactúan entre sí y los negativos, aquellos que se sabe que no interactúan.

El total de ejemplos para cada análisis fue la cantidad de fungicidas que interactúa específicamente con estas proteínas. Teniendo 7 fungicidas que interactúan con RNA Polimerasa I, 9 con Beta Tubulina, 23 con Succinato Deshidrogenasa y 21 con Citocromo BC1, quedó un total de 60 fungicidas para realizar esta evaluación.

De esta forma, el set de datos utilizado para evaluar la interacción proteína-ligando consistió en 60 ejemplos, 186 atributos y dos categorías: Positiva, si interactúa con la proteína, y Negativa, si es que no (Tabla 5.2).

Tareas de evaluación	Nº de atributos	Nº ejemplos con clase positiva	Nº de ejemplos con clase negativa
Interacción con RNA Polimerasa I	186	7	53
Interacción con Beta Tubulina	186	9	51
Interacción con Succinato Deshidrogenasa	186	23	37
Interacción con Citocromo BC1	186	21	39

Tabla 5.2: Descripción de los sets de datos para el análisis en interacción a proteína con atributos calculados por PLPG3D.

El método de selección de atributos, Mutual-information, se aplicó a estos sets de datos y cada uno generó un ranking de atributos en base a su relevancia para distinguir entre las categorías.

Para cada tarea, con estos rankings se buscó el número mínimo de atributos que entrega el mejor desempeño en la evaluación con Random Forest y el método de validación cruzada Leave-One-Out.

La tabla 5.3 muestra los resultados obtenidos de la evaluación de exactitud considerando el número óptimo de atributos seleccionados en cada caso. Los resultados muestran que las tareas de interacción con RNA Polimerasa I y Beta Tubulina alcanzaron una exactitud del 96.67% cada una. Además, se observan que en ambas tareas los modelos de clasificación alcanzan tasas de sobre el 85% en términos de sensibilidad para las distintas clases.

Tareas de evaluación	Nº de Atributos	Exactitud	Sensibilidad
Interacción con RNA Polimerasa I	13	96.67%	Clase (+): 85.71% Clase (-): 98.11%
Interacción con Beta Tubulina	13	96.67%	Clase (+): 88.89% Clase (-): 98.04%

Interacción con Succinato Deshidrogenasa	11	70%	Clase (+): 43.48% Clase (-): 86.49%
Interacción con Citocromo BC1	9	80%	Clase (+): 66.67% Clase (-): 87.18%

Tabla 5.3: Resultados del mejor desempeño según número de atributos.

Sin embargo, en las tareas relacionadas con la interacción a Succinato Deshidrogenasa y Citocromo BC1, se encontraron problemas relacionados al desbalance de las clases. Se observa claramente una baja significativa en la performance general de estas tareas, lo que se debe a bajos valores en la sensibilidad de la predicción de la clase positiva.

Para enfrentar esta situación se utilizó la función SMOTE para igualar el número de ejemplos para ambas clases (ver tabla 5.4).

Con este nuevo set de datos se realizó el mismo análisis anterior seleccionando los mejores atributos según Mutual information, utilizando el modelo de clasificación de Random Forest y el método de validación cruzada Leave-One-Out para estimar exactitud.

Tareas de evaluación	Nº ejemplos con clase positiva	Nº de ejemplos con clase negativa
Interacción con Succinato Deshidrogenasa	37	37
Interacción con Citocromo BC1	39	39

Tabla 5.4: Número de ejemplos por clase posterior al proceso de balance de clases con el método SMOTE.

Los resultados de esta nueva evaluación del desempeño para estas tareas, posterior al balance de clases, se puede visualizar en la tabla 5.5. Se observa un aumento significativo de la exactitud en ambos casos. Al mismo tiempo se observa un equilibrio en cuanto a tasas de sensibilidad para cada una de las clases.

Tareas de evaluación	Nº de Atributos	Exactitud	Sensibilidad
Interacción con Succinato Deshidrogenasa	13	79.73%	Clase (+): 78.38% Clase (-): 81.08%
Interacción con Citocromo BC1	12	87.18%	Clase (+): 89.74% Clase (-): 84.62%

Tabla 5.5: Resultados del mejor desempeño según número de atributos con clases balanceadas.

En la tabla 5.6 se pueden observar los tipos de atributos que son más relevantes para la construcción de cada uno de los modelos de clasificación generados que están asociados a las distintas tareas definidas anteriormente.

Proteína de Estudio	Tipos de Atributos relevantes
RNA Polimerasa I	<ul style="list-style-type: none"> - Energía - Interacciones - Distancia - Torsiones
Beta Tubulina	<ul style="list-style-type: none"> - Torsiones - Composición atómica - Energía - Interacciones - Composición aminoacídica - Volumen
Succinato Deshidrogenasa	<ul style="list-style-type: none"> - Torsiones - Energía - Volumen - Composición atómica - Distancia

	<ul style="list-style-type: none"> - Ángulos
Complejo BC1	<ul style="list-style-type: none"> - Volumen - Energías - Alpha Shapes - Distancias - Ángulos - Interacciones

Tabla 5.6: Atributos más relevantes para las tareas relacionadas con la interacción a proteína.

En el anexo 8.1 se pueden observar los atributos específicos que fueron considerados diferenciadores para cada tarea. Para corroborar la importancia relativa de estos atributos, las Figuras 8.1-8.4 muestra gráficos de tipo boxplot para comparar la distribución de valores de los distintos atributos en los ejemplos definidos para cada tarea. Se observa que, para lograr un buen desempeño de los modelos de clasificación, lo que implica alta exactitud y tasas equilibradas de sensibilidad para cada clase, se requiere considerar el efecto combinado de varios atributos.

5.2 Resultado de predicción de los niveles en efectos colaterales.

En una segunda etapa de esta memoria se buscó estudiar si es posible relacionar la estructura molecular de los fungicidas identificados, con algunos efectos colaterales que producen su uso. Los efectos colaterales estudiados fueron riesgo a resistencia, ecotoxicidad, daño al medio ambiente y en la salud humana.

Se observa que no todos los fungicidas tenían una clasificación asignada para los distintos efectos colaterales estudiados. La cantidad de fungicidas con un nivel identificado se encuentra en la tabla 5.7. Los efectos secundarios son medidos en distintos grados. FRAC entrega los niveles de riesgo a resistencia que presentan los fungicidas y PPDB el grado de daño ecotóxico, medioambiental y en el ser humano.

Efecto colateral	Nº de fungicidas
Riesgo a resistencia	191
Ecotoxicidad	163
Daño medioambiental	153
Daño a la salud humana	167

Tabla 5.7: Número de fungicidas con niveles identificados en cada efecto colateral estudiado.

Los niveles identificados fueron utilizados como categorías para hacer el análisis con modelos de aprendizaje supervisado. Todas las categorías se muestran en la tabla 5.8.

Tareas de evaluación	Nº de categorías	Nivel de efecto colateral (Categoría)
Nivel de riesgo a resistencia	5	Alto, Medio-Alto, Medio, Medio-Bajo, Bajo
Nivel ecotóxico	2	Alto, Medio
Grado de año medioambiental	2	Alto, Medio

Grado de daño a la salud humana	2	Alto, Medio
---------------------------------	---	-------------

Tabla 5.8: Efectos colaterales del uso de fungicidas y sus niveles identificados en el Comité de Acción al Riesgo a Resistencia (FRAC) y la Base de Datos de Propiedades de Pesticidas (PPDB).

El cálculo con Mordred entregó un total de 1826 atributos para cada fungicida. Sin embargo, algunos atributos y fungicidas obtuvieron altos números de datos nulos en los cálculos realizados por este software, por lo que se decidió eliminarlos del set de datos para no ocasionar problemas en las evaluaciones del modelo. El nuevo número de fungicidas por cada efecto colateral y número de atributos se presenta en la tabla 5.9.

Tarea de evaluación	Nº atributos	Nº de fungicidas	Nº de fungicidas por nivel
Nivel de riesgo a resistencia	1654	175	Clase Alto: 35 Clase Medio-Alto: 35 Clase Medio: 54 Clase Medio-Bajo: 33 Clase Bajo: 18
Nivel ecotóxico	1654	151	Clase Alto: 62 Clase Medio: 89
Grado de año medioambiental	1654	139	Clase Alto: 59 Clase Medio: 80
Grado de daño a la salud humana	1654	135	Clase Alto: 56 Clase Medio: 79

Tabla 5.9: Número de fungicidas y atributos posterior al proceso de limpieza.

En este caso se utilizó el mismo procedimiento que en la sección anterior para entrenar y evaluar modelos de clasificación de Random Forest. Seleccionando los atributos más relevantes para cada tarea definida, y evaluando los modelos generados mediante el método de validación cruzada Leave-One-Out para estimar tasas de exactitud y sensibilidad de cada clase definida.

Los resultados de esta evaluación se presentan en la Tabla 5.10. Se observa que en estas tareas de clasificación la exactitud alcanza valores menores al 75%. Sin embargo, se observa que en todas las situaciones el problema de desbalance de clases afecta la sensibilidad de las clases minoritarias en cada una de las tareas.

Tareas de evaluación	Nº de Atributos	Exactitud	Sensibilidad
Nivel de riesgo a resistencia	32	72.26%	Clase Alto: 91.43% Clase Medio-Alto: 51.43% Clase Medio: 87.04% Clase Medio-Bajo: 54.55% Clase Bajo: 66.67%
Nivel ecotóxico	14	68.87%	Clase Alto: 51.61% Clase Medio: 80.90%
Grado de año medioambiental	13	73.38%	Clase Alto: 64.40% Clase Medio: 80%
Grado de daño a la salud humana	9	65.93%	Clase Alto: 48.21% Clase Medio: 78.48%

Tabla 5.10: Resultado del desempeño de Random Forest para diferenciar entre los grados de los efectos colaterales por el uso de fungicidas.

Para enfrentar esta situación se utilizó nuevamente la función SMOTE para igualar el número de ejemplos de las distintas clases. Con este nuevo set de datos se realizó el mismo análisis anterior seleccionando los mejores atributos según Mutual information, utilizando el modelo de clasificación de Random Forest y el método de validación cruzada Leave-One-Out para estimar exactitud.

Los resultados de esta nueva evaluación del desempeño para estas tareas, posterior al balance de clases, se puede visualizar en la tabla 5.11. Se observa un aumento significativo de la exactitud en todas las tareas de clasificación asociadas al nivel de efectos colaterales de los fungicidas. Al mismo tiempo, se observa un equilibrio en cuanto a tasas de sensibilidad para cada una de las clases definidas en cada tarea.

Tareas de evaluación	Nº de Atributos	Exactitud	Sensibilidad
Nivel de riesgo a resistencia	24	81.48%	Clase Alto: 88.89% Clase Medio-Alto: 79.63% Clase Medio: 81.48% Clase Medio-Bajo: 70.37% Clase Bajo: 87.04%
Nivel ecotóxico	24	76.97%	Clase Alto: 78.65% Clase Medio: 75.28%
Grado de año medioambiental	21	75.63%	Clase Alto: 75% Clase Medio: 76.25%
Grado de daño a la salud humana	27	70.89%	Clase Alto: 74.68% Clase Medio: 67.09%

Tabla 5.11: Resultado del desempeño de Random Forest con set de datos balanceado para diferenciar entre los grados de los efectos colaterales por el uso de fungicidas.

Finalmente, la tabla 5.12 presenta los tipos de atributos que son más relevantes para la construcción de cada uno de los modelos de clasificación generados en esta etapa. Debido al alto número de atributos seleccionados en cada caso, los atributos específicos y los gráficos de tipo bloxplot que compara valores de los distintos atributos seleccionados y sus clases, se presentan en el anexo 8.2.

Efecto colateral	Tipos de Atributos relevantes
Nivel de riesgo a resistencia	- BCUT - DetourMatrix - BaryszMatrix - WalkCount - DistanceMatrix - ExtendedTopochemicalAtom - Chi

Nivel ecotóxico	<ul style="list-style-type: none"> - BaryszMatrix - MoeType - AutoCorrelation - PathCount - Chi - MoRSE - WalkCount
Grado de año medioambiental	<ul style="list-style-type: none"> - Framework - AutoCorrelation - MolecularId - EState - MoeType - AdjacencyMatrix - BaryszMatrix - InformationContent - TopoPSA - Chi
Grado de daño a la salud humana	<ul style="list-style-type: none"> - Autocorrelation - DetourMatrix - MoeType - BCUT - BaryszMatrix - Chi - DistanceMatrix - MoRSE - TopologicalIndex - TopologicalCharge

Tabla 5.12: Atributos más relevantes para las tareas relacionadas con el nivel de efecto colateral.

6 DISCUSIÓN

6.1 Recolección de datos

La recolección es uno de los pasos más importantes en la minería de datos, ya que desde estos se obtendrán conclusiones para una posterior toma de decisiones o desarrollo de modelos de predicción.

Durante este trabajo, la recolección de datos se basó principalmente en el Comité de Acción de Resistencia a Fungicidas (FRAC) y la Base de Datos de Propiedades de Pesticidas (PPDB). Las dos entregan información respecto a los fungicidas utilizados en la actualidad, FRAC entrega clasificaciones respecto al modo de acción con proteínas y el riesgo a resistencia de los fungicidas y PPDB, categoriza los niveles de daños que pueden producir los fungicidas en la salud humana, ambiental y ecológico. Ambas fuentes ya han sido utilizadas para estudios según las áreas en las que se enfocan, resistencia a fungicidas (Miyamoto et al., 2020) y daños al medio ambiente (Möhring et al., 2021).

FRAC y PPDB solo aportaban con la identificación de los fungicidas en nombres, proteínas dianas y categorías mencionadas. Por este motivo, fue necesario recurrir a programas que realizaran cálculos estructurales de las moléculas y complejos proteínas-ligando, en los cuales fueron utilizados Mordred y PLPG3D, respectivamente.

Ambos programas han sido utilizados para estudios en minería de datos, Mordred para el desarrollo de modelos predictivos aplicados a moléculas con el fin de saber si estas atravesarían la barrera hematoencefálica (BHE) (Meng et al., 2021) y PLPG3D en la caracterización de interacciones proteínas-ligandos en organismo mesófilos y termófilos (Bravo Díaz, 2017). Sin embargo, a diferencia de Meng et al., 2021, en este trabajo sí se utilizaron atributos tridimensionales de las moléculas.

Los cálculos en PLPG3D tenían la limitante que las estructuras debían tener complejos formados, es decir, que la estructura cristalizada de la proteína incluyera al ligando. Las proteínas encontradas en PDB no cumplían con este requerimiento, por lo que fue necesario realizar los estudios de acoplamiento molecular.

Los sitios de unión para esta evaluación fueron encontradas en literatura para las cuatro proteínas, RNA Polimerasa I (Engel et al., 2013), Beta Tubulina (Mühlethaler et al., 2021), Succinato Deshidrogenasa (Sun et al., 2005) y el Citocromo BC1 (Esser et al., 2004).

6.2 Selección de atributos

En todas las tareas de clasificación analizadas en este trabajo, la cantidad de atributos estimados era mayor a la del número de ejemplos disponibles. Esto en general produce un problema denominado *maldición de la dimensionalidad*, que ocasiona errores en los cálculos con técnicas en minería de datos, afectando al análisis de estos (Berisha et al., 2021).

Para reducir la dimensión de los sets de datos se recurrió al método denominado “Mutual Information” (Ross, 2014), que permite seleccionar aquellos atributos más relevantes, es decir, aquellos que se diferencian más entre las categorías o clases definidas para cada escenario estudiado. Los atributos seleccionados fueron evaluados por el modelo de clasificación Random Forest, el cual se ha usado anteriormente para procesos de selección de atributos (Ayadanta & Adiwijaya, 2018).

El resultado de este procedimiento redujo notoriamente la cantidad de atributos necesarios para que los distintos modelos de clasificación alcanzaran un buen desempeño. Obteniendo en la mayoría de los escenarios evaluados un nivel de exactitud superior al 80%. De esta manera, se logró superar el problema de la maldición de la dimensionalidad y al mismo tiempo destacar los atributos más relevantes en cada escenario.

6.3 Desempeño de los modelos

En este trabajo se estudiaron diversas tareas de clasificación asociadas a distintos tipos de mecanismos de acción de fungicidas aplicados generalmente en la fruticultura. Se evaluaron modelos para discriminar de manera eficiente entre los distintos tipos de interacciones entre fungicidas y proteínas de patógenos. Adicionalmente, se generaron modelos para inferir sobre efectos colaterales del uso de fungicidas, en relación con niveles de riesgo a resistencia, ecotoxicidad y el grado de daño al medio ambiente y salud humana.

Se escogió utilizar el método de Random Forest como modelo de clasificación para todas las tareas definidas. Esto debido principalmente a la versatilidad que este algoritmo entrega para trabajar con problemas de alta dimensionalidad y con un número reducido de ejemplos. En general, este tipo de algoritmos entrena modelos predictivos que no sobreajustan al set de entrenamiento utilizado.

Las tareas de clasificación definidas presentaban en general un problema de desbalance de clases. Para enfrentar esta situación se utilizó el método llamado SMOTE, que frecuentemente se utiliza en estudios de minería de datos para equilibrar el número de ejemplos de cada clase (Fernández et al., 2018).

Los modelos de clasificación generados en este estudio presentan altas tasas de exactitud, de sobre el 80% para todos los escenarios considerados en cuanto a interacciones entre fungicidas y proteínas de patógenos, y de sobre un 70% de exactitud para escenarios asociados a efectos colaterales de fungicidas. Así mismo, los problemas de alta dimensionalidad y desbalance de clases son eficientemente abordados, al obtener medidas equilibradas de sensibilidad para las distintas clases.

Esta buena performance obtenida demostraría que el grupo de características y atributos seleccionados para entrenar los distintos modelos, son relevantes y permiten caracterizar los mecanismos de unión de fungicidas a proteínas y también explicarían los efectos colaterales estudiados.

Estos resultados abren el camino para encontrar nuevas moléculas que tengan potencial como fungicida. Estos modelos generados podrían ser evaluados contra una base de datos de moléculas conocidas, y encontrar un grupo de nuevas moléculas candidatas a ser evaluadas como potenciales fungicidas. Sin embargo, un mayor análisis y estudios específicos deben ser conducidos en el futuro para demostrar experimentalmente estos descubrimientos.

7 CONCLUSIÓN

En este estudio se exploraron diversas tareas relacionadas al mecanismo de acción de fungicidas generalmente utilizados en fruticultura y sus efectos colaterales, implementando herramientas para la selección de atributos diferenciadores y técnicas de minería de datos.

Por un lado, se analizó información asociada a los mecanismos de interacción entre moléculas fungicidas y proteínas blanco ya conocidas. Específicamente se estudiaron 4 tareas de clasificación asociadas a las interacciones con las proteínas RNA Polimerasa I, Beta Tubulina, Succinato Deshidrogenasa y Citocromo BC1. Se estimaron atributos en base a las propiedades estructurales, fisicoquímicas y geométricas de las zonas de unión a ligando con el software PLPG3D.

Por otro lado, se estudiaron efectos colaterales provocados por el uso de fungicidas, considerando 4 nuevas tareas de clasificación asociadas con el riesgo a resistencia, ecotoxicidad, efectos al medioambiente y efectos a la salud humana. Para este caso se estimaron atributos asociados a las propiedades estructurales de los fungicidas con el software Mordred.

Se implementaron y evaluaron modelos de clasificación basados en el algoritmo de Random Forest para cada una de las tareas definidas. En cada uno de los casos evaluados se logró identificar un grupo acotado de atributos y características que permitían generar modelos de clasificación eficientes.

El método de selección de atributos desarrollado permitió evaluar y generar un ranking de relevancia de aquellos atributos que permitían diferenciar/discriminar entre las distintas categorías o clases definidas en cada caso. De lo anterior, se concluye que los softwares PLPG3D y Mordred son herramientas útiles para la caracterización de estructuras moleculares y de las zonas de unión en los complejos proteína-ligando. Junto con esto, se confirma la hipótesis planteada que con estos tipos de datos se logra diferenciar entre las categorías estudiadas, permitiendo el futuro análisis para la selección y desarrollo de nuevos fungicidas.

8 ANEXOS

8.1 Anexos tarea de interacción a proteína

Se dan a conocer los atributos más relevantes para las tareas relacionadas con interacción a proteína y sus valores representados con boxplots. Estos fueron considerados diferenciadores por el método de selección de atributo Mutual-Information.

Tabla 8.1: Se dan a conocer los atributos más relevantes para la tarea de interacción con RNA Polimerasa I.

Atributos	Tipos de Atributos
Solvation Polar Total Energy Solvation Hydrophobic backbone clash Sidechain Hbond entropy main chain energia_de_captación c7	Energía
II-II	Interacciones
Distancia CB-A-C4 Distancia_CB-A-C2	Distancia
Omega_aa3 Psi_aa2 Torsion_aa3	Torsiones

Figura 8.1: Boxplots de los atributos más relevantes para la tarea de interacción con RNA Polimerasa I. POL1 representa la clase positiva, es decir, a los fungicidas que se sabe interactúan con la proteína. NOTPOL1 representa a la clase negativa. (Parte 1/2)

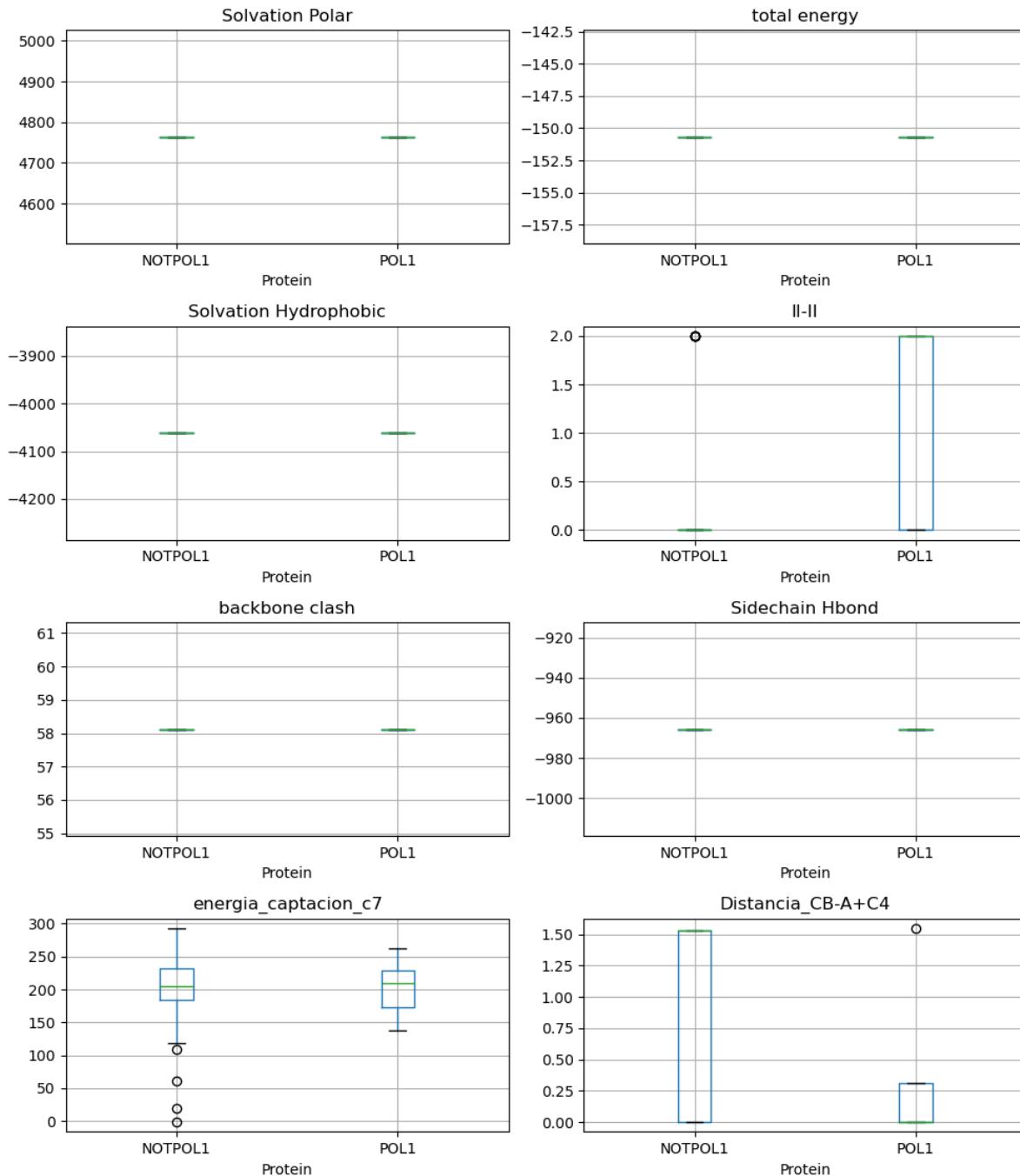


Figura 8.1: Boxplots de los atributos más relevantes para la tarea de interacción con RNA Polimerasa I. POL1 representa la clase positiva, es decir, a los fungicidas que se sabe interactúan con la proteína. NOTPOL1 representa a la clase negativa. (Parte 2/2)

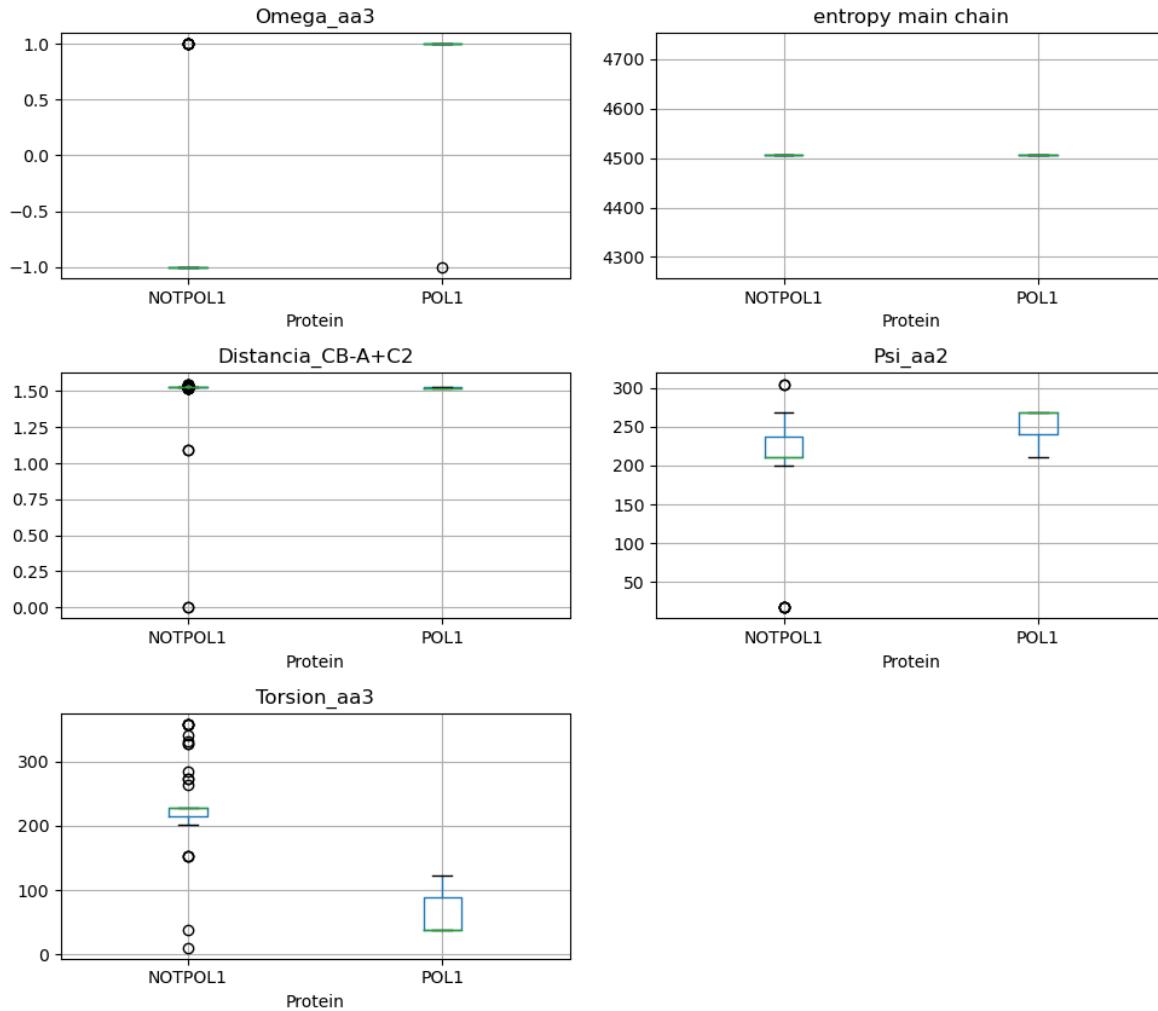


Tabla 8.2: Se dan a conocer los atributos más relevantes para la tarea de interacción con Beta Tubulina.

Atributos	Tipos de Atributos
Phi_aa1 Psi_aa1 Omega_aa1 Psi_aa2	Torsiones
%S_c7 %S_c3	Composición atómica
energia_captacion_c6 energia_captacion_c4 energia_captacion_c3 torsional clash	Energía
I-IV	Interacciones
Polares_positivos	Composición aminoacídica
volumen_c1	Volumen

Figura 8.2: Boxplots de los atributos más relevantes para la tarea de interacción con Beta Tubulina. BT representa la clase positiva, es decir, a los fungicidas que se sabe interactúan con la proteína. NOTBT representa a la clase negativa. (Parte 1/2)

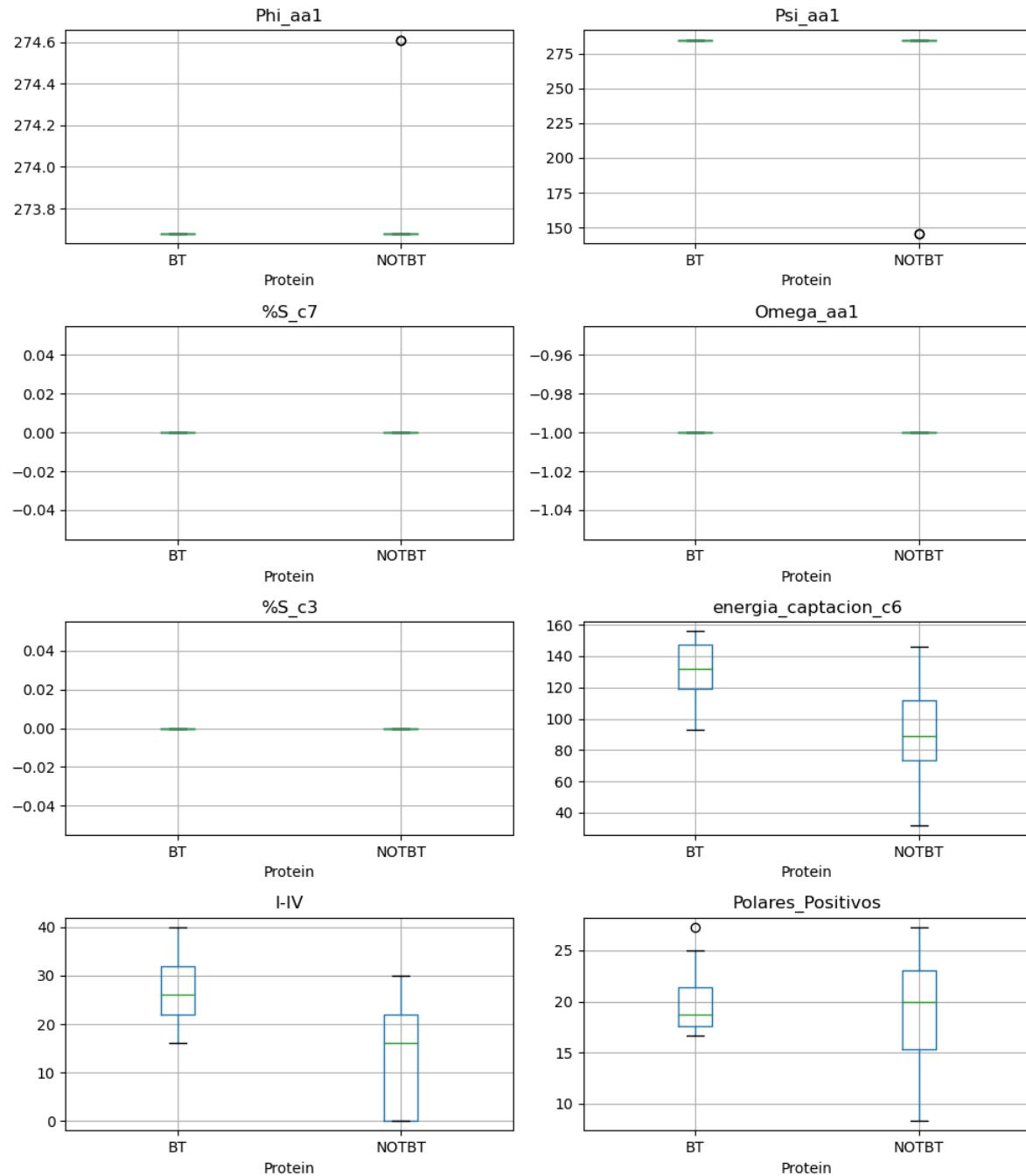


Figura 8.2 Boxplots de los atributos más relevantes para la tarea de interacción con Beta Tubulina. BT representa la clase positiva, es decir, a los fungicidas que se sabe interactúan con la proteína. NOTBT representa a la clase negativa. (Parte 2/2)

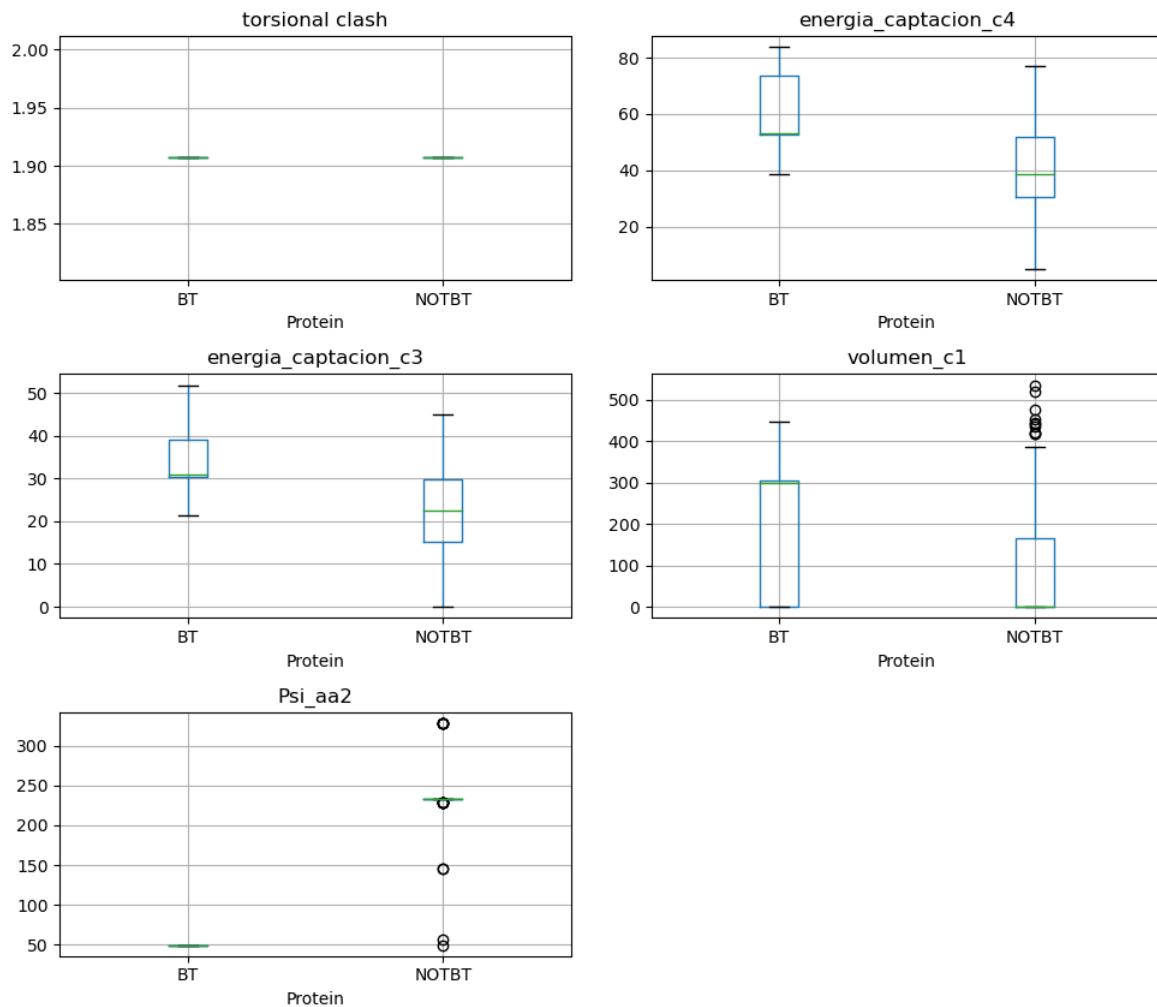


Tabla 8.3: Se dan a conocer los atributos más relevantes para la tarea de interacción con Succinato Deshidrogenasa.

Atributos	Tipos de Atributos
Psi_aa2 Phi_aa3 Omega_aa2	Torsiones
Energia_captacion_c1 Energia_captacion_c2 Energia_captacion_c4 Energia_captacion_c5	Energía
volumen_c6	Volumen
%N_c5 %O_c3 Atomos_c2	Composición atómica
Distancia_AC1-AC4 Angulo_CB-CA3	Distancia

Figura 8.3: Boxplots de los atributos más relevantes para la tarea de interacción con Succinato Deshidrogenasa. SD representa la clase positiva, es decir, a los fungicidas que se sabe interactúan con la proteína. NOTSD representa a la clase negativa. (1/2)

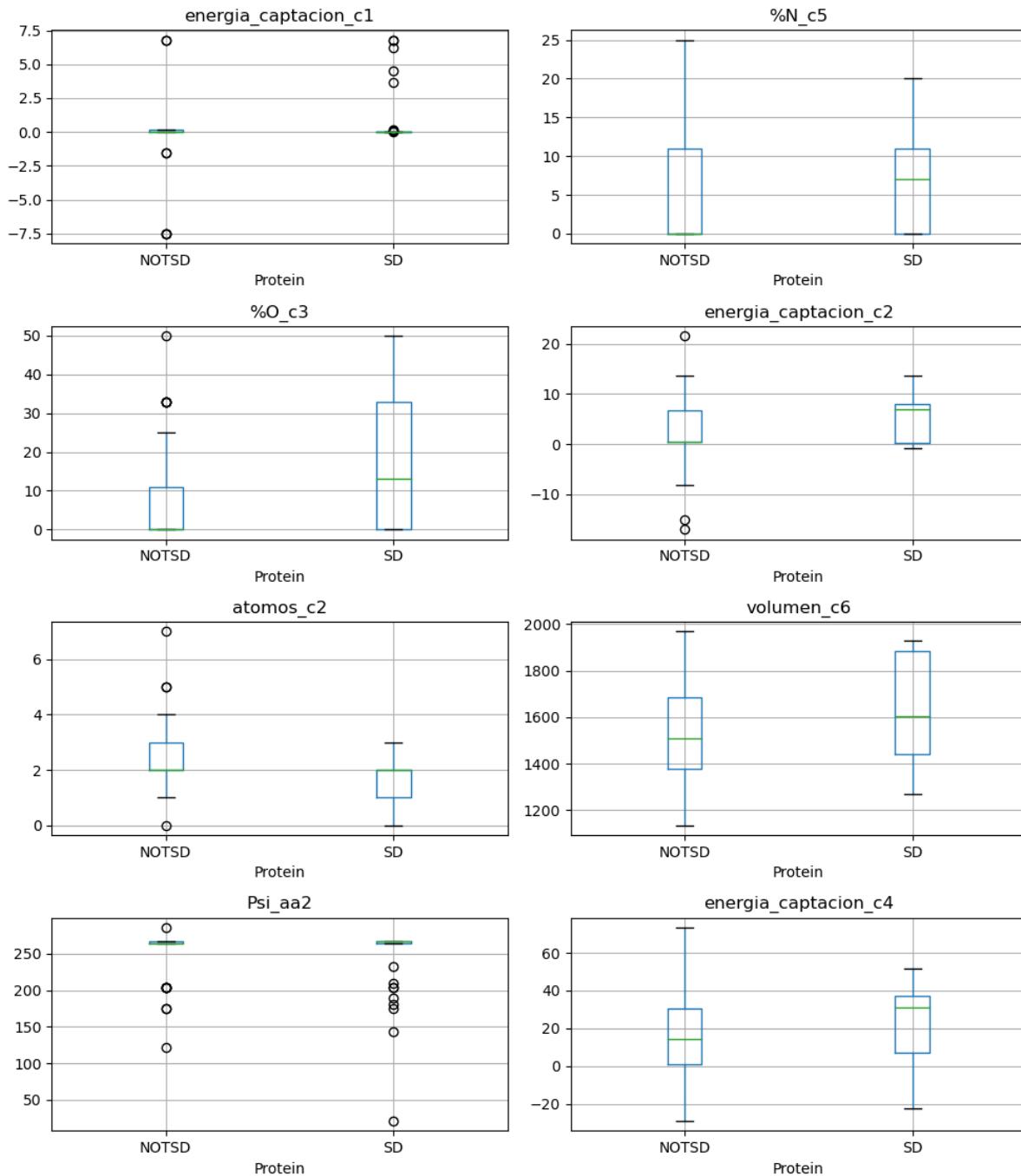


Figura 8.3: Boxplots de los atributos más relevantes para la tarea de interacción con Succinato Deshidrogenasa. SD representa la clase positiva, es decir, a los fungicidas que se sabe interactúan con la proteína. NOTSD representa a la clase negativa. (2/2)

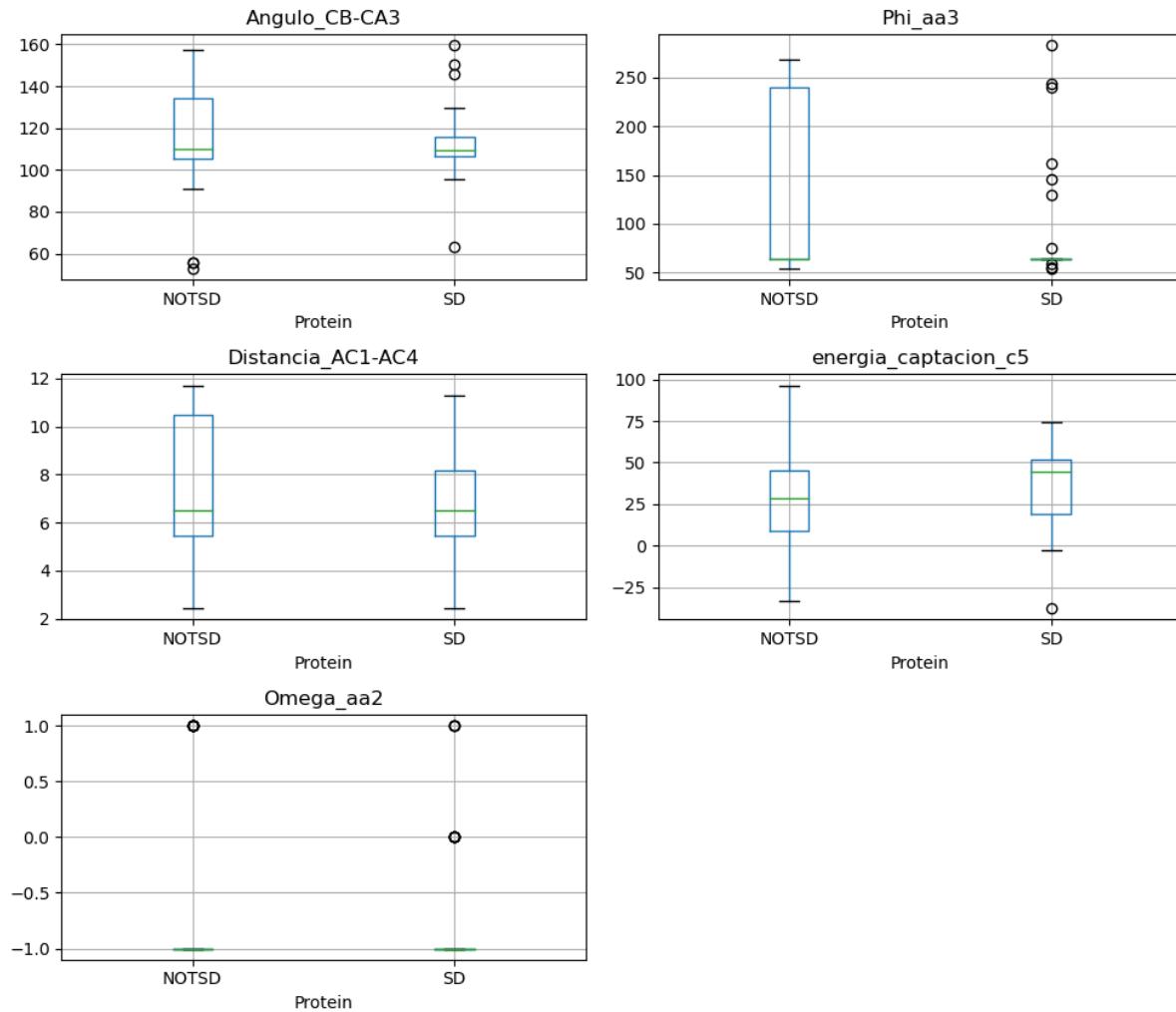


Tabla 8.4: Se dan a conocer los atributos más relevantes para la tarea de interacción con Citocromo BC1.

Atributos	Tipos de Atributos
volumen_c3	Volumen
volumen_c2	
volumen_c1	
Energia_captacion_c3	Energía
Smallest Alpha	Alpha Shapes
Distancia_CA1-CA4	Distancias
Distancia_AC3-AC4	
Angulo_AC2-AC4	Ángulos
Angulo_CA2-CA4	
Angulo_CB-CA3	
Angulo_CB-AC3	
IV-IV	Interacciones

Figura 8.4: Boxplots de los atributos más relevantes para la tarea de interacción con Citocromo BC1. BC1 representa la clase positiva, es decir, a los fungicidas que se sabe interactúan con la proteína. NOTBC1 representa a la clase negativa. (Parte 1/2)

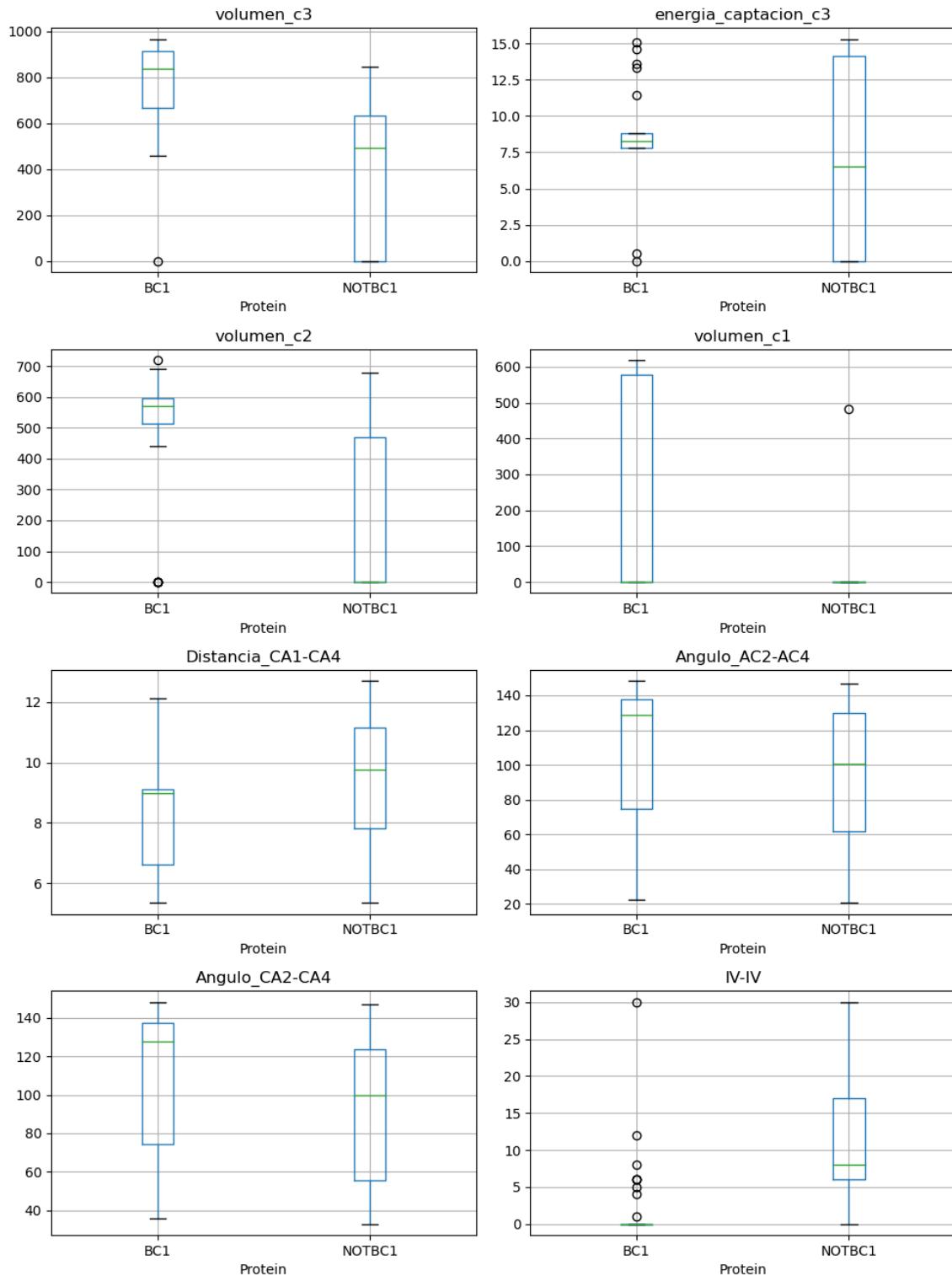
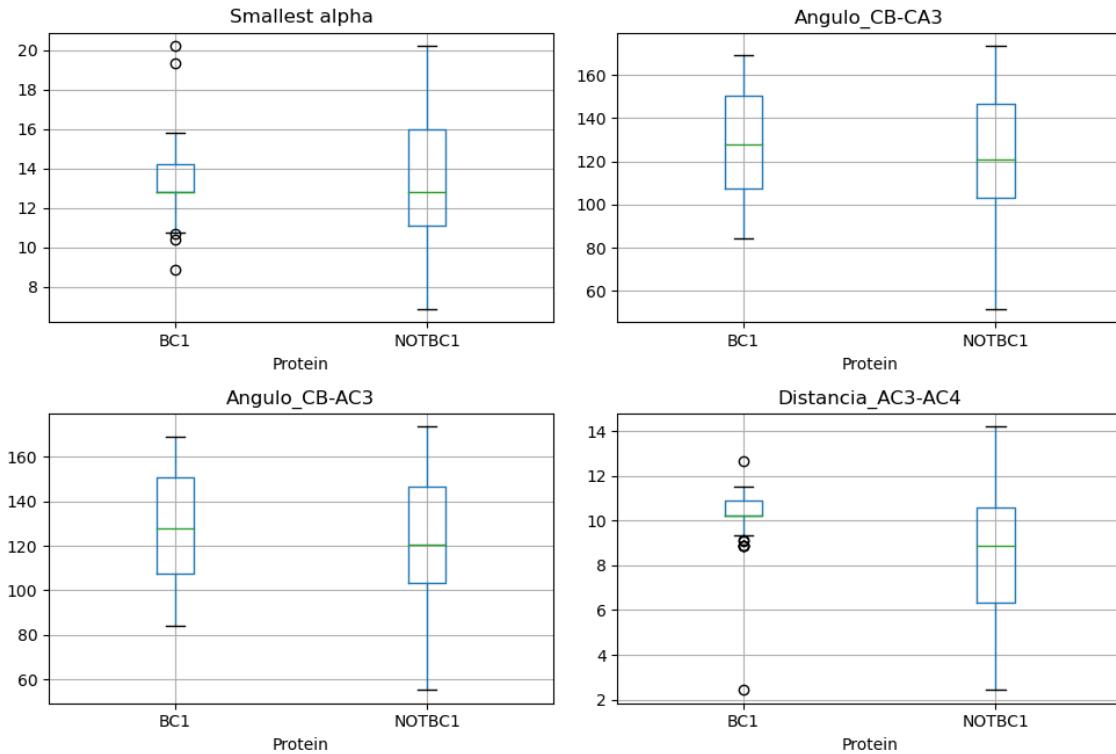


Figura 8.4: Boxplots de los atributos más relevantes para la tarea de interacción con Citocromo BC1. BC1 representa la clase positiva, es decir, a los fungicidas que se sabe interactúan con la proteína. NOTBC1 representa a la clase negativa. (Parte 2/2)



8.2 Anexos tarea de nivel de efecto colateral

Se dan a conocer los atributos más relevantes para las tareas relacionadas con el nivel de efectos colaterales y sus valores representados con boxplots. Estos fueron considerados diferenciadores por el método de selección de atributo Mutual-Information.

Tabla 8.5: Se dan a conocer los atributos más relevantes para la tarea de diferenciar el nivel de riesgo a resistencia de los fungicidas.

Atributo	Tipo de atributo
BCUTd-1h BCUTp-1l	BCUT
VE2_Dt SM1_Dt	DetourMatrix
VE2_Dzi SM1_Dzm VE2_Dzpe SM1_DzZ SpMax_DzZ SpMax_Dzm VE2_Dzse SM1_Dzi VE1_Dzi VE2_DzZ VE2_Dzm VE1_Dzpe VE1_DzZ VE1_Dzm	BaryszMatrix
AMW	WalkCount
VE2_D LogEE_D VE1_D	DistanceMatrix
AETA_eta_R	ExtendedTopochemicalAtom
Xp-0d	Chi

Figura 8.5: Boxplots de los atributos más relevantes para la tarea de diferenciar los niveles de riesgo a resistencia de fungicidas. (Parte 1/3)

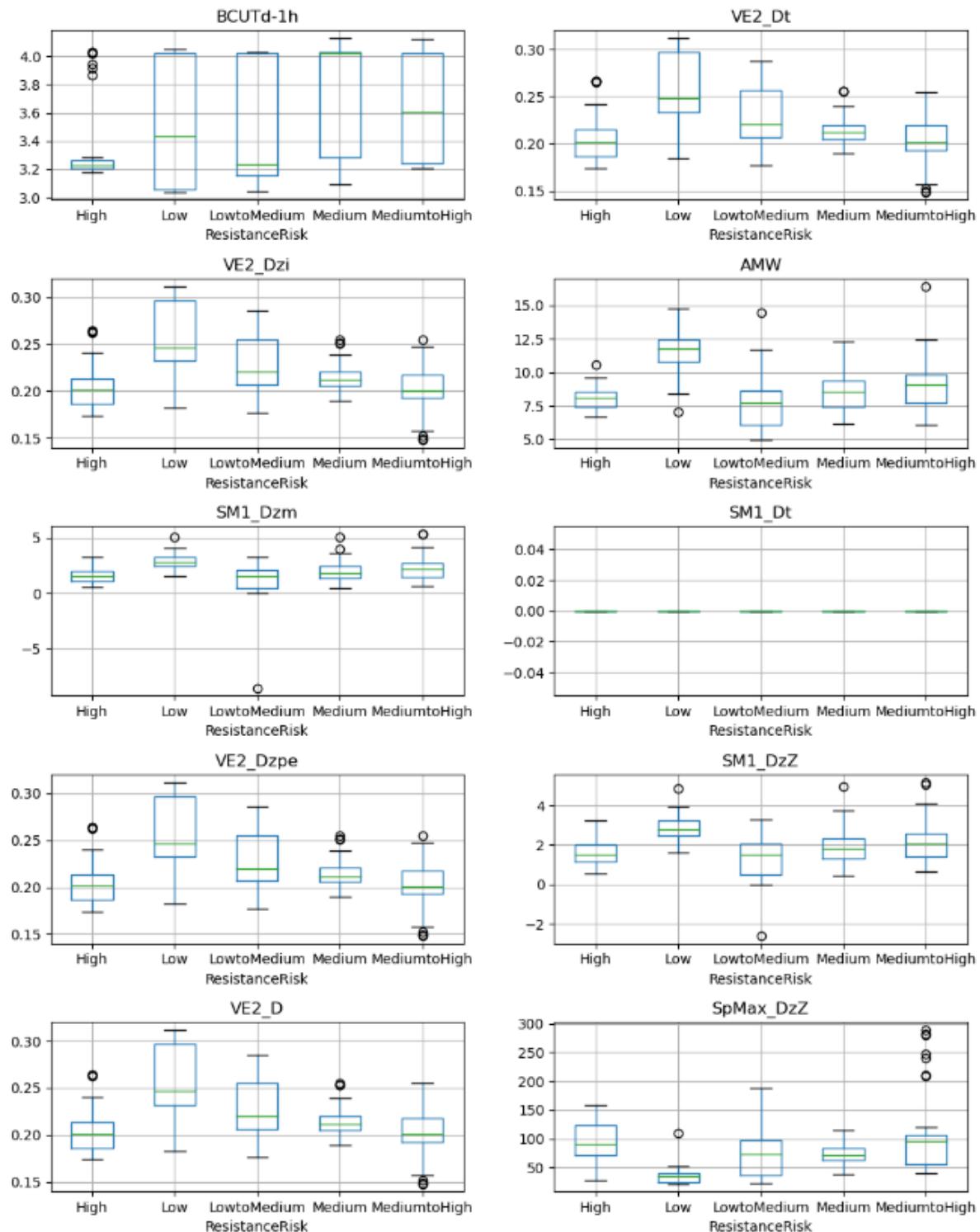


Figura 8.5: Boxplots de los atributos más relevantes para la tarea de diferenciar los niveles de riesgo a resistencia de fungicidas. (Parte 2/3)

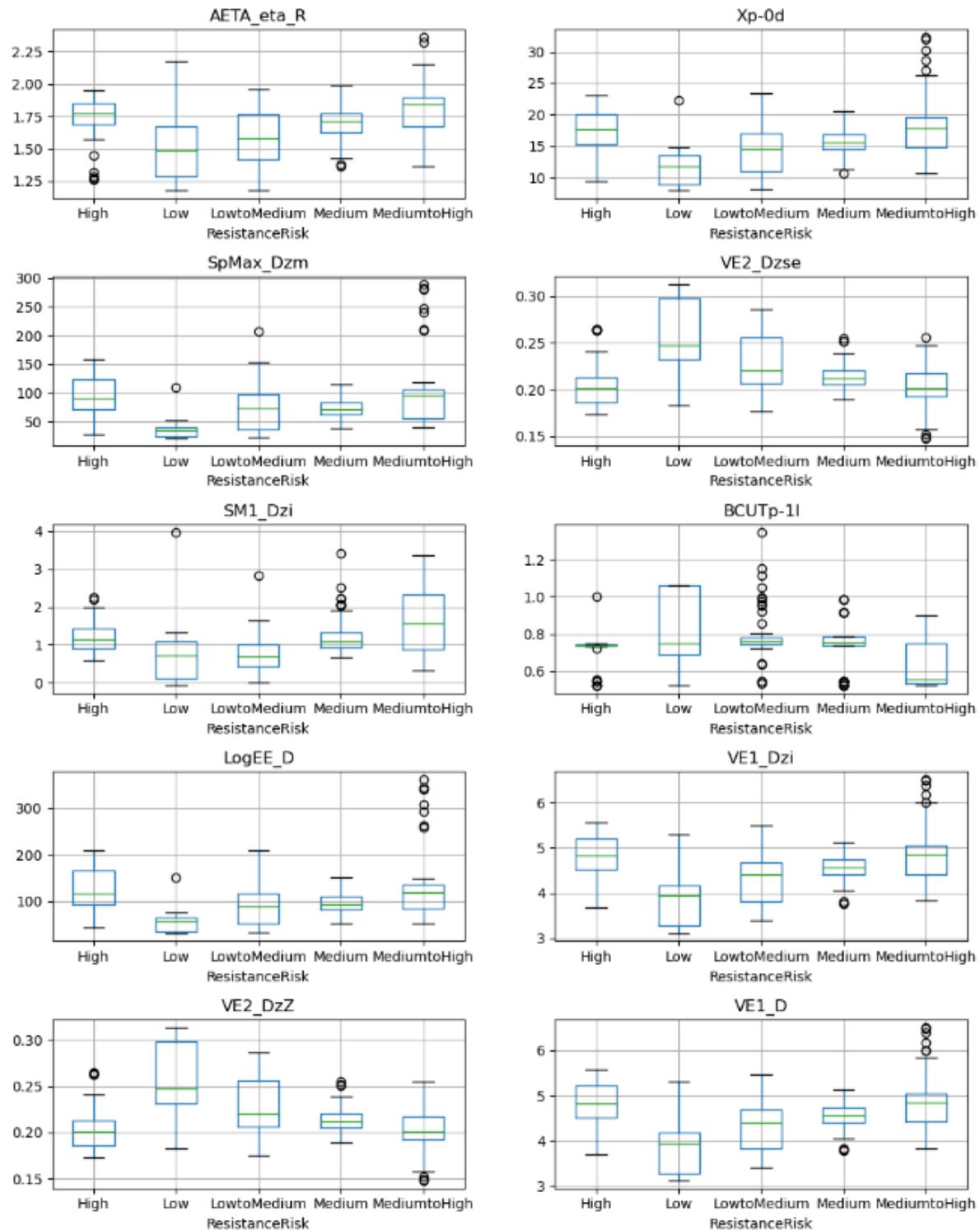


Figura 8.5: Boxplots de los atributos más relevantes para la tarea de diferenciar los niveles de riesgo a resistencia de fungicidas. (Parte 3/3)

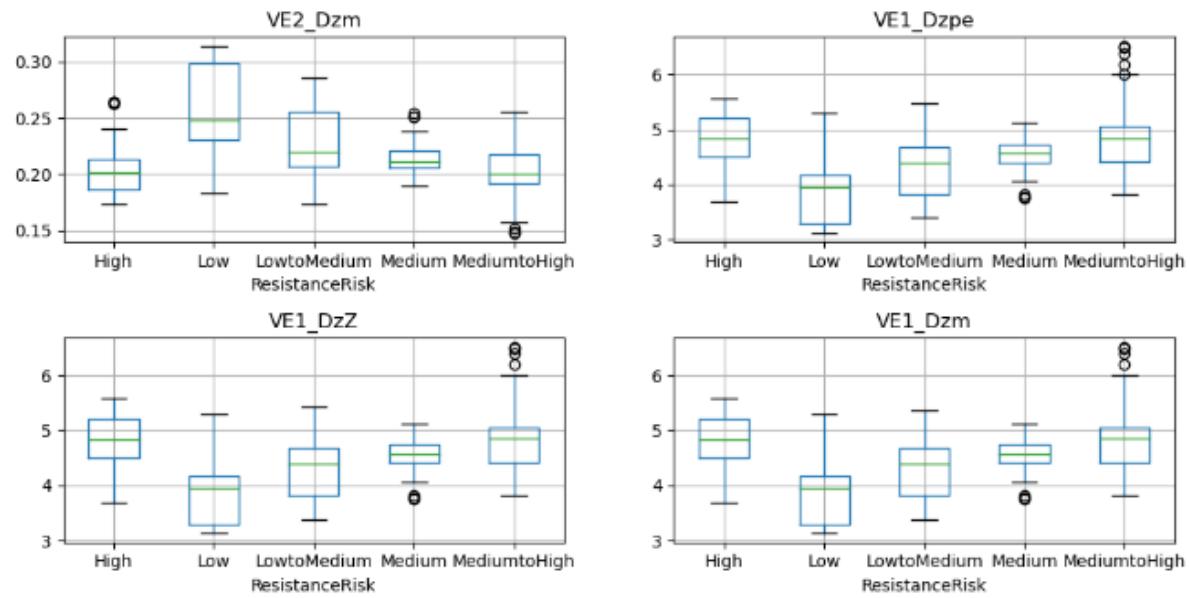


Tabla 8.6: Se dan a conocer los atributos más relevantes para la tarea de diferenciar el nivel ecotóxico de los fungicidas.

Atributo	Tipo de atributo
SM1_DzZ SM1_Dzm VE2_DzZ VE2_Dzm	BaryszMatrix
VSA_EState8 PEOE_VSA1	MoeType
MATS4i ATS6d ATSC4dv MATS2se AATSC6m AATS3v MATS5p ATSC7se MATS6p MATS2are	AutoCorrelation
piPC5 piPC2 piPC7	PathCount
Xp-6d Xp-3d	Chi
Mor25se Mor11v	MoRSE
MWC06	WalkCount

Figura 8.6: Boxplots de los atributos más relevantes para la tarea de diferenciar los niveles de ecotoxicidad de los fungicidas. (Parte 1/3)

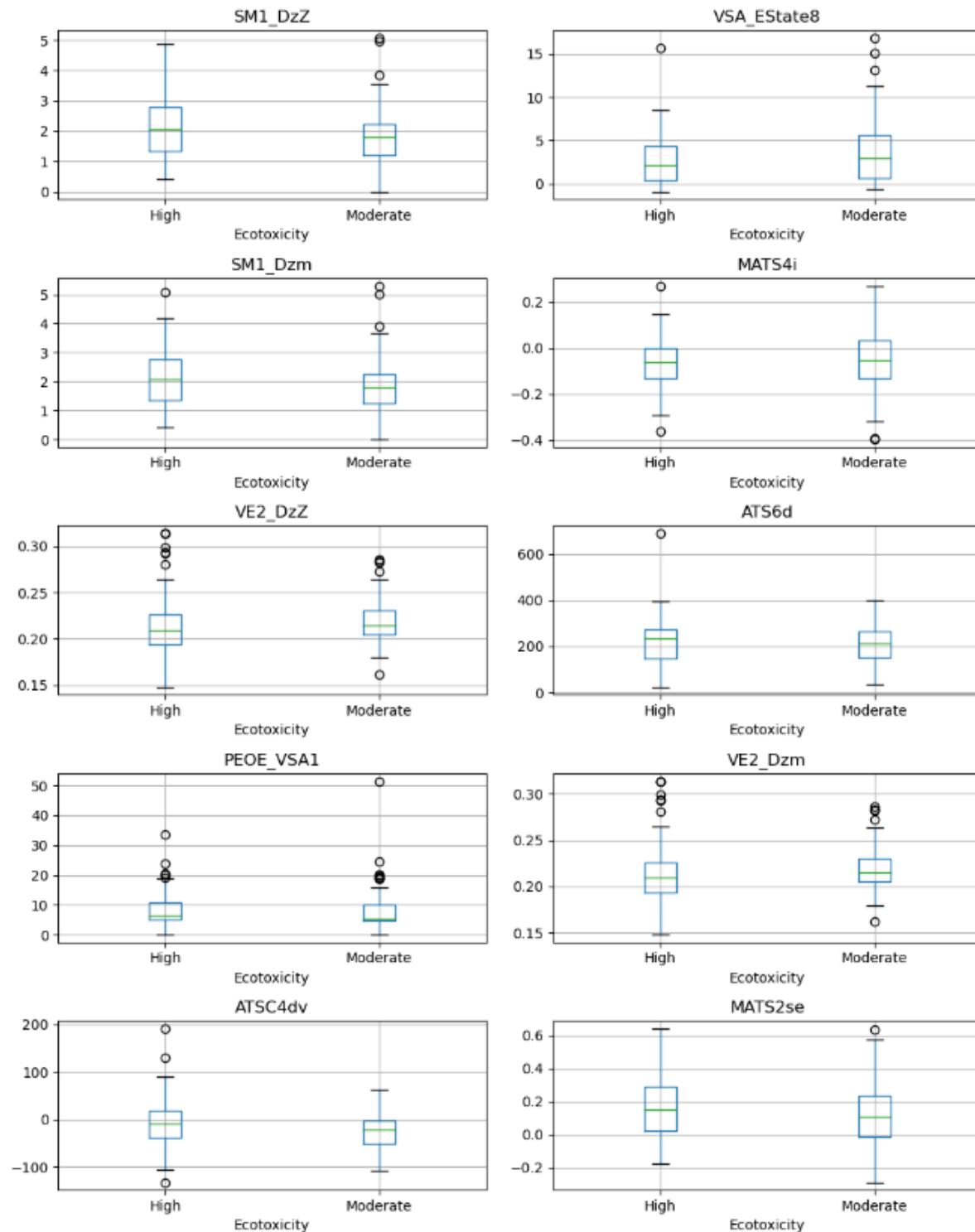


Figura 8.6: Boxplots de los atributos más relevantes para la tarea de diferenciar los niveles de ecotoxicidad de los fungicidas. (Parte 2/3)

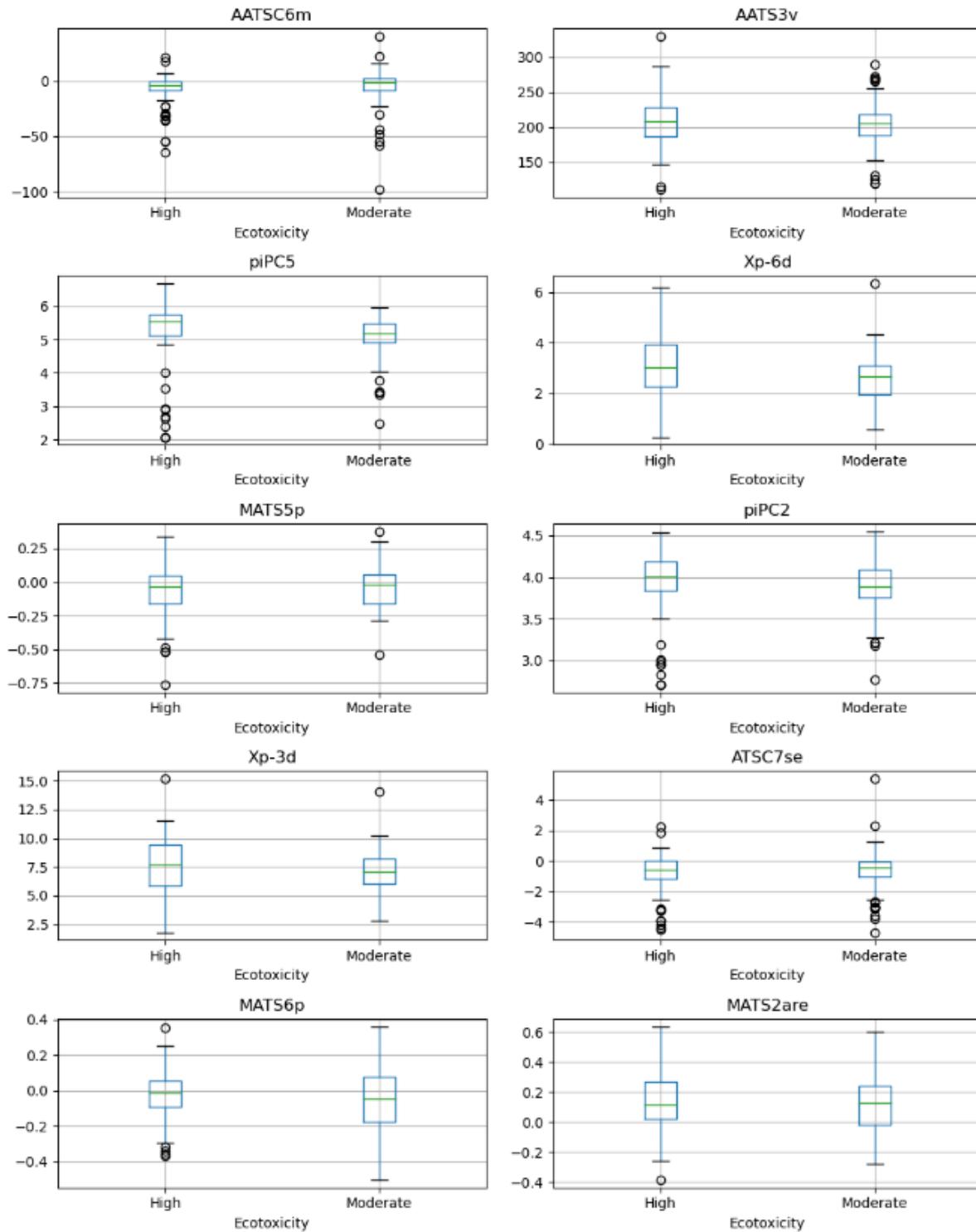


Figura 8.6: Boxplots de los atributos más relevantes para la tarea de diferenciar los niveles de ecotoxicidad de los fungicidas. (Parte 3/3)

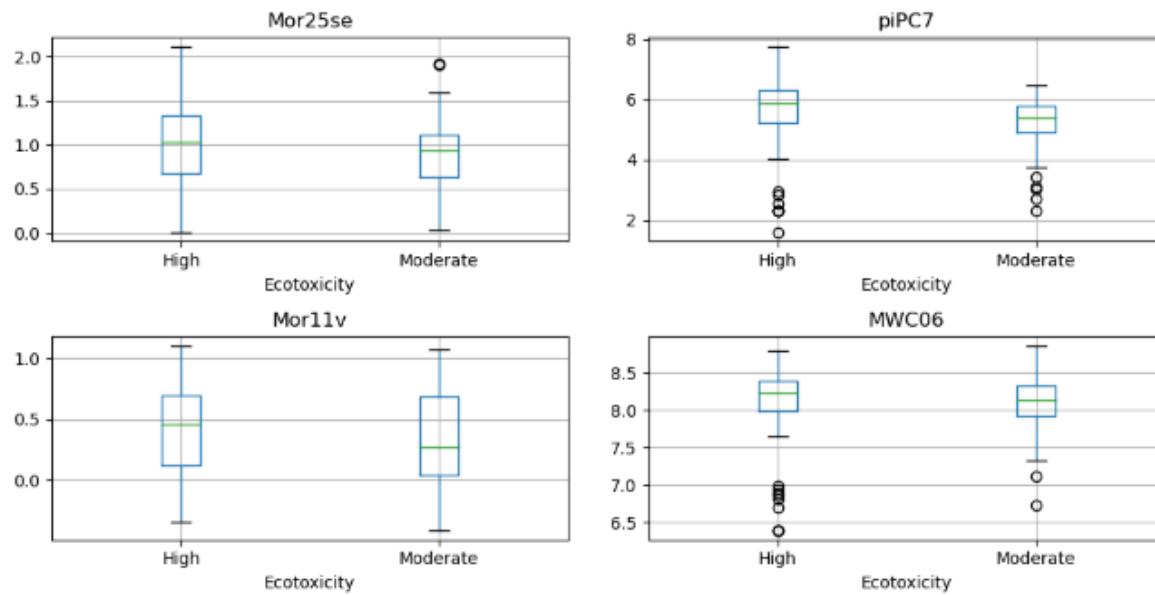


Tabla 8.7: Se dan a conocer los atributos más relevantes para la tarea de diferenciar el grado de daño medioambiental de los fungicidas.

Atributo	Tipo de atributo
fMF	Framework
GATS1s GATS7dv AATS7i ATS6dv AATS3m ATS5d ATS1s ATS5i GATS5d	AutoCorrelation
MID_O	MolecularId
SddsN SaaO SdS	EState
PEOE_VSA2	MoeType
SpAD_A	AdjacencyMatrix
LogEE_Dzm SpMax_Dzp	BaryszMatrix
ZMIC4	InformationContent
TopoPSA(NO)	TopoPSA
Xpc-6dv	Chi

Figura 8.7: Boxplots de los atributos más relevantes para la tarea de diferenciar el grado de daño medioambiental de los fungicidas. (Parte 1/3)

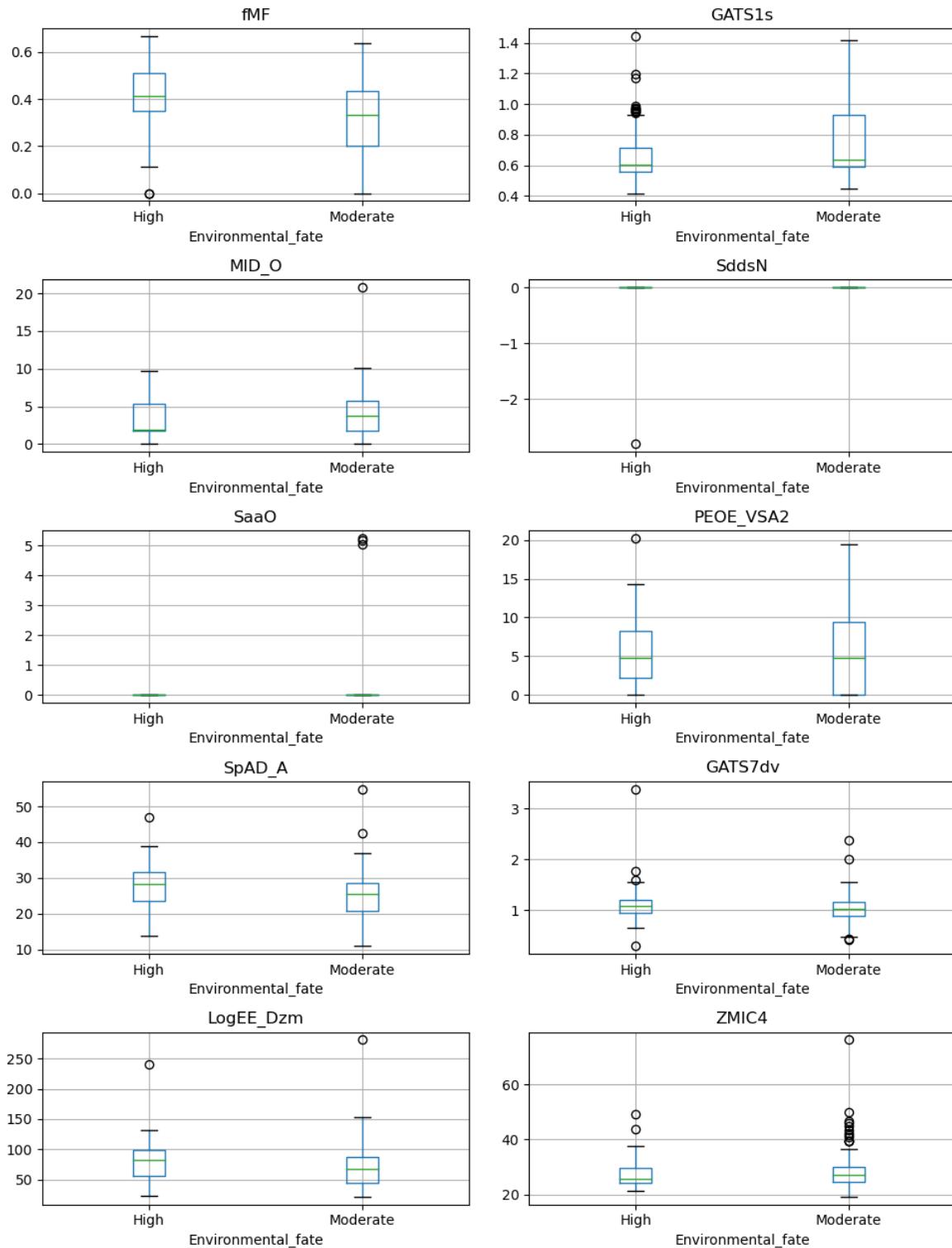


Figura 8.7: Boxplots de los atributos más relevantes para la tarea de diferenciar el grado de daño medioambiental de los fungicidas. (Parte 2/3)

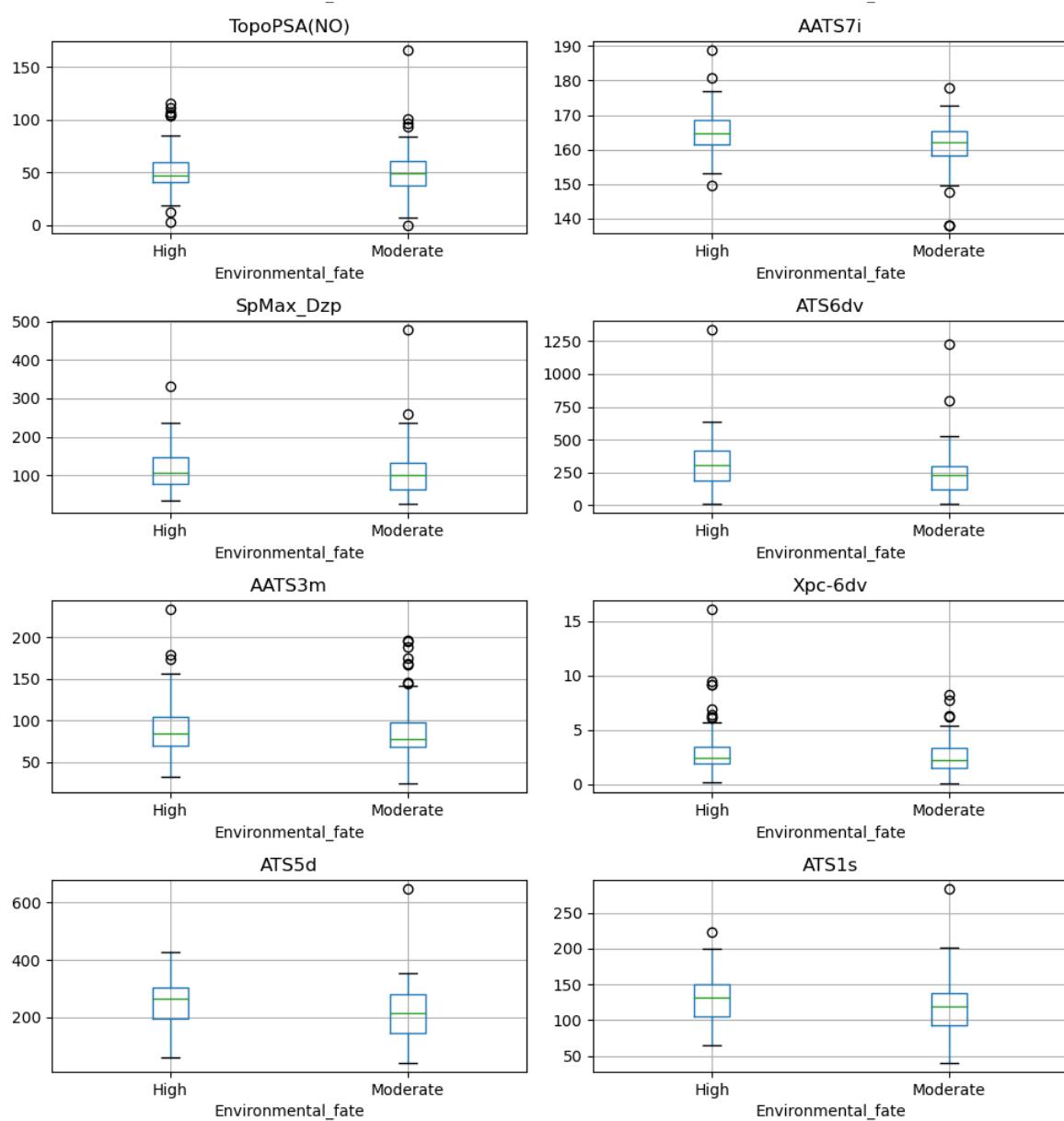


Figura 8.7: Boxplots de los atributos más relevantes para la tarea de diferenciar el grado de daño medioambiental de los fungicidas. (Parte 3/3)

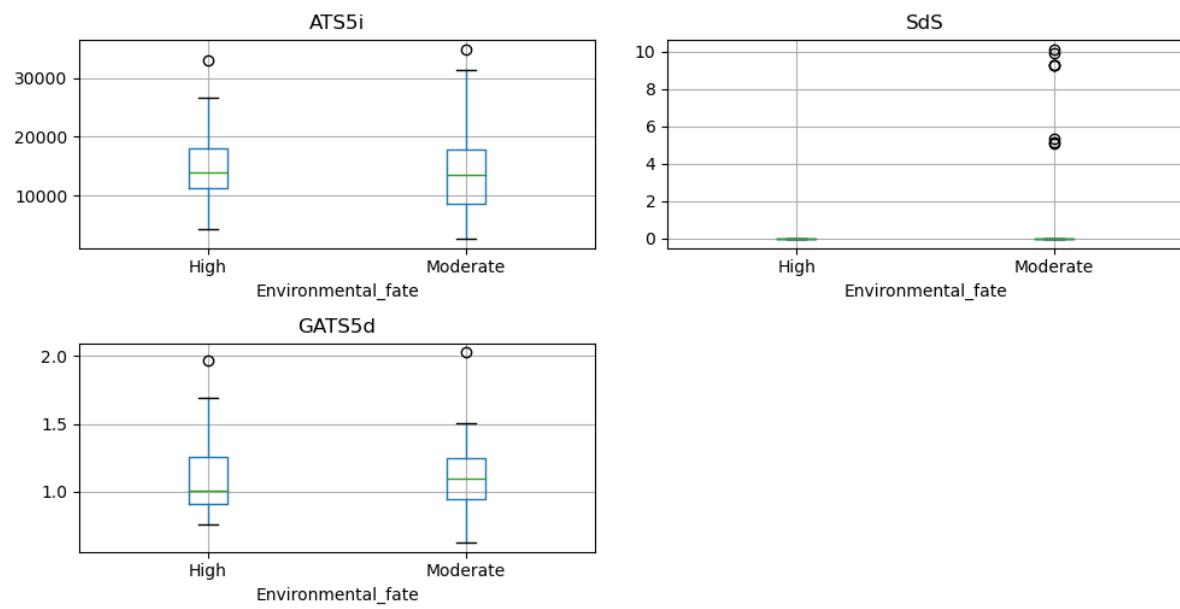


Tabla 8.8: Se dan a conocer los atributos más relevantes para la tarea de diferenciar el grado de daño en salud humana de los fungicidas.

Atributo	Tipo de atributo
MATS5p ATSC4c AATSC4c AATS3v ATSC6m GATS8m AATSC4Z ATSC3Z AATSC5c AATS7se ATS2pe MATS4pe MATS2Z	Autocorrelation
VR1_Dt	DetourMatrix
SlogP_VSA6	MoeType
BCUTZ-1h	BCUT
SM1_Dzv LogEE_Dzi SpDiam_Dzv SpMax_Dzare	BaryszMatrix
Xpc-5dv	Chi
SpAD_D	DistanceMatrix
Mor22m	MoRSE
PetitjeanIndex	TopologicalIndex
GGI3	TopologicalCharge

Figura 8.8: Boxplots de los atributos más relevantes para la tarea de diferenciar el grado de daño en salud humana de los fungicidas. (Parte 1/3)

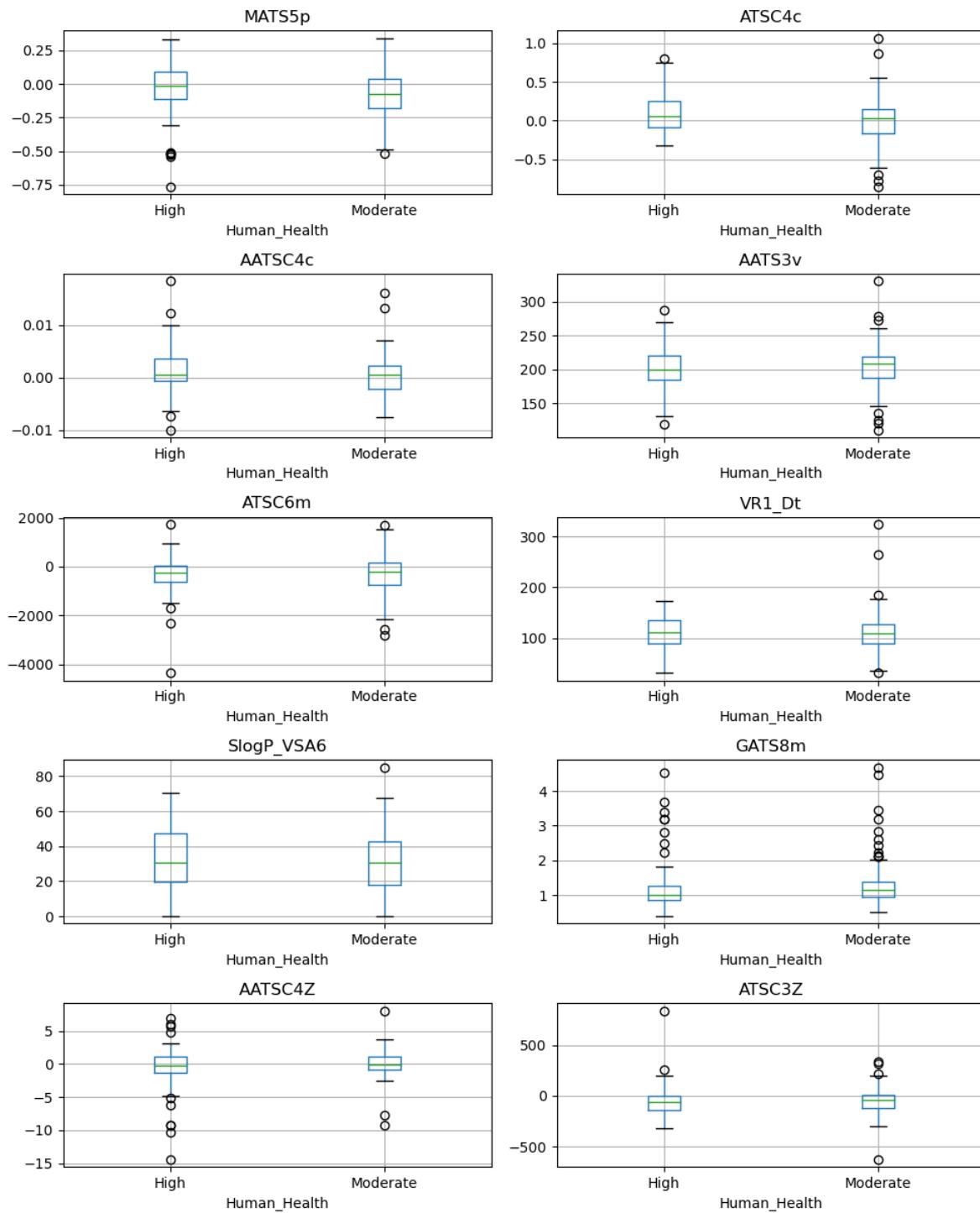


Figura 8.8: Boxplots de los atributos más relevantes para la tarea de diferenciar el grado de daño en salud humana de los fungicidas. (Parte 2/3)

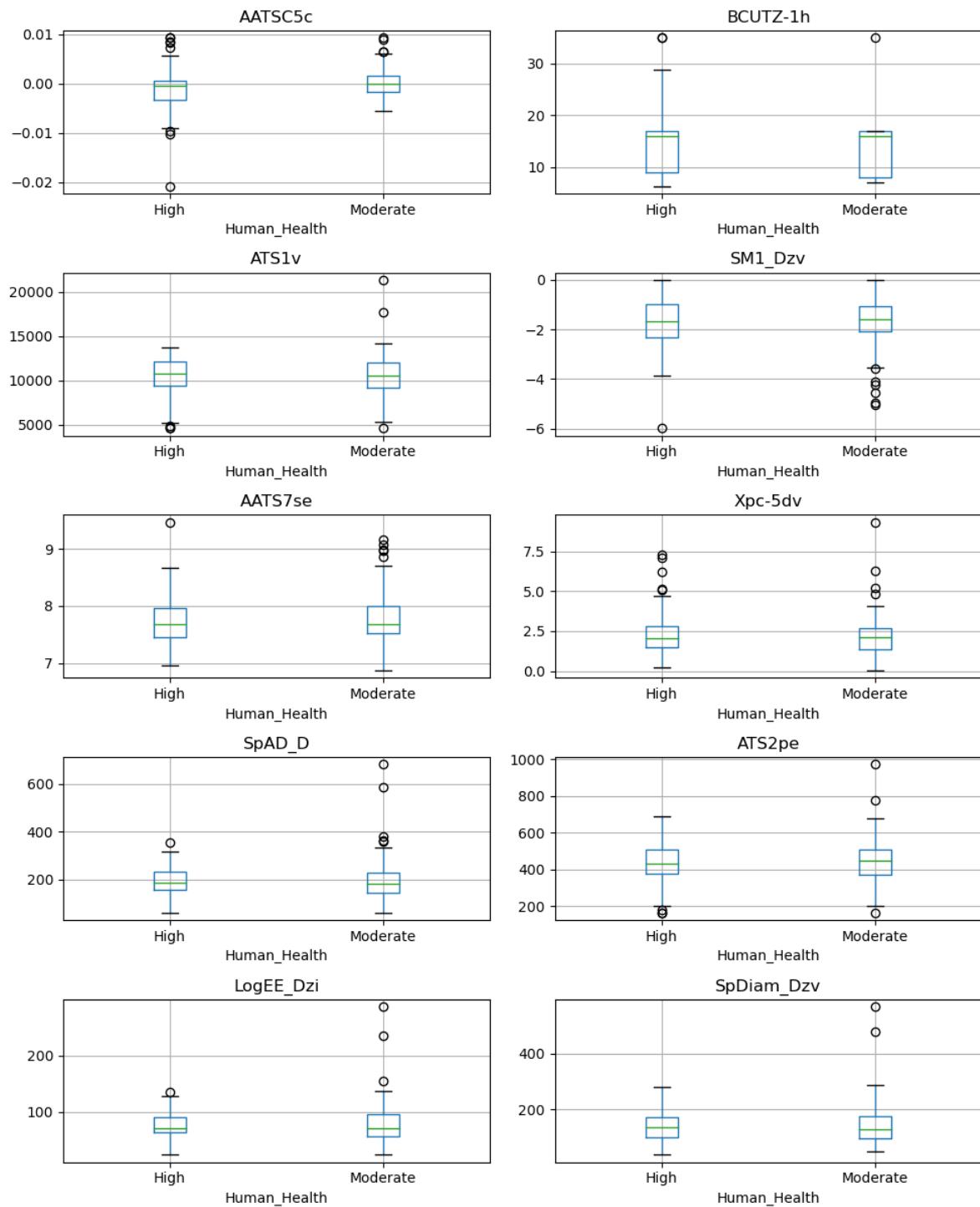
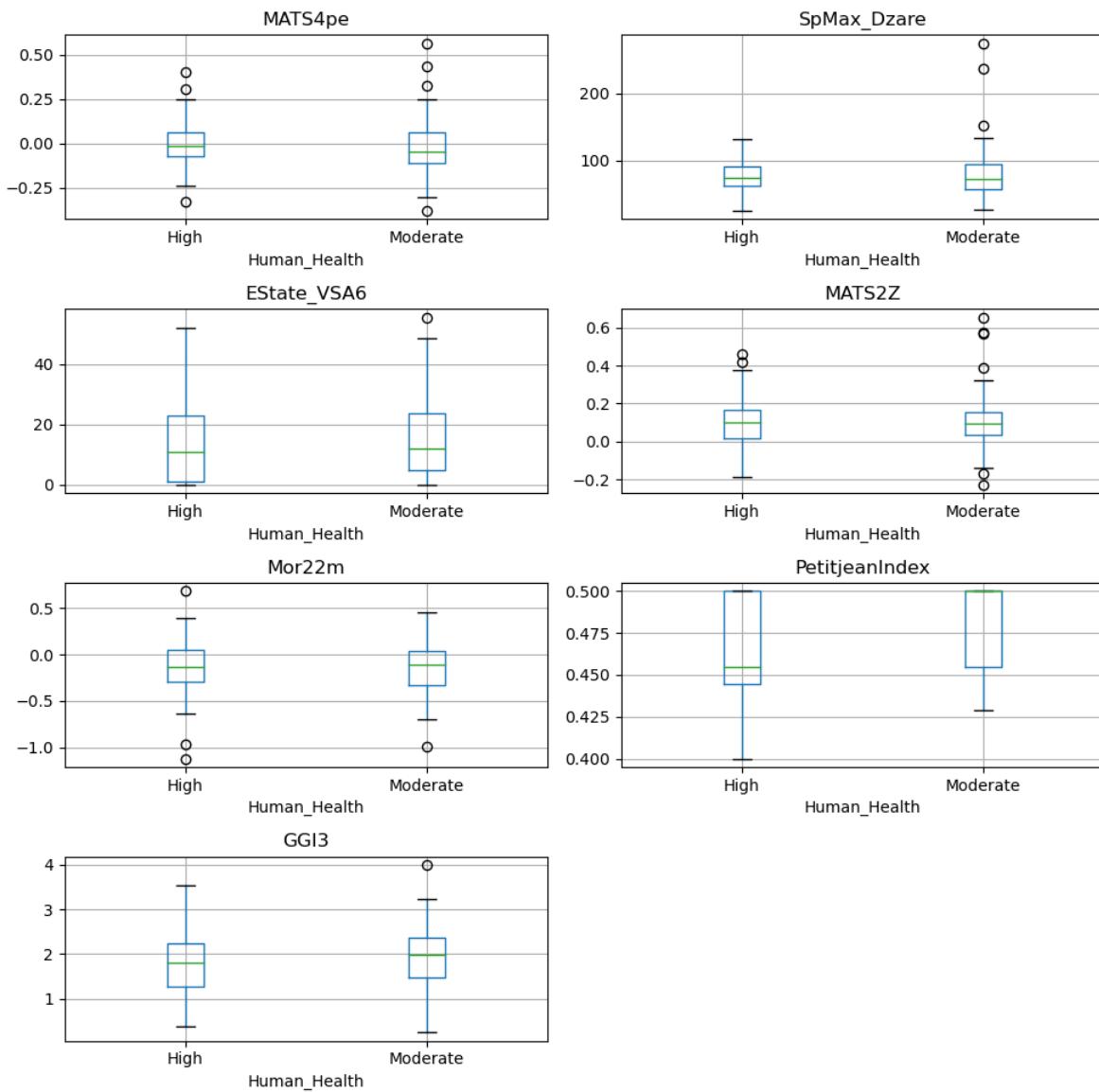


Figura 8.8: Boxplots de los atributos más relevantes para la tarea de diferenciar el grado de daño en salud humana de los fungicidas. (Parte 3/3)



REFERENCIAS

- Apey Guzmán, A. (2019). *La fruticultura en Chile: tendencias productivas y su expresión territorial.* 23. https://sgs.odepa.gob.cl/bitstream/handle/20.500.12650/2613/Artículo-Fruticultura_mayo-1.pdf?sequence=1&isAllowed=y
- Attwood, T. K., Blackford, S., Brazas, M. D., Davies, A., & Schneider, M. V. (2019). A global perspective on evolving bioinformatics and data science training needs. *Briefings in Bioinformatics*, 20(2), 398–404. <https://doi.org/10.1093/bib/bbx100>
- Auger, J., Pozo, L., Rubilar, M., Briceño, N., Osorio-Navarro, C., & Esterio, M. (2021). First report of canker and branch dieback of sweet cherry trees caused by calosphaeria pulchella in Chile. *Plant Disease*, 105(1), 217. <https://doi.org/10.1094/PDIS-05-20-1026-PDN>
- Azeroual, O., Saake, G., & Abuosba, M. (2019). Data quality measures and data cleansing for research information systems. ArXiv, 16(1).
- Baibakova, E. V., Nefedjeva, E. E., Suska-Malawska, M., Wilk, M., Sevriukova, G. A., & Zheltobriukhov, V. F. (2019). Modern Fungicides: Mechanisms of Action, Fungal Resistance and Phytotoxic Effects. *Annual Research & Review in Biology*, 32(3), 1–16. <https://doi.org/10.9734/arrb/2019/v32i330083>
- Bateman, A., Martin, M. J., O'Donovan, C., Magrane, M., Alpi, E., Antunes, R., Bely, B., Bingley, M., Bonilla, C., Britto, R., Bursteinas, B., Bye-AJee, H., Cowley, A., Da Silva, A., De Giorgi, M., Dogan, T., Fazzini, F., Castro, L. G., Figueira, L., ... Zhang, J. (2017). UniProt: The universal protein knowledgebase. *Nucleic Acids Research*, 45(D1), D158–D169. <https://doi.org/10.1093/nar/gkw1099>
- Bertetti, D., Monchiero, M., Garibaldi, A., & Gullino, M. L. (2020). Monitoring activities on fungicide resistance in Botrytis cinerea carried out in vineyards in North-West Italy in 2018. *Journal of Plant Diseases and Protection*, 127(1), 123–127. <https://doi.org/10.1007/s41348-019-00263-3>

Bhachoo, J., & Beuming, T. (2017). Investigating protein-peptide interactions using the Schrödinger computational suite. *Methods in Molecular Biology*, 1561, 235–254. https://doi.org/10.1007/978-1-4939-6798-8_14

Bravo Díaz (2017). Caracterización de las interacciones proteína-ligando en organismos mesófilos y termófilos utilizando técnicas de minería de datos [Tesis de pregrado]. Universidad de Talca

Candian, V., Pansa, M. G., Santoro, K., Spadaro, D., Tavella, L., & Tedeschi, R. (2020). Photoselective exclusion netting in apple orchards: effectiveness against pests and impact on beneficial arthropods, fungal diseases and fruit quality. *Pest Management Science*, 76(1), 179–187. <https://doi.org/10.1002/ps.5491>

Canner, J. E., McEligot, A. J., Pérez, M.-E., Qian, L., & Zhang, X. (2017). Enhancing Diversity in Biomedical Data Science. *Ethnicity & Disease*, 27(2), 107. <https://doi.org/10.18865/ed.27.2.107>

Carugo, O., & Eisenhaber, F. (2010). Data Mining Techniques for the Life Sciences. In O. Carugo & F. Eisenhaber (Eds.), Biochemistry (Moscow) (Vol. 609, Issue 4). Humana Press. <https://doi.org/10.1007/978-1-60327-241-4>

Cheng, Y., Lin, Y., Cao, H., & Li, Z. (2020). Citrus postharvest green mold: Recent advances in fungal pathogenicity and fruit resistance. *Microorganisms*, 8(3), 1–18. <https://doi.org/10.3390/microorganisms8030449>

Chrisfield, B. J., Hopfer, H., & Elias, R. J. (2021). Impact of copper-based fungicides on the antioxidant quality of ethanolic hop extracts. *Food Chemistry*, 355(March), 129551. <https://doi.org/10.1016/j.foodchem.2021.129551>

Contigliani, E. V., Jaramillo-Sánchez, G., Castro, M. A., Gómez, P. L., & Alzamora, S. M. (2018). Postharvest Quality of Strawberry Fruit (*Fragaria x Ananassa* Duch cv. Albion) as Affected by Ozone Washing: Fungal Spoilage, Mechanical Properties, and Structure. *Food and Bioprocess Technology*, 11(9), 1639–1650. <https://doi.org/10.1007/s11947-018-2127-0>

CropLife International. (2021). *Fungal control agents sorted by cross resistance pattern and mode of action (including coding for FRAC Groups on product labels)*. 1–17. https://www.frac.info/docs/default-source/publications/frac-code-list/frac-code-list-2021--final.pdf?sfvrsn=f7ec499a_2

De Curtis, F., Ianiri, G., Raiola, A., Ritieni, A., Succi, M., Tremonte, P., & Castoria, R. (2019). Integration of biological and chemical control of brown rot of stone fruits to reduce disease incidence on fruits and minimize fungicide residues in juice. *Crop Protection*, 119(December 2018), 158–165. <https://doi.org/10.1016/j.cropro.2019.01.020>

Delgado-Cerrone, L., Mondino-Hintz, P., & Alaniz-Ferro, S. (2016). Botryosphaeriaceae species associated with stem canker, die-back and fruit rot on apple in Uruguay. *European Journal of Plant Pathology*, 146(3), 637–655. <https://doi.org/10.1007/s10658-016-0949-z>

Díaz, G. A., Auger, J., Besoain, X., Bordeu, E., & Latorre, B. A. (2013). Prevalence and pathogenicity of fungi associated with grapevine trunk diseases in Chilean vineyards. *Ciencia e Investigación Agraria*, 40(2), 327–339. <https://doi.org/10.4067/s0718-16202013000200008>

Díaz, G. A., Mostert, L., Halleen, F., Lolas, M., Gutierrez, M., Ferrada, E., & Latorre, B. A. (2019). Diplodia seriata associated with botryosphaeria canker and dieback in apple trees in Chile. *Plant Disease*, 103(5), 2014–2018. <https://doi.org/10.1094/PDIS-10-18-1785-PDN>

Engel, C., Sainsbury, S., Cheung, A. C., Kostrewa, D., & Cramer, P. (2013). RNA polymerase i structure and transcription regulation. *Nature*, 502(7473), 650–655. <https://doi.org/10.1038/nature12712>

Esser, L., Quinn, B., Li, Y. F., Zhang, M., Elberry, M., Yu, L., Yu, C. A., & Xia, D. (2004). Crystallographic studies of quinol oxidation site inhibitors: A modified classification of inhibitors for the cytochrome bc1 complex. *Journal of Molecular Biology*, 341(1), 281–302. <https://doi.org/10.1016/j.jmb.2004.05.065>

Esterio, M., Copier, C., Román, A., Araneda, M. J., Rubilar, M., Pérez, I., & Auger, J. (2017). Frecuencia de poblaciones de botrytis cinerea resistentes a fungicidas en uva de mesa ‘thompson seedless’ en el valle central de Chile. *Ciencia e Investigacion Agraria*, 44(3), 295–306. <https://doi.org/10.7764/rcia.v44i3.1721>

Fernández, A., García, S., Herrera, F., & Chawla, N. V. (2018). SMOTE for Learning from Imbalanced Data: Progress and Challenges, Marking the 15-year Anniversary. *Journal of Artificial Intelligence Research*, 61, 863–905. <https://doi.org/10.1613/jair.1.11192>

Forli, S., Huey, R., Pique, M. E., Sanner, M. F., Goodsell, D. S., & Olson, A. J. (2016). Computational protein–ligand docking and virtual drug screening with the AutoDock suite. *Nature Protocols*, 11(5), 905–919. <https://doi.org/10.1038/nprot.2016.051>

Franco, D., de Goes, A. D., & Pereira, F. D. (2020). Sources and concentrations of cupric fungicides for the control of citrus black spot. *Revista Caatinga*, 33(1), 1–8. <https://doi.org/10.1590/1983-21252020v33n101rc>

Greene, A. C., Giffin, K. A., Greene, C. S., & Moore, J. H. (2016). Adapting bioinformatics curricula for big data. *Briefings in Bioinformatics*, 17(1), 43–50. <https://doi.org/10.1093/bib/bbv018>

Gupta, P. K. (2018). Toxicity of Fungicides. In *Veterinary Toxicology: Basic and Clinical Principles: Third Edition* (Third Edit). Elsevier Inc. <https://doi.org/10.1016/B978-0-12-811410-0.00045-3>

Gupta, P. K. (2017). Herbicides and fungicides. In *Reproductive and Developmental Toxicology*. Elsevier Inc. <https://doi.org/10.1016/B978-0-12-804239-7.00037-8>

Hermann, D., & Stenzel, K. (2019). FRAC Mode-of-action Classification and Resistance Risk of Fungicides. In *Modern Crop Protection Compounds* (Vol. 1, pp. 589–608). Wiley-VCH Verlag GmbH & Co. KGaA. <https://doi.org/10.1002/9783527699261.ch14>

Hessa, T., Kim, H., Bihlmaier, K., Lundin, C., Boekel, J., Andersson, H., Nilsson, I., White, S. H., & von Heijne, G. (2005). Recognition of transmembrane helices by the endoplasmic reticulum translocon. *Nature*, 433(7024), 377–381. <https://doi.org/10.1038/nature03216>

Ji, J. Y., Yang, J., Zhang, B. W., Wang, S. R., Zhang, G. C., & Lin, L. N. (2020). Sodium pheophorbide a controls cherry tomato gray mold (*Botrytis cinerea*) by destroying fungal cell structure and enhancing disease resistance-related enzyme activities in fruit. *Pesticide Biochemistry and Physiology*, 166(April), 104581. <https://doi.org/10.1016/j.pestbp.2020.104581>

Kim, S., Chen, J., Cheng, T., Gindulyte, A., He, J., He, S., Li, Q., Shoemaker, B. A., Thiessen, P. A., Yu, B., Zaslavsky, L., Zhang, J., & Bolton, E. E. (2019). PubChem 2019 update: Improved access to chemical data. *Nucleic Acids Research*, 47(D1), D1102–D1109. <https://doi.org/10.1093/nar/gky1033>

Kumar, K., Woo, S. M., Siu, T., Cortopassi, W. A., Duarte, F., & Paton, R. S. (2018). Cation- π interactions in protein-ligand binding: Theory and data-mining reveal different roles for lysine and arginine. *Chemical Science*, 9(10), 2655–2665. <https://doi.org/10.1039/c7sc04905f>

Kwon, J.-H., Won, S.-J., Moon, J.-H., Lee, U., Park, Y.-S., Maung, C. E. H., Ajuna, H. B., & Ahn, Y. S. (2021). *Bacillus licheniformis* PR2 Controls Fungal Diseases and Increases Production of Jujube Fruit under Field Conditions. *Horticulturae*, 7(3), 49. <https://doi.org/10.3390/horticulturae7030049>

Lewis, K., & Tzilivakis, J. (2017). Development of a data set of pesticide dissipation rates in/on various plant matrices for the pesticide properties database (PPDB). *Data*, 2(3). <https://doi.org/10.3390/data2030028>

Lin, S., Taylor, N. J., & Hand, F. P. (2018). Identification and characterization of fungal pathogens causing fruit rot of deciduous holly. *Plant Disease*, 102(12), 2430–2445. <https://doi.org/10.1094/pdis-02-18-0372-re>

Mendoza, L., Navarro, F., Melo, R., Báez, F., & Cotoras, M. (2019). Characterization of polyphenol profile of extracts obtained from grape pomace and synergistic effect of these extracts and fungicides against Botrytis Cinerea. *Journal of the Chilean Chemical Society*, 64(4), 4607–4609. <https://doi.org/10.4067/S0717-97072019000404607>

Meng, F., Xi, Y., Huang, J., & Ayers, P. W. (2021). A curated diverse molecular database of blood-brain barrier permeability with chemical descriptors. *Scientific Data*, 8(1), 1–11. <https://doi.org/10.1038/s41597-021-01069-5>

Miyamoto, T., Hayashi, K., & Ogawara, T. (2020). First report of the occurrence of multiple resistance to Flutianil and Pyriofenone in field isolates of *Podosphaera xanthii*, the causal fungus of cucumber powdery mildew. *European Journal of Plant Pathology*, 156(3), 953–963. <https://doi.org/10.1007/s10658-020-01946-6>

Moriwaki, H., Tian, Y. S., Kawashita, N., & Takagi, T. (2018). Mordred: A molecular descriptor calculator. *Journal of Cheminformatics*, 10(1), 1–14. <https://doi.org/10.1186/s13321-018-0258-y>

Möhrling, N., Kudsk, P., Jørgensen, L. N., Ørum, J. E., & Finger, R. (2021). An R package to calculate potential environmental and human health risks from pesticide applications using the ‘Pesticide Load’ indicator applied in Denmark. *Computers and Electronics in Agriculture*, 191(November), 0–2. <https://doi.org/10.1016/j.compag.2021.106498>

Mühlethaler, T., Gioia, D., Prota, A. E., Sharpe, M. E., Cavalli, A., & Steinmetz, M. O. (2021). Comprehensive Analysis of Binding Sites in Tubulin. *Angewandte Chemie - International Edition*, 60(24), 13331–13342. <https://doi.org/10.1002/anie.202100273>

Narayanan, H., Dingfelder, F., Butté, A., Lorenzen, N., Sokolov, M., & Arosio, P. (2021). Machine Learning for Biologics: Opportunities for Protein Engineering, Developability, and Formulation. *Trends in Pharmacological Sciences*, 42(3), 151–165. <https://doi.org/10.1016/j.tips.2020.12.004>

Pefaur Lepe, J. (2020). *Evolución de la Fruticultura Chilena en los Últimos 20 Años*.
<https://bibliotecadigital.odepa.gob.cl/bitstream/handle/20.500.12650/70234/evolucionFruticulturachilena.pdf>

Piatetsky-Shapiro, G., Sydney, Q. U., Bu, M. S., Kershberg, L., Quinlan, R., & Langley, P. (2000). Knowledge Discovery in Databases: 10 years after Applications and concluded with a summary panel discussion by. *Discovery*, 1, 59–61.

Ramakrishnan, G., Gowri, V. S., Mudgal, R., Chandra, N. R., & Srinivasan, N. (2014). *Chapter 1: Mining the Sequence Databases for Homology Detection: Application to Recognition of Functions of Trypanosoma brucei brucei Proteins and Drug Targets*. 3–31.
https://doi.org/10.1142/9789814551014_0001

Rojo, C., Becerra, V., France, A., Paredes, M., Buddie, A., & Balzarini, M. (2017). Genetic diversity of Chondrostereum purpureum (Pers.) Pouzar causing silverleaf disease on blueberries in Chile. *Gayana. Botánica*, 74(ahead), 0–0. <https://doi.org/10.4067/s0717-66432017005000217>

Rose, P. W., Prlić, A., Altunkaya, A., Bi, C., Bradley, A. R., Christie, C. H., Di Costanzo, L., Duarte, J. M., Dutta, S., Feng, Z., Green, R. K., Goodsell, D. S., Hudson, B., Kalro, T., Lowe, R., Peisach, E., Randle, C., Rose, A. S., Shao, C., ... Burley, S. K. (2017). The RCSB protein data bank: Integrative view of protein, gene and 3D structural information. *Nucleic Acids Research*, 45(D1), D271–D281. <https://doi.org/10.1093/nar/gkw1000>

Schorn, M. A., Verhoeven, S., Ridder, L., Huber, F., Acharya, D. D., Aksенов, A. A., Aleti, G., Moghaddam, J. A., Aron, A. T., Aziz, S., Bauermeister, A., Bauman, K. D., Baunach, M., Beemelmanns, C., Beman, J. M., Berlanga-Clavero, M. V., Blacutt, A. A., Bode, H. B., Boullie, A., ... van der Hooft, J. J. J. (2021). A community resource for paired genomic and metabolomic data mining. *Nature Chemical Biology*, 17(April).
<https://doi.org/10.1038/s41589-020-00724-z>

Schweigel, H., Wicht, M., & Schwendicke, F. (2016). Salivary and pellicle proteome: A datamining analysis. *Scientific Reports*, 6, 1–12. <https://doi.org/10.1038/srep38882>

Sessa, L., Abreo, E., Bettucci, L., & Lupo, S. (2016). Botryosphaeriaceae species associated with wood diseases of stone and pome fruits trees: symptoms and virulence across different hosts in Uruguay. *European Journal of Plant Pathology*, 146(3), 519–530. <https://doi.org/10.1007/s10658-016-0936-4>

Shaffer, J. G., Mather, F. J., Wele, M., Li, J., Tangara, C. O., Kassogue, Y., Srivastav, S. K., Thiero, O., Diakite, M., Sangare, M., Dabitao, D., Toure, M., Djimde, A. A., Traore, S., Diakite, B., Coulibaly, M. B., Liu, Y., Lacev, M., Lefante, J. J., ... Doumbia, S. O. (2019). Expanding research capacity in sub-Saharan Africa through informatics, bioinformatics, and data science training programs in Mali. *Frontiers in Genetics*, 10(APR), 1–13. <https://doi.org/10.3389/fgene.2019.00331>

Silva, C. J., van den Abeele, C., Ortega-Salazar, I., Papin, V., Adaskaveg, J. A., Wang, D., Casteel, C. L., Seymour, G. B., & Blanco-Ulate, B. (2021). Host susceptibility factors render ripe tomato fruit vulnerable to fungal disease despite active immune responses. *Journal of Experimental Botany*. <https://doi.org/10.1093/jxb/eraa601>

Silva-Valderrama, I., Toapanta, D., Miccono, M. de los A., Lolas, M., Díaz, G. A., Cantu, D., & Castro, A. (2021). Biocontrol Potential of Grapevine Endophytic and Rhizospheric Fungi Against Trunk Pathogens. *Frontiers in Microbiology*, 11(January), 1–13. <https://doi.org/10.3389/fmicb.2020.614620>

Souleyre, E. J. F., Bowen, J. K., Matich, A. J., Tomes, S., Chen, X., Hunt, M. B., Wang, M. Y., Illeperuma, N. R., Richards, K., Rowan, D. D., Chagné, D., & Atkinson, R. G. (2019). Genetic control of α -farnesene production in apple fruit and its role in fungal pathogenesis. *Plant Journal*, 100(6), 1148–1162. <https://doi.org/10.1111/tpj.14504>

Speck-Planche, A., Kleandrova, V. V., Luan, F., & Cordeiro, M. N. D. S. (2012). Predicting multiple ecotoxicological profiles in agrochemical fungicides: A multi-species chemoinformatic approach. *Ecotoxicology and Environmental Safety*, 80, 308–313. <https://doi.org/10.1016/j.ecoenv.2012.03.018>

Sun, F., Huo, X., Zhai, Y., Wang, A., Xu, J., Su, D., Bartlam, M., & Rao, Z. (2005). Crystal structure of mitochondrial respiratory membrane protein Complex II. *Cell*, 121(7), 1043–1057. <https://doi.org/10.1016/j.cell.2005.05.025>

Triantafyllidis, V., Zotos, A., Kosma, C., & Kokkotos, E. (2020). Environmental Implications from Long-term Citrus Cultivation and Wide Use of Cu Fungicides in Mediterranean Soils. *Water, Air, and Soil Pollution*, 231(5). <https://doi.org/10.1007/s11270-020-04577-z>

Valdés-Gómez, H., Acevedo-Opazo, C., Pañitrur-De la Fuente, C., Verdugo-Vásquez, N., Bratti, J., & Donoso, E. (2017). Evaluation of three control strategies against grapevine powdery mildew in the central region of Chile. *GiESCO, November*, 1001–1006.

Van der Aalst, W. (2016). Process mining: Data science in action. *Process Mining: Data Science in Action, April 2014*, 1–467. <https://doi.org/10.1007/978-3-662-49851-4>

Villoutreix, P. (2021). What machine learning can do for developmental biology. *Development* (Cambridge), 148(1). <https://doi.org/10.1242/dev.188474>

Wood, A. R. (2017). Fungi and invasions in South Africa. *Bothalia*, 47(2), 1–16. <https://doi.org/10.4102/abc.v47i2.2124>

Yu, S., Guan, Y., & Dai, Y. (2003). Application of data mining in ESMC. *Jisuanji Gongcheng/Computer Engineering*, 29(19), 90. <https://doi.org/10.4028/www.scientific.net/amm.0.1776>